BOOTSTRAP3D: IMPROVING MULTI-VIEW DIFFUSION MODEL WITH SYNTHETIC DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent years have witnessed remarkable progress in multi-view diffusion models for 3D content creation. However, there remains a significant gap in image quality and prompt-following ability compared to 2D diffusion models. A critical bottleneck is the scarcity of high-quality 3D data with detailed captions. To address this challenge, we propose **Bootstrap3D**, a novel framework that automatically generates an arbitrary quantity of multi-view images to assist in training multi-view diffusion models. Specifically, we introduce a data generation pipeline that employs (1) 2D and video diffusion models to generate multi-view images based on constructed text prompts, and (2) our fine-tuned 3D-aware **MV-LLaVA** for filtering high-quality data and rewriting inaccurate captions. Leveraging this pipeline, we have generated 1 million high-quality synthetic multi-view images with dense descriptive captions to address the shortage of high-quality 3D data. Furthermore, we present a **Training Timestep Reschedule (TTR)** strategy that leverages the denoising process to learn multi-view consistency while maintaining the original 2D diffusion prior. Extensive experiments demonstrate that Bootstrap3D can generate high-quality multi-view images with superior aesthetic quality, image-text alignment, and maintained view consistency.

1 INTRODUCTION

3D content creation stands as a fundamental challenge within the generative domain, boasting widespread applications in augmented reality (AR) and game modeling. Unlike 2D image generation, the dearth of high-quality 3D models persists as a significant hurdle to overcome. In the realm of 2D image generation, the pivotal role of training on billion-scale image-text pairs (Schuhmann et al., 2022) has been firmly established (Betker et al., 2023; Rombach et al., 2022; Li et al., 2024; Chen et al., 2023a; 2024a). However, in 3D content generation, the scarcity of high-quality 3D models often compels reliance on the priors of 2D diffusion models. The predominant methodologies in this domain can be categorized into two main streams: 1) Gaining optimized neural representations from fixed 2D diffusion models via Score Distillation Sampling (SDS) loss (Poole et al., 2022; Shi et al., 2023b; Liu et al., 2023b; Shi et al., 2023a; Liu et al., 2023a; Wang et al., 2024a), which are time-intensive, lacking diversity and suffer from low robustness although capable of producing high-quality 3D objects. 2) Fine-tuning 2D diffusion models to achieve multi-view generation (Li et al., 2023a; Shi et al., 2023a;b) , directly synthesizing 3D objects through sparse reconstruction models (Li et al., 2023a; Wang et al., 2023b; Xu et al., 2024a;b; Tang et al., 2024a; Wei et al., 2024). With recent improvements in large-scale sparse view reconstruction models and 3D representations (Kerbl et al., 2023), the second stream is garnering increasing attention.

Fine-tuning 2D diffusion models for multi-view generation remains challenging owing to the insufficiency in both data quality and quantity. Previous methods (Qiu et al., 2023; Li et al., 2023a; Shi et al., 2023b; Deitke et al., 2024) primarily train on a filtered subset of high-quality data from Objaverse (Deitke et al., 2023) and Objaverse-XL (Deitke et al., 2024). The scarcity of high-quality data often introduces various shortcomings. In single-view based novel view synthesis (Liu et al., 2023b; Shi et al., 2023a; Wang & Shi, 2023; Voleti et al., 2024), if the input images deviate from the distribution of the training data, it can induce issues such as motion blurring, object distortion and deformation (Shi et al., 2023a).

