

Bootstrap3D: Improving Multi-view Diffusion Model with Synthetic Data

Supplementary Material

A. Evaluation on wild prompts from real users

The results of the main part of the paper are only tested on GPT generated prompts. To test our work’s capability in wild cases, we also collect real user prompts and compare our method with Instant3D [49]. Specifically, we randomly collect 100 prompts from <https://www.meshy.ai/> and test the CLIP-R precision as well as GPT based evaluation (detailed in Sup. H). Results and some qualitative cases are shown in Tab. 9 and Fig. 9. We highlight that our Bootstrap3D excels Instant3D [49] when tested on real user prompts through training on synthetic data.

B. More Visualization Compared to other Methods.

We show more visualization of the quantitative experiments shown in the main paper in Fig. 10.

For Image-to-3D methods, they can sometimes produce significant motion blurring and fails when the input image is out-of-distribution (like the 3rd cartoon style case). We re-sample the high-quality segment of the distribution of generated images using quality filtering based on MLLM methods. Furthermore, by employing TTR, we limit the impact of these data when training multi-view diffusion models, allowing our model to produce much clearer results. In addition, we use a caption rewriting method, enabling finer prompt control for the generated multi-view images.

C. Detailed Analysis of Training Time-step Reschedule (TTR)

We first propose to use a frequency-domain energy metric, Motion Blur Frequency Energy (MBFE), to quantify the degree of motion blur in images. This metric leverages the attenuation of high-frequency components in the Fourier domain, where MBFE is defined as the proportion of high-frequency energy relative to the total spectral energy:

$$\text{MBFE} = \frac{\sum_{\rho > r} P(u, v)}{\sum P(u, v)}, \quad (3)$$

where $P(u, v) = |F(u, v)|^2$ represents the power spectral density (PSD) of the image in the frequency domain, and $\rho = \sqrt{u^2 + v^2}$ is the radial frequency. We empirically set the high-frequency threshold r to $0.4f_{\max}$, where f_{\max} is the highest spatial frequency in the Fourier transform. A lower MBFE value indicates a stronger motion blur effect, making it a reliable quantitative indicator for blur assessment.

We evaluate the proposed metric on a dataset of 500 real-world motion-blurred images and their corresponding clean counterparts from the GoPro dataset [58]. The clean images exhibit an average MBFE of 0.3724, while the blurred images have an average MBFE of 0.1372. This significant difference validates the effectiveness of MBFE as a quantitative measure for evaluating motion blur intensity.

To further investigate the correlation between motion blur and different timesteps in the diffusion model, we employ DDIM inversion using the PixArt- α [14] model on 100 motion-blurred images generated by SV3D [82], followed by resampling to obtain a new clean image (z'_0). During inversion, we use the text prompt "blurred image", while during resampling, we employ the text prompt "clean and sharp image". This setup allows us to systematically analyze the extent of residual blur present in z_0 and its progression across different timesteps z_t .

As illustrated in Fig. 5 and summarized in Tab. 8, when the inversion timestep is large, indicating greater information loss in z_t relative to z_0 (with z_{1000} representing pure random noise), the MBFE value is low. We observe a steady increase in MBFE from $t = 700$ to $t = 200$, suggesting that as z_t retains more structural details from the original image, the resampled z_0 becomes progressively clearer and sharper. However, in the range $t = 200$ to $t = 0$, the motion blur effect persists in the inverted latent representation, leading to high-fidelity output images but still contain residual blur. Based on this quantitative analysis, we set the training timestep reschedule to $t = 200$, ensuring that only the low-frequency components of the blurred images contribute to the training of the multi-view diffusion model. This strategy leads to the generation of sharp, high-quality output images with minimal motion blur artifacts.

D. Data Statistics

D.1. Data Diversity Analysis

We conduct a comprehensive analysis of data diversity in Tab. 6. Specifically, we compare the diversity of our dataset against large-scale 3D object data from Objaverse [22] and high-quality 2D image data from Laion-Aesthetics [69]. To quantify diversity, we randomly sample 20K images from each dataset and extract CLIP-B/16 [66] features from both text and images, as well as DINO-B/8 [11] features from images, reporting feature variance as a diversity metric. As shown in Tab. 6, our dataset exhibits greater diversity compared to rendered images from Objaverse. Moreover, the caption diversity generated by MV-LLaVA significantly

	Dataset	DINO [11] variance	CLIP _{img} variance	CLIP _{text} variance
	Objaverse [22]	0.251	0.279	0.251
	BootStrap w/o quality filter	0.375	0.312	0.496
	BootStrap w quality filter	0.376	0.305	0.489
	Laion Aesthetic [69]	0.431	0.359	0.696

Table 6. **Data Diversity Analysis** on 20K sampled image-text pairs from different dataset. With diversity measured by CLIP [66] and DINO [11] feature variance. The quality filtering process does not compromise diversity, ensuring a rich and varied training set.

Method	Zero123	SyncDreameer	HarmonyView	Instant3D	Zero123++	CRM	Objaverse only	Bootstrap3D
CD score	0.752	0.722	0.804	0.751	0.754	0.770	0.731	0.823

Table 7. **CD score evaluation on diversity analysis.** The test images are collected by SyncDreameer [55].

surpasses that of Cap3D [57]. Notably, our data quality filtering process does not compromise diversity, ensuring a rich and varied training set. This diversity plays a crucial role in the effective large-scale training of our multi-view diffusion model.

GPT tends to generate repetitive prompts when at scale, limiting diversity. To mitigate this, we systematically structure grammar and concepts to produce a broader range of diverse, grammatically valid prompts. Among the 100K generated prompts, 74.05% are retained after MV-LLaVA filtering, ensuring high-quality and diverse prompt selection.

D.2. Caption Analysis

Fig. 11 and 12 provide a visualization of the root noun-verb pairs for the captions generated by GPT-4V [59] and MV-LLaVA. It’s clear to see that the diversity and linguistic expression of the captions produced by MV-LLaVA are highly matched with those of GPT-4V. We believe the highly detailed description focusing on object’s texture, shape and color have potential usage beyond training multi-view diffusion model in the field like object texturing [30] and stylization [70] in Computer Graphics. MV-LLaVA can also serve as free and efficient 3D object assistant comparable with GPT-4V for future research of 3D content creation.

Fig. 13 visualizes the histogram of caption length compared with Cap3D [57]. We fine-tune MV-LLaVA to generate two different lengths suitable for different diffusion architecture, namely CLIP-based text encoding [8, 62] with 77 token length and T5 based text encoding [14, 15] with 120 token length. Both excel the length of Cap3D with less hallucinations.

D.3. Estimated Quality Analysis

For direct grasp of the quality of objaverse data and synthetic data used to train diffusion model, we randomly picked some of multi-view images from different score rank. Results are shown in Fig. 17, Fig. 18 and Fig. 19.

We use high quality data with score 4 and 5 for the training of multi-view diffusion model.

We count the number of multi-view images from different data sources, namely 660K from Objaverse, 500K from SV3D [82] and 500K from Zero123++ [71] generated by our Bootstrap3D pipeline. Result are shown in Fig.14. For Objaverse and SV3D, the assigned score is normal and we use score 4 and score 5 multi-view images as high quality data for training. However, for Zero123++, most objects are assigned with score greater than 3. We attribute this phenomenon to the fact that Zero123++ tend to generate objects with less motion blurring but more stretching and deformation compared to SV3D. Joint training of MV-LLaVA on three different data source lead to higher and more focused distribution for Zero123++’s multi-view images. For this part of synthetic data, we leave only score 5 multi-view images as high quality data.

E. Quality of MV-LLaVA

E.1. Choice of number of unfrozen layers of vision encoder.

Inspired by ShareGPT-4V [16], we unfreeze selected final layers of the CLIP [52] vision encoder during the initial phase of vision language alignment. The CLIP-L/14 model used for LLaVA [52] contains 24 transformer layers. We selectively unfreeze some of final layers to enable the CLIP model to focus more on details such as texture of multi-view images. After qualitative manual screening, we select to unfreeze eight layers to yield better results. Fig. 15 illustrates the differences between unfreezing eight layers and not unfreezing any (the original training setting of LLaVA [52]). The red sections highlight the erroneous hallucinations occurring when the vision encoder remains fully unchanged, while the green sections indicate accurate descriptions of the image content. This demonstrates that partially unfreezing the vision encoder can produce more precise captions and reduce some hallucinations.

Inversion Step	700	600	500	400	300	200	100	0 (ori image)	GoPro blur	GoPro clean
MBFE	0.2144	0.2310	0.2512	0.2896	0.3011	0.3417	0.2554	0.1973	0.1372	0.3724

Table 8. **Motion Blur Frequency Energy (MBFE) of DDIM inverted and resampled images generated by SV3D.** We observe a steady increase in MBFE from $t = 700$ to $t = 200$, suggesting that as z_t retains more structural details from the original image, the resampled z_0 becomes progressively clearer and sharper. However, in the range $t = 200$ to $t = 0$, the motion blur effect persists in the inverted latent representation, leading to high-fidelity output images but still contain residual blur.

Method	CLIP based metric	GPTEval3D	
	CLIP-R score	image-text alignment	texture detail
Instant3D (unofficial)	77.0	22.0%	24.5%
Bootstrap3D	83.5	78.0%	75.5%

Table 9. **Test results of in the wild cases.** Bootstrap3D also excels Instant3D [49] in generating high quality images according to real user prompts.

E.2. Quantitative quality study

To test the quality of our MV-LLaVA. We propose two quantitative study over the quality of captions and the alignment of quality estimation with human experts. In first study, we randomly picked 200 object from Objaverse [22] and exclude training data of MV-LLaVA. We use GPT4-V [59] and MV-LLaVA to generate descriptive captions for each object. We invite human volunteers to choose their preference over shuffled captions. Results are shown in Tab. 11, where MV-LLaVA shows comparable captioning ability with powerful GPT4-V [59], which is essential to generate millions of high quality image-text pairs for the training of text to multi-view image diffusion model.

Second experiment studies MV-LLaVA’s ability in quality estimation of both 3D assets and generated multi-view images. We invite human volunteers to estimate the quality of multi-view images rendered from Objaverse [22] or generated by SV3D [82]. As there is no golden standard for multi quality classification, We ask them to separate the randomly select multi-view images into approximately two half and serve as GT quality. We use MV-LLaVA to estimate the quality of these images and generate confusion matrix. Results are shown in Tab. 1. Given the great amount of source data of 3D assets and infinite synthetic data, we care more about the false positive rate, as these data will be mixed into training data. In this observation, we highlight the false positive rate of over 20% for SV3D [82] generated multi-view images. This result align with the observation of inevitable motion blurring of SV3D [82]. To leverage this part of data source for data diversity without hurting the final quality. We propose Training Noise Reschedule to avoid samplings from these synthetic data when time step is small.

E.3. Qualitative caption quality study

We selective compare some of the captions generated by Cap3D [57] and MV-LLaVA in Fig. 16. Our MV-LLaVA can generate more detailed descriptive captions with less hallucinations.

F. Details of Prompt Design

F.1. Prompts for GPT-4V for Quality Check

Detailed prompts are shown in Fig. 20.

F.2. Prompts for MV-LLaVA Instruct Tuning

Detailed prompts are shown in Fig. 13 and Fig. 14

G. Generation Diversity Analysis.

We follow HarmonyView [89] measuring the diversity using the CD score with test image collected by SyncDreamer [55]. As reported in Tab. 7, the model trained on Objaverse only shows lower CD score while the model trained with Bootstrap3D significantly improves CD score compared to HarmonyView [89]. This results aligns with the study of data diversity in Appendix D.1.

H. GPT-4V based 3D Object Generation Evaluation.

We adopt method proposed in GPTEval3D [91] for more thorough and human-aligned evaluation of the quality of generated object by different methods. A full test case is shown in Fig. 21. Left 9-view image is rendered from object generated by Bootstrap3D and the right one generated by Instant3D [49]. We ask GPT-4V to mainly evaluate through comparison based on three dimensions: text-image alignment, low-level texture quality and 3D plausibility. The answer of GPT-4V shows its in depth perception ability of

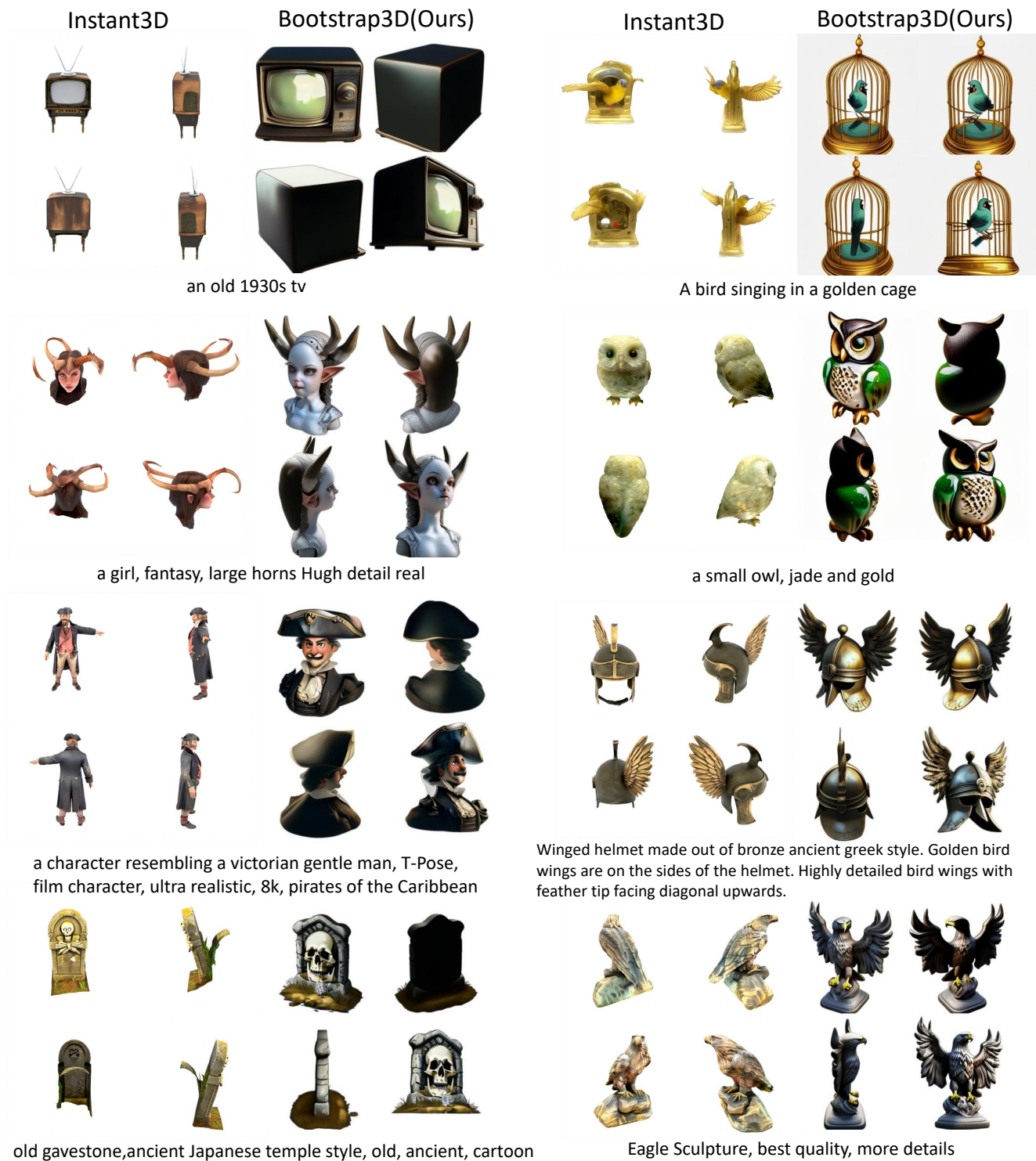


Figure 9. **Real user prompt cases** visualization compared to Instant3D [49]

given reasonable comparison well aligned with human preference. We thus choose to use GPT-4V rather than human volunteers to give reasonable evaluation.

We adopt the 110 test prompts proposed in GPTEval3D [91] to test Bootstrap3D generated object comparing with Instant3D [49], Zero123++ [71] and MVDream [72].

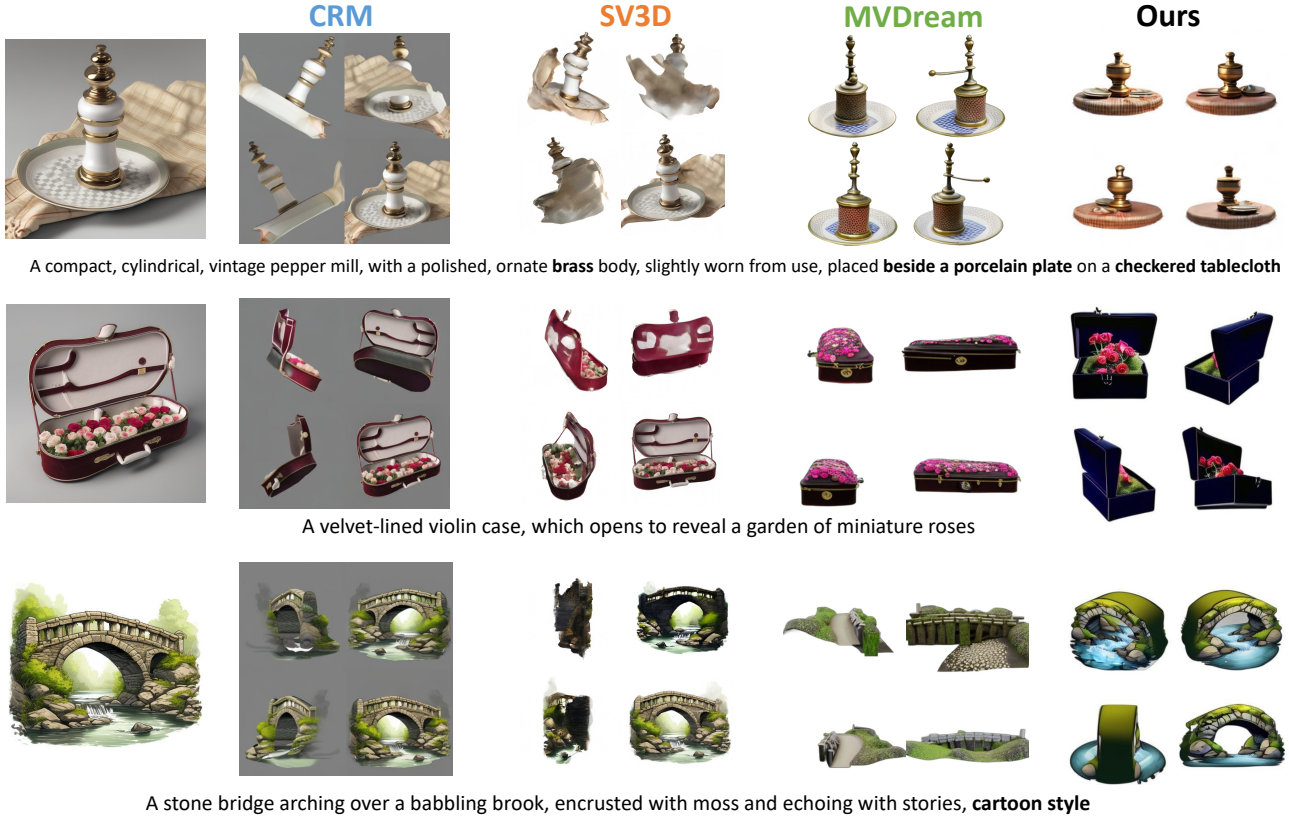


Figure 10. **Generated multiview images compare to other methods.** Our method can generate multi-view images with long text control without encountering blurring effect from data generated by SV3D thanks to TTR and quality filtering.

Lexical	n.	adj.	adv.	v.	num.	prep.
GPT-4V [59]	29.1%	16.0%	1.5%	11.1%	0.5%	9.0%
BS-Description	28.5%	16.0%	1.4%	10.8%	0.3%	8.6%
BS-Caption	30.2%	23.0%	0.3%	5.6%	0.1%	8.9%

Table 10. **Comparison of lexical composition of the captions** generated by GPT4-Vision and Share-Captioner.

For each methods, we conditioned model based on 110 test prompts with 4 different seeds, with each methods generates 440 objects, we make 1-to-1 comparison following aforementioned test setting. Results are reported in Tab. 16. Except MVDream [72] (SDS) (which generates single object consuming 30 mins while Bootstrap3D only need 5 seconds.). Bootstrap3D excels in all three evaluation dimensions, which proves the ability of Bootstrap3D in creating high quality 3D objects.

I. User Preference Study.

User preference study of Bootstrap3D winning rate with 47 volunteers of college students and ordinary people on 220 objects generated by different methods with text prompts

from GPTeval3D [91] compared to other methods are reported in Tab. 15, where Bootstrap3D exhibits strong performance according to object quality and object-text alignment.

J. Improving Direct 3D Generative Models

In addition to fine-tuning the multiview diffusion model, we also evaluate our framework on direct 3D generative models, circumventing the use of multi-view images as intermediaries. For this purpose, we selected the Shape-E [40] model for experiment and assess the outcomes following the testing method the same to Cap3D [57]. Specifically, we fine-tune Shape-E using 250K BS-Objaverse data, ensuring that all entries scored greater than 3, accompa-

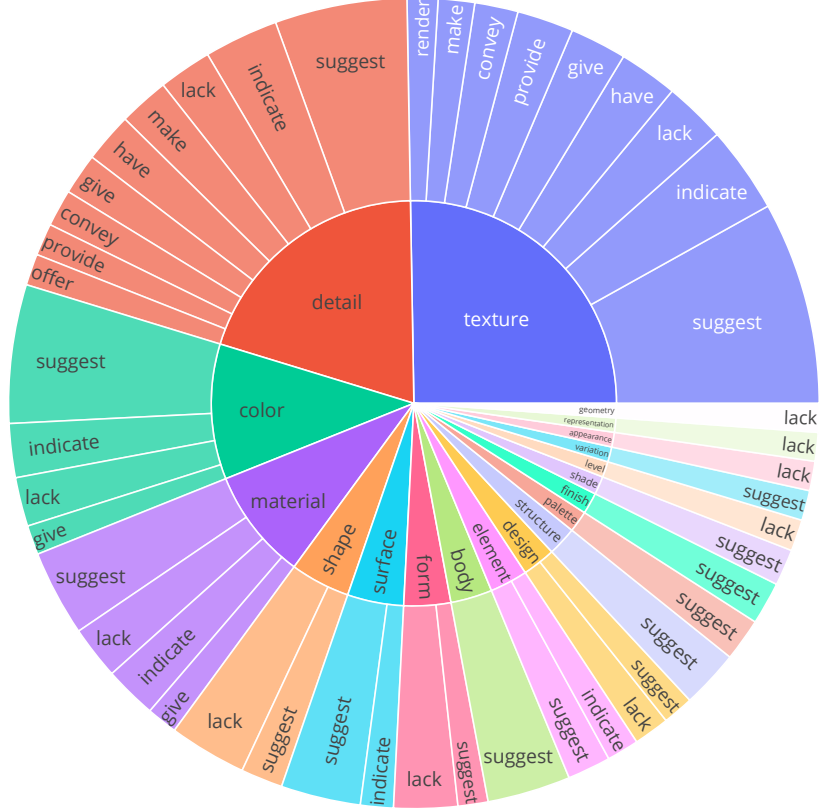


Figure 11. **Visualized analysis of dense reasoning descriptions generated by GPT4-Vision [59]** of the root noun-verb pairs (occurring over 1%) of the descriptions

Preference	GPT4-Vision [59]	MV-LLaVA	Comparable
Percentage	39.5%	34.5%	26.0%

Table 11. **Human evaluation** on the quality of generated captions from MV-LLaVA vs. GPT4-Vision [59] over 200 validation samples from Objaverse [22].

Method	FID ↓	CLIP score ↑	CLIP-R-precision ↑
Shape-E	37.2	80.4	20.3
Cap3D	35.5	79.1	20.0
Ours	35.3	81.2	22.1

Table 12. **Test results on Shape-E.** More accurate and descriptive 3D caption help model to achieve better object-text alignment.

nied by more precise and descriptive captions. The metrics for training and testing are consistent with those employed in Cap3D [57]. Some qualitative results are presented in Fig.22, where our finetuned version can generate object that follow text prompt more precisely. Quantitative results are detailed in Tab.12, where more accurate and descriptive captions than Cap3D can significantly im-

prove metrics like CLIP score [74]. Our findings indicate that improved data quality can significantly enhance object-text alignment and visual quality of Shape-E. This experiment substantiates that our pipeline, characterized by detailed captions and quality filtering, is also effective for direct 3D objects generation represented by neural field.

K. More Results Visualization

K.1. Comparison with Other Methods

As shown in Fig. 23, Fig. 24, Fig. 25, Fig. 26, Fig. 27.

K.2. Visualization of Generated objects with Different Styles

As shown in Fig. 28, Fig. 29, Fig. 30.

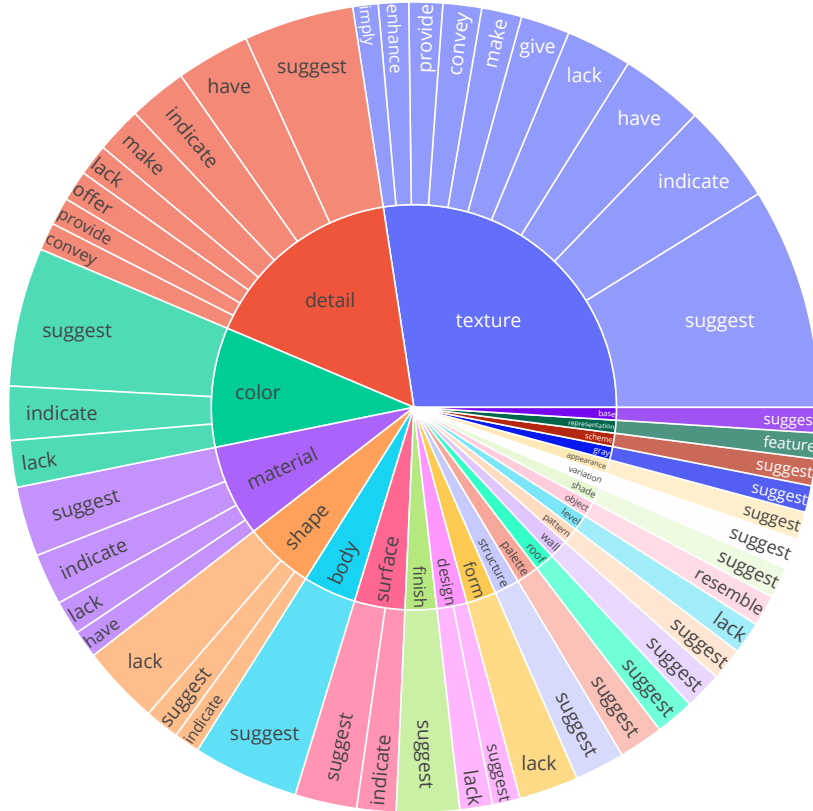


Figure 12. **Visualized analysis of dense reasoning descriptions generated by our MV-LLaVA** of the root noun-verb pairs (occurring over 1%) of the descriptions

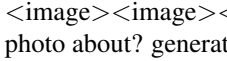
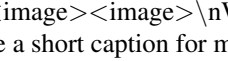
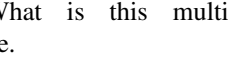

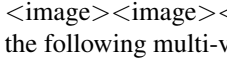
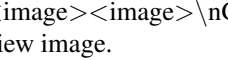
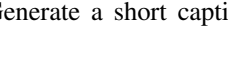

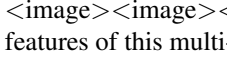
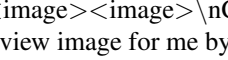
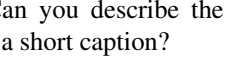

prompt type	prompt
generate caption	    <p>What is this multi-view photo about? generate a short caption for me.</p>     <p>Generate a short caption of the following multi-view image.</p>     <p>Can you describe the main features of this multi-view image for me by a short caption?</p>
reasoning	<p>How about the view consistency of this synthesized multi-view image?</p> <p>Do some comments about the view consistency of this synthesized multi-view image.</p> <p>What do you think about the view consistency of this synthesized multi-view image?</p>
quality estimation	<p>What do you think about the overall quality of view consistency of three synthesized novel views? Choosing from "poor", "relatively poor", "boardline", "relatively good", "good", "perfect".</p>

Table 13. **Instruct tuning prompt for SV3D [82] and Zero123++ [71] multi-view images**

L. Extended Evaluation with Different T2I Backbones

To validate the generalizability of our method under different T2I backbones, we conduct evaluations using SD-v2,

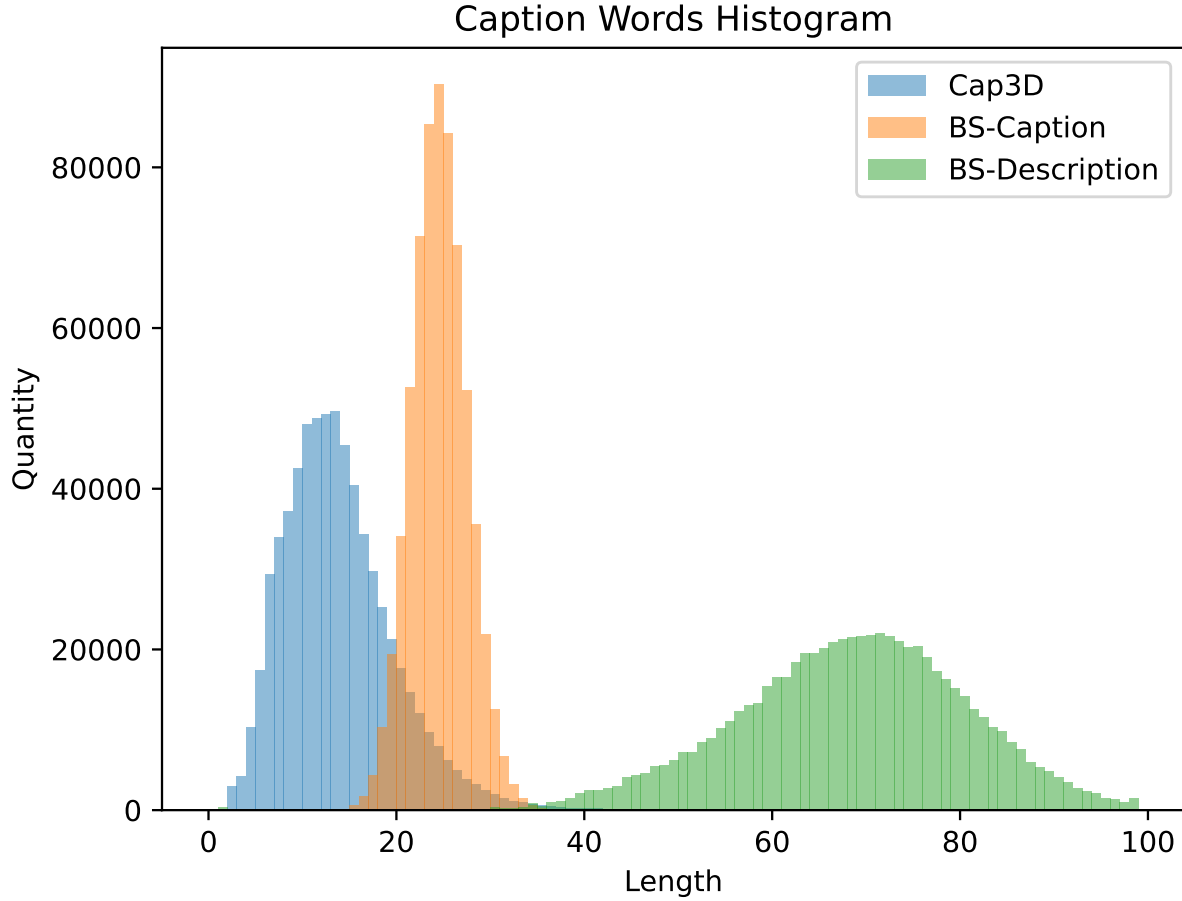


Figure 13. **Histogram Visualization of the Caption Length** compared with Cap3D [57]

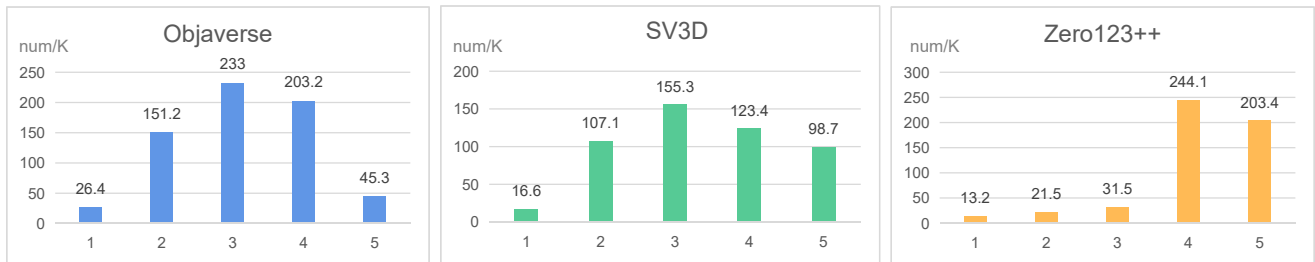


Figure 14. **Quality score statistics of different data source.**

SD-v2.1, and Pixart- α . All models are trained with our synthetic bootstrap data. The results are shown in Tab. 17.

All backbones benefit significantly from our bootstrapped data under the same training pipeline.

M. Prompt Category Analysis and LLM Translation Impact

We categorize prompts into seven types and assess performance improvements introduced by LLM-based prompt translation, using both GPT-4o and Qwen2.5-72B. Figure 33 shows CLIP-R and FID performance across categories.




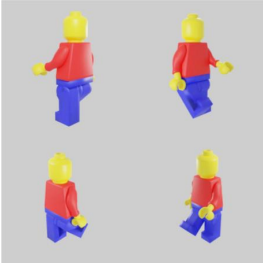
input image				
unfreeze 0 layers	The model displays a basic medieval helmet in matte olivegreen with a visor pattern and nasal guard, suitable for simple visualizations or game assets with a fantasy style. Tag: [single object] [Cartoon]	A cartoonish turtle model with a smooth, creamy yellow body, large eyes, and a friendly expression, featuring soft shading and a glossy finish for a playful appearance. Tag: [single object] [Cartoon]	A detailed classic wingback chair model, with realistic dark charcoal fabric and weathered brown wooden legs and frame, conveying an authentic and elegant aesthetic. Tag: [single object] [Photo realistic]	Stylized humanoid figure with a playful design, featuring a yellow head and hands, red torso, blue legs, and a smiling face , with a glossy finish for animation or gaming. Tag: [single object] [Cartoon]
unfreeze 8 layers	Stylized humanoid head with a green, matte finish, featuring a white symbol and purple lines, with darker green hair, suitable for fantasy or historical themes. Tag: [single object] [Cartoon]	The model displays a minimalist, cartoonish humanoid with a uniform cream color and simple black eyes, suggesting a basic prototype or abstract character design. Tag: [single object] [Cartoon]	The model showcases a wingback chair with detailed worn leather in dark brown , contrasting matte wooden legs, and a shiny, curved wooden frame, suitable for realistic interior visualizations. Tag: [single object] [Photo realistic]	A stylized humanoid figure with a glossy yellow head, red torso, and blue legs, featuring a minimalistic face and smooth surfaces, ideal for animation or game usage. Tag: [single object] [Cartoon]

Figure 15. **Qualitative results of unfreeze final layers of CLIP [66] vision encoder** compared to original fixed vision encoder setting in LLaVA [52].

prompt type	prompt
long description	<image><image><image><image>\nWhat is this multi-view photo about? generate a long descriptive caption for me. <image><image><image><image>\nGenerate a long descriptive caption of the following multi-view image. <image><image><image><image>\nCan you describe the main features of this multi-view image for me by a long descriptive caption?
caption	<image><image><image><image>\nWhat is this multi-view photo about? generate a short caption for me. <image><image><image><image>\nGenerate a short caption of the following multi-view image. <image><image><image><image>\nCan you describe the main features of this multi-view image for me by a short caption?
quality estimation	What do you think about the overall quality of this 3D model? Choosing from "poor", "relatively poor", "boardline", "relatively good", "good", "perfect".
scale tag	What do you think about the scale of the 3D model represents? Choosing from "single_object", "multi-object", "small_scene", "large_scene".
style tag	What do you think about the overall style of the 3D model? Choosing from "CAD", "Cartoon", "Photo_realistic".

Table 14. **Instruct tuning prompt for Objaverse [22] rendered multi-view images**

Model	Text-Object Alignment	Aesthetic Quality
MVDream [72]	93.15 %	75.40%
SyncDreamer [55]	96.31 %	80.19%
HarmonyView [89]	94.32 %	85.25%
CRM [86]	85.43 %	79.31%

Table 15. **User preference study** of Bootstrap3D winning rate with 47 volunteers of college students and ordinary people on 220 objects generated by different methods with text prompts from GPTeval3D [91].

	Image-text alignment	Texture quality	3D plausibility
Compared to Instant3D [49] (unofficial)	247 / 116	202 / 162	259 / 110
Compared to Zero123++ [71]	192 / 143	210 / 161	231 / 139
Compared to MVDream [72] (GRM)	290 / 71	245 / 131	284 / 102
Compared to MVDream [72] (SDS)	188 / 155	173 / 190	192 / 150

Table 16. **GPT-4V based evaluation result**, the result is in format of "number of objects preferred generated by Bootstrap3D/ that of other methods". Cases when GPT cannot answer the question or generates "cannot decide" answer are excluded.

Method	CLIP-R Score \uparrow	FID (FLUX) \downarrow	MEt3R Score \downarrow
SDv2 + Objaverse	78.3	86.1	0.2695
SDv2 + Bootstrap	87.4	49.3	0.2362
SDv2.1 + Bootstrap	87.1	47.2	0.2359

Table 17. Comparison under different backbones with and without synthetic data.

Findings:

- Our method narrows the gap between complex and simple prompts.
- Translation via LLMs improves prompt structure and generation quality, especially when paired with our synthetic data.

N. Scaling Behavior with Data Volume

To analyze the impact of training data size, we vary the volume of our synthetic dataset and evaluate performance using CLIP-R and FID. Results are presented in Fig. 34.

We observe steady improvements in both fidelity and alignment, suggesting that our method can continue to scale with more data, which is essential for generalizable generation tasks.

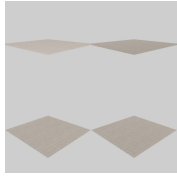
O. Broader Impacts

Potential positive societal impacts: The proposed framework, Bootstrap3D, enhances the quality and consistency of 3D models, which can benefit various industries such as entertainment, education, virtual reality, and digital art. By generating and sharing a large synthetic dataset of high-quality synthetic multi-view images, We will promotes open access to resources that can accelerate progress in the field. The model and data can serve as educational tools for stu-

dents and researchers, fostering learning and innovation in machine learning and 3D modeling.

Potential negative societal impacts: High-quality 3D models could be used to create deepfakes or misleading content, which may contribute to disinformation or malicious activities. Monitoring and Defense Mechanisms: Developing tools to detect and prevent the misuse of the generated 3D models, particularly in contexts like disinformation and surveillance. There may be unintended biases in the generated data or models, leading to unfair treatment of specific groups if the technology is deployed in applications affecting societal decision-making.

Score: 1



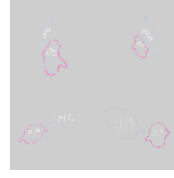
Cap3D: 3D rendering of a square, tiled floor and ceiling.

Ours: The model shows a plain with nearly identical flat surfaces with a uniform beige color, lacking any distinctive features or textures. There are no discernible details indicating what the surfaces represent or their potential use, and thus the model serves as an extremely basic representation of a flat object.



Cap3D: A collection of drawings featuring a jellyfish, a girl with various accessories, a fish, and a unicorn.

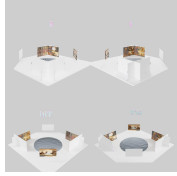
Ours: The model is a highly abstract representation of a character's head with an unclear purpose. It features thin, rod-like lines and sparse clusters of pink and blue colors, lacking any discernible texture, material quality, or detailed features.



Cap3D: A 3D object featuring elements of maps, a fish, a heart, a person, and a ghost, all adorned with pink and blue dots.

Ours: The model is composed of faint, scattered dots and lacks discernible shape or texture, making it difficult to determine its intended representation. It appears as a collection of sparse, flat details without any clear function or recognizability.

Score: 2



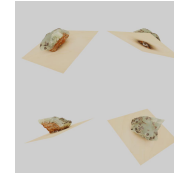
Cap3D: A 3D rendering of a white exhibition space featuring a round table, pictures, a water feature, and a chair.

Ours: The model represents a minimalistic art gallery scene with discernible but simplified elements. Semi-transparent white walls and floors suggest a clean, empty space. Slight color variations hint at lighting nuances. Sparse objects including picture frames are rendered in simplistic grayscale, lacking texture but offering a rudimentary representation of an art display.



Cap3D: A low poly 3D model of a white ghost, small bird, and teddy bear.

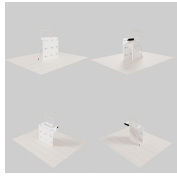
Ours: The model is a low-polygon representation of a character's head with a simplified form, featuring a broad, flat surface with minimal texture. The main body of the model is white, while the facial features are depicted in a flat grey tone, lacking depth and detail. It suggests a basic, monochromatic material, possibly indicative of clay or a digital sculpting tool...



Cap3D: 3D models of a fruit, paper airplane, rock, and green leaf with various surfaces and interactions.

Ours: The model represents a rock-like object with a textured surface, depicting moss or lichen patches in shades of green and gray. The underlying stone surface exhibits rust tones suggesting iron-rich sedimentary rock. However, the model lacks finer details such as specific rock strata or additional variations that could enhance its realism.

Score: 3



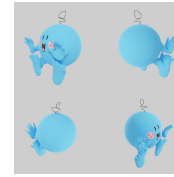
Cap3D: A 3D rendering of a white cabinet, staircase, small box, chair, desk, ceiling-mounted light fixture, and ceiling panel.

Ours: The 3D model represents a minimalist architectural form, possibly a building or pavilion with an external staircase. It's composed of white and grey tones, where white suggests a smooth concrete or plaster and grey represents metallic or stone features. The model lacks intricate textures but has discernible geometric shapes and shadows, offering a basic sense of depth and spatial relationships.



Cap3D: A 3D wooden toy car model featuring a stove, chair, pipe, spout, bottle, and steering wheel.

Ours: The model represents a whimsical, simplified vehicle consisting of a brown wagon with yellow wheels and a red, toy-like trailer being pulled. The wagon's body has a flat, wood-like texture, while the trailer's body shows a smooth, glossy finish. The handle and wheels of the trailer are colored in bright yellow, enhancing its toy-like appearance...



Cap3D: A 3D blue sphere with pink eyes and hands, resembling a cartoon character.

Ours: The model is a whimsical, cartoon-style character with a spherical blue body and exaggerated, bulbous limbs. The facial features include a wide, cheerful smile and large, expressive eyes with a simple white sclera and black pupils. The limbs' ends have been stylized into points, enhancing the playful nature of the model...

Score: 4



Cap3D: 3D model of a Japanese garden village with trees, a bridge, and a hill, featuring a plane flying overhead.

Ours: This model portrays a small, stylized scene of a Japanese garden with a pond and cherry blossom trees. The pond is depicted in a tranquil shade of blue, surrounded by vibrant green grass. Pink cherry blossom trees, a symbol of spring in Japan, are scattered around the pond with varying shades of pink. A small bridge with a red railing crosses the pond, adding a pop of color to the scene. The textures are simple yet distinct, with the grass and trees having a flat appearance...



Cap3D: A 3D model of a white, yellow, and gold machine-like spaceship.

Ours: This model is a robotic chair with a minimalist design, rendered in a clean, modern style. The seat and backrest are white, implying a smooth, possibly plastic or painted metal surface. The mechanical arm and base are a metallic gray, suggesting a metallic texture with a matte finish. Gold-colored accents, likely meant to represent controls or adjustments, add a touch of color to the model...



Cap3D: Yvy toilet paper box, Yui air purifier, Yoyo toner cartridges, Yui vitamins, and Yui sanitary napkins in yellow and white packaging with respective logos.

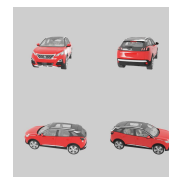
Ours: The model is a packaged product, most likely an air purifier, with a recognizable brand and model identifiable from the images. The packaging is a vibrant yellow with white accents, which gives a sense of a plastic or laminated paper material. The product image on the box shows a white device with a blue air filter, conveying a clean and modern aesthetic...

Score: 5



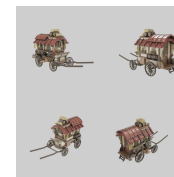
Cap3D: A 3D horn-shaped jug with a wooden handle and writing on it.

Ours: This 3D model represents a wooden tankard with a detailed wood grain texture in natural tan shades, featuring realistic knots and variations in the wood. The handle and rim of the tankard are well-defined, and the model exhibits a rustic, handcrafted appearance. The carved lines and wood texture give it an authentic look...



Cap3D: Red Peugeot 208 SUV 3D model.

Ours: The model is a red SUV with a high level of detail, including realistic shapes for the body, windows, and wheels. The body has a glossy finish, simulating automotive paint, while the black trim and silver rims have a matte texture. The wheels are detailed with intricate silver spokes that resemble alloy wheels, and the tires are a vibrant green with tread patterns adding to the model's realism...



Cap3D: 3D model of a wagon with a roof and wheels, featuring a horse.

Ours: This 3D model is a charming, medieval-style vendor cart with a high degree of texture detail, suggesting materials like weathered wood for the main structure, rusty metal for the fittings and wheel rims, and tattered fabric for the awning. The cart is laden with goods such as sacks, barrels, and hanging items...

Figure 16. **Caption comparison with Cap3D [57].** Our MV-LLaVA can generate long captions that faithfully describing 3D assets from different perspectives like color, geometry and texture.

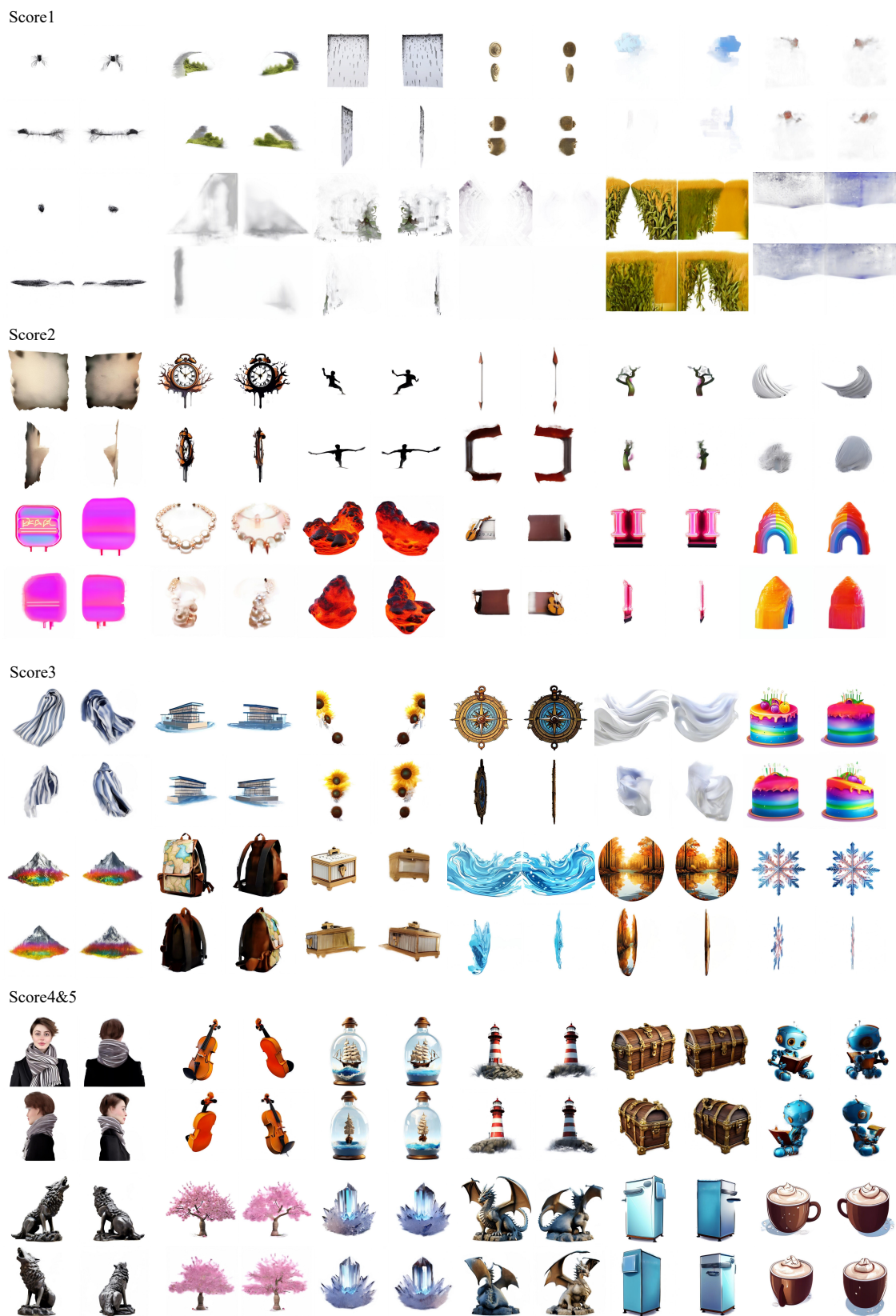


Figure 17. Randomly picked multi-view images with different scores from 500k synthetic data generated by SV3D [82].

Score: 1



Score: 2



Score: 3

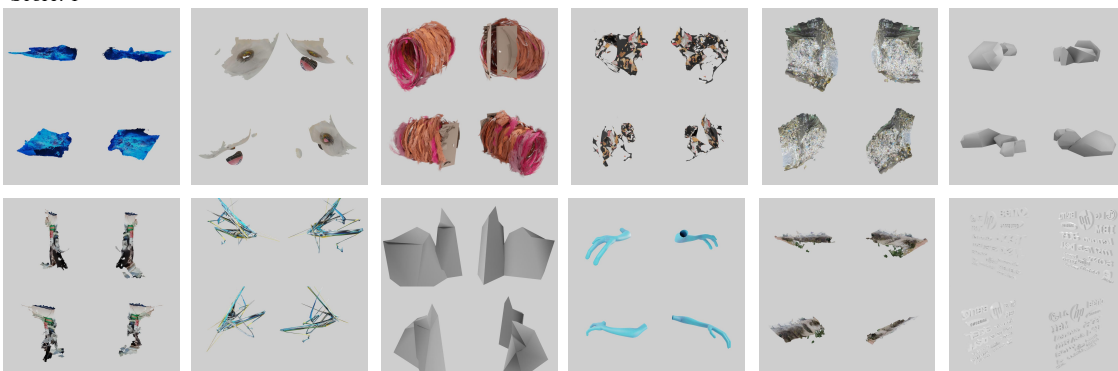


Score: 4&5

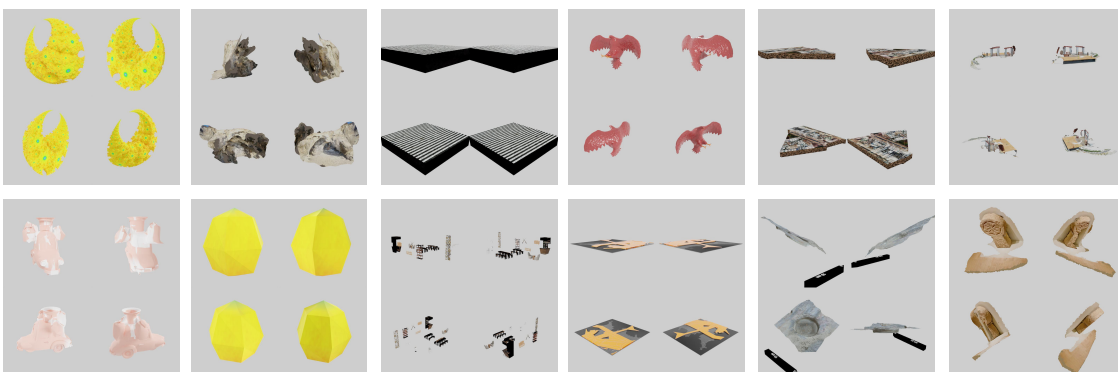


Figure 18. Randomly picked multi-view images with different scores from 500k synthetic data generated by Zero123++ [71].

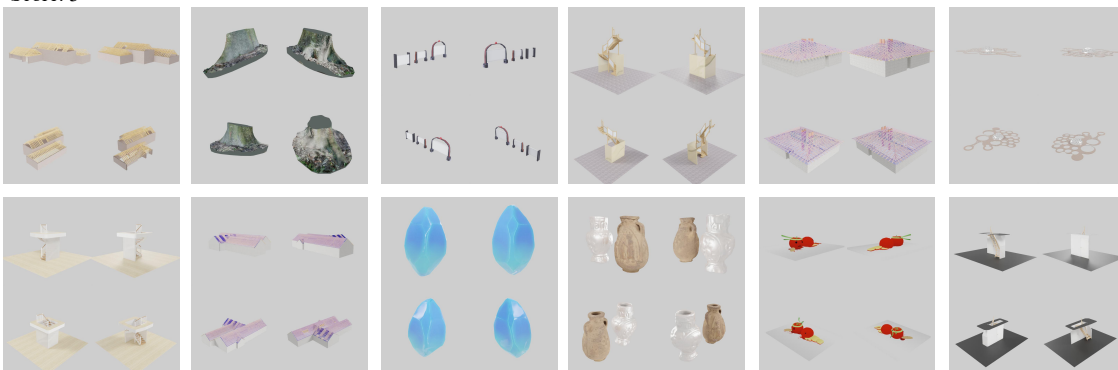
Score: 1



Score: 2



Score: 3



Score: 4&5

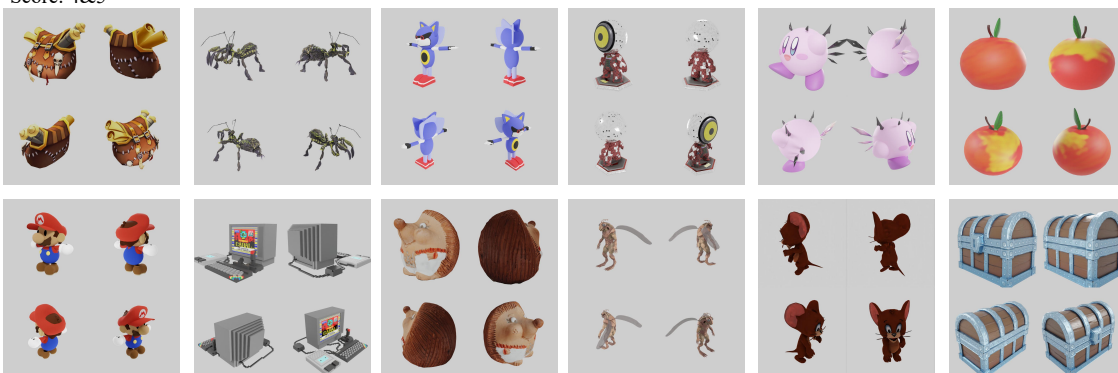


Figure 19. Randomly picked multi-view images with different scores from 660k Objaverse [22] 3D assets.

Assume you are a quality checker of a diffusion model. This diffusion model is trained to achieve novel view synthesis. I give this model the image in the upper-left side and it generate novel views in the rest three images(upper-right, lower-left, lower-right). You should tell me the quality of the generated novel view images. The score ranges from 1 to 5, representing the quality of the model from low to high. The detailed evaluation criteria are as follows:

1. The novel views are difficult to discern what the image supposed to be, lacking in recognizability. It has no usable value.
2. The novel views are distinguishable, clearly determine what the object/scene is similar to the given ground truth image. However, there is obvious inconsistency between the novel view synthesized images and ground-truth image. There are many obvious areas of image is blurred or indicating rotation.
3. The novel views are relatively good, the inconsistency between novel view synthesized images with ground-truth image is not obvious. The blurring area indicating rotation or uncertainty is acceptable for usage.
4. The novel views are pretty good, although the might be blurring areas or less resolution. the view consistency is well maintained.
5. The novel views are excellent. It is hard to tell which image from four is ground-truth and which is synthesised.

You should give me the overall score with one score number, with reason in next line. besides the quality check, I need you to generate a long descriptive caption for the scene/object from 4 different view. focusing on the part/object relative position, color, number of objects and so on with no more than 50 words and no less than 30 words. DO NOT MENTION MULTI-VIEW IMAGES FROM DIFFERENT PERSPECTIVE since it is a single scene/object. you should rearrange your result in a JSON format. if all the images(include the ground-truth image) are of low quality, just output a lowest score.

Here is an example for you:

```
{
  "score": 4,
  "reason": "The novel views generated from the model are quite convincing with a high degree of consistency in terms of texture, lighting, and color when compared to the ground-truth image. There is some minor distortion in shape and perspective, but the overall quality is high, and it maintains the realism of the scene.",
  "caption": "A cluster of shiny five apples, ranging from deep red to sunny yellow, sits comfortably within a rustic woven basket. Their smooth, round forms are grouped closely, reflecting light and casting soft shadows that accentuate their voluminous curves and vibrant colors."
}
```

This is a quad image generated from rendering a SINGLE 3D model FROM FOUR DIFFERENT views. I would like you to score the quality of this models to evaluate its current state. The score ranges from 1 to 5, representing the quality of the model from low to high. The detailed evaluation criteria are as follows:

- 1 point: The overall quality of the model is quite poor, making it difficult to discern what it is supposed to be, lacking in recognizability. The model is almost one solid block, or extremely scattered, or in fragments. It has no usable value.
- 2 points: The overall quality of the model is relatively poor, but it is possible to guess what it is, possessing low recognizability. It preliminarily has some geometric shape and can be considered a prototype model element, lacking identifiable material information, and almost has no usable value.
- 3 points: The overall quality of the model is average, it is possible to determine what it is, having certain recognizability. Different areas use different materials (colors), it preliminarily has usable value, and initially has aesthetic value.
- 4 points: The overall quality of the model is relatively high, it can be clearly determined what it is, with high recognizability. It preliminarily has certain texture details, and different parts of a model can be clearly distinguished, having high usable value and certain aesthetic value.
- 5 points: The overall quality of the model is extremely high, allowing for the classification of the model's type at a very fine granularity. It has high texture details, is a fully formed 3D model that can be used for games, simulations, or even animations, and has high aesthetic value.

After scoring, please also generate a description of the current model. If the model quality is low, only a brief description is needed; when the model quality is high, a complete description of the different details of the model is required. The description process should focus on color, material, texture details as much as possible. You can also recommended to suggest overall style. With NO MORE THAN 120 words. Especially describe color and material of different parts concretely and faithfully, let the reader easily imagine the same model.

Finally, I hope you can annotate two kinds of tags for the model. Tag1 is about the style of overall model. You can choose from [photo-realistic], [carton] and [CAD]. Tag model as [CAD] when seems like a preliminary work build by CAD software and not real. Tag model [carton] when it is good enough with carton style. Tag model [photo-realistic] when model seems like real object in the world; Tag2 is about the scale that the model represents, you can choose from [single object], [multi-object], [small scene] and [large scene]. Assign model [large scene] when it represents scene like urban street, park, etc. Assign it as [small scene] when it represents scene like inner structure or design of a house, small area, etc. Assign it as [multi-object] when it represents combination of multi objects. Assign it as [single object] when it represents single object.

Here are three examples. You should follow this format:

e.g. 1
Score: 1
Description: The model depicts a very basic and abstract urban planning concept with indistinct structures and simplistic landscaping, lacking detail and texture, appropriate for early-stage design or conceptual visualization.
Tag: [Photorealistic] [large scene]

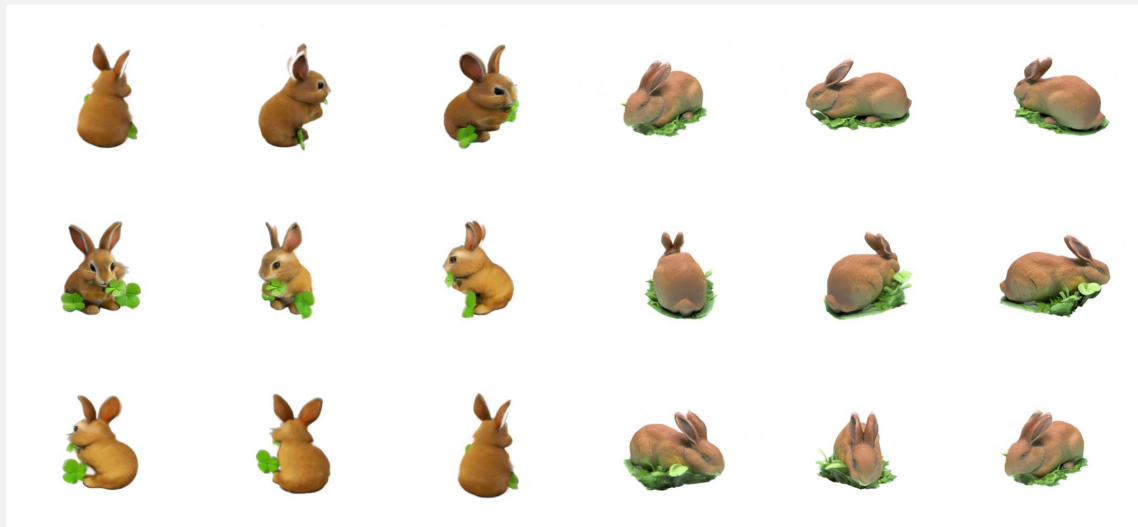
e.g. 2
Score: 2
Description: The object is a simple sphere with a homogeneous speckled texture, suggesting a stone-like material. The colors vary slightly between shades of dark gray, brown, and rust, with a matte finish. It lacks specific features or details that would indicate a higher level of complexity or function.
Tag: [Photorealistic] [single object]

e.g. 3
Score: 3
Description: The model appears to represent an architectural structure with two levels. Different colors suggest varied materials: translucent white for the structural framework, solid blue representing walls or glass panels, and yellow for interior elements, possibly stairs or floors. The style seems utilitarian, potentially for preliminary construction visualization.
Tag: [CAD] [small scene]

e.g. 4
Score: 4
The model depicts a metallic livestock handling equipment known as a cattle chute. It is rendered in shades of dark gray, conveying a metallic texture consistent with steel or iron. The structure is detailed with bolts, bars, and sliding gates, implying a sturdy construction. Text labels like "METALCORP" and "CATTLE MASTER" in blue enhance realism, suggesting a commercial quality model suitable for simulations or instructional material. The style is industrial and pragmatic.
Tag: [Photorealistic] [single object]

e.g. 5
Score: 5
The model is a stylized, anime-inspired character with a cheerful expression. Hair is rendered in a turquoise shade, contrasting with ribbons in alternate hues of pink and blue. Skin tone is in a soft peach, while the outfit combines white, grey, and gold tones, with a large yellow flower accessory. Surfaces show subtle shading, indicating variations in material. The playful, colorful appearance suggests a light-hearted, fantasy aesthetic.
Tag: [Cartoon] [single object]

Figure 20. Prompt for GPT-4V to generate caption and estimate quality of multi-view images from SV3D [82], zero123++ [71] and Objaverse [22].



Our task here is to compare two 3D objects, both generated from the same text description. I will provide you with 9 specific views of two models, where the left part of it are image rendering and normal rendering of 3D object 1, and the right part denotes those of 3D object 2.

We want to decide which one is better according to the provided 3 criteria:

1. **Text prompt and 3D shape Alignment**. This evaluates how well the 3D model corresponds to the text description. Please first describe each of the two models, and then evaluate how well it covers ALL the attributes in the original text prompt.

2. **Low-Level Texture Quality**. Focus on LOCAL parts of the RGB images: whose texture is sharper, more realistic, with high resolution, and with more details? Remember that the local texture quality could be high even if the text-3D alignment is weak, this is independently evaluated.

3. **3D Plausibility**. Look at both the RGB and normal images and imagine a 3D model from the multi-view images. Which model is more normal, solid, and plausible? Decide if there are strange body problems, duplicated parts, any noisy or meaningless 3D structures.

Take a really close look at each of the images for each of the two 3D objects before providing your answer. When evaluating these aspects, focus on one of them at a time and do not let them affect each other. Make independent decisions between these criteria.

Provide a short analysis for each of the abovementioned evaluation criteria before giving your answer. The analysis should be very concise and accurate.

For each of the criteria, you need to make a decision using these three options:

1. Left (object 1) is better;
2. Right (object 2) is better;
3. Cannot decide.

IMPORTANT: PLEASE USE THE THIRD OPTION SPARSELY.

In the last row, summarize your final decision by "<option 1> <option 2> <option 3>".

Current text prompt is: **Brown rabbit nibbling on a clover leaf**



Analysis:

1. **Text prompt & 3D Shape Alignment**: The left model seems to align more closely with the text description, with a more defined **nibbling** action on the clover leaf and a more distinct brown color. The right model also shows nibbling on a leaf, but the action is not as pronounced, and the color of the rabbit appears **less brown** and **more uniform**. The left one is better.

2. **Low-Level Texture Quality**: The left model features sharper and more detailed textures **with visible fur strands and leaves details**. The right model has smoother and less detailed textures, especially noticeable on the rabbit's fur which appears less realistic. The left one is better.

3. **3D Plausibility**: Both models seem plausible and solid from a 3D perspective. However, the left one has **more natural poses and anatomically correct rabbit shapes**, with no awkward angles or duplicated parts. The right one has a less natural resting pose and slightly unusual proportions on the head and limbs. The left one is better.

Final answer:

1 1 1

Figure 21. A test conversation with GPT-4V [59] of evaluating generated objects.

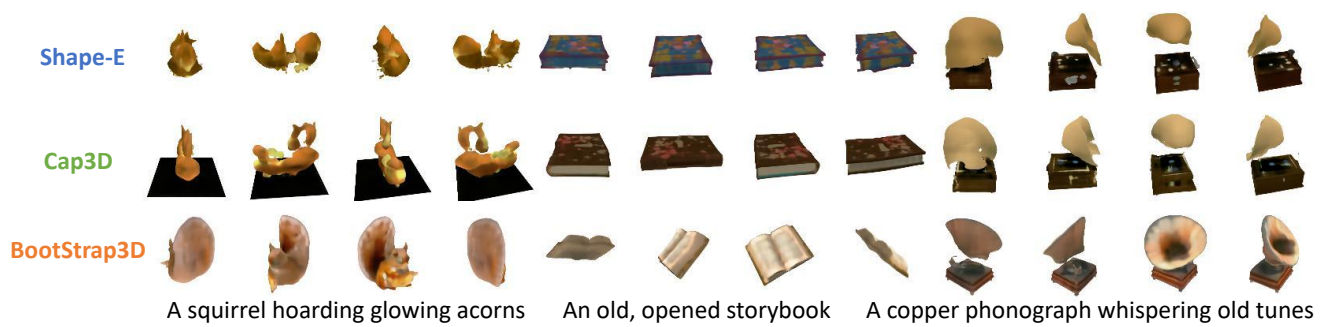


Figure 22. **Fine tuned Shape-E generation results** that shows better object-text alignment than original Shape-E [40] and finetuned version in Cap3D [57].

**A chibi phoenix reborn from ashes, flames gently
flickering around it**

A cracked teapot heating on an old stove

Zero123++

Instant3D

Ours

Zero123++

Instant3D

Ours



Figure 23. Visualization of generated objects compared to other edge-cutting methods

A galactic lighthouse guiding travelers through space-time anomalies

A miniature robot companion, poised for adventure with glowing eyes

Zero123++

Instant3D

Ours

Zero123++

Instant3D

Ours

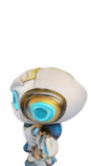


Figure 24. **Visualization of generated objects compared to other edge-cutting methods**

A tranquil, winter cabin

A serene, celestial observatory

Zero123++

Instant3D

Ours

Zero123++

Instant3D

Ours

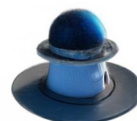
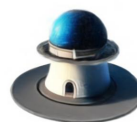
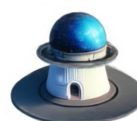


Figure 25. Visualization of generated objects compared to other edge-cutting methods

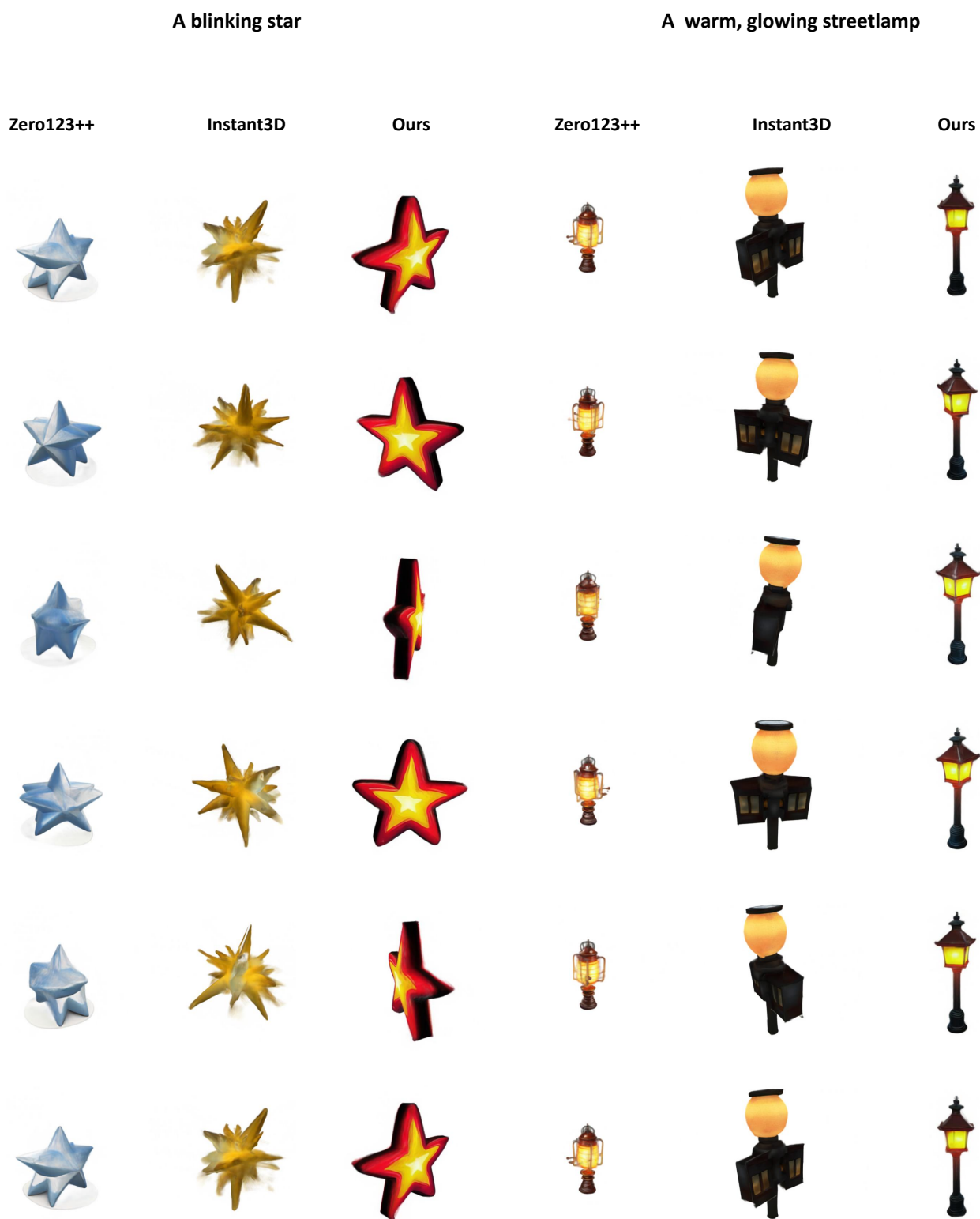


Figure 26. Visualization of generated objects compared to other edge-cutting methods

A teal cup, steaming with hot tea.

A teal cup, steaming with hot tea.

Ours



Figure 27. Visualization of generated objects compared to other edge-cutting methods



Figure 28. Visualization of generated objects compared to other edge-cutting methods with different style control.



Figure 29. Visualization of generated objects compared to other edge-cutting methods with different style control.



Figure 30. Visualization of generated objects compared to other edge-cutting methods with different style control.



Figure 31. Visualization of generated objects compared to other edge-cutting methods with different style control.



Figure 32. Visualization of generated objects compared to other edge-cutting methods with different style control.

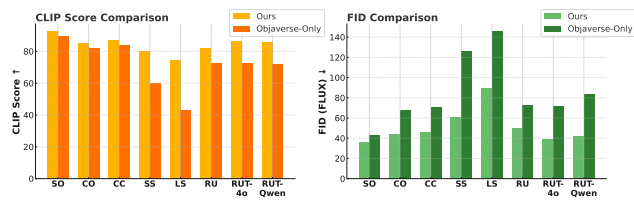


Figure 33. CLIP-R and FID by prompt type: SO (single object), CO (combined objects), CC (creature), SS/LS (small/large scene), RU (real user), RUT-4o (GPT-4o translation), RUT-Qwen (Qwen2.5-72B translation).

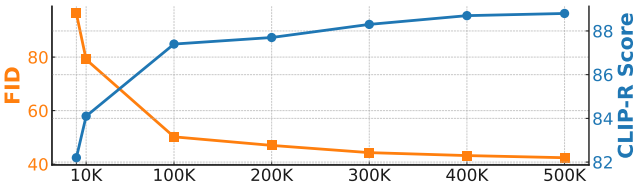


Figure 34. FID and CLIP-R vs. synthetic data volume. Larger datasets consistently lead to improved fidelity and alignment.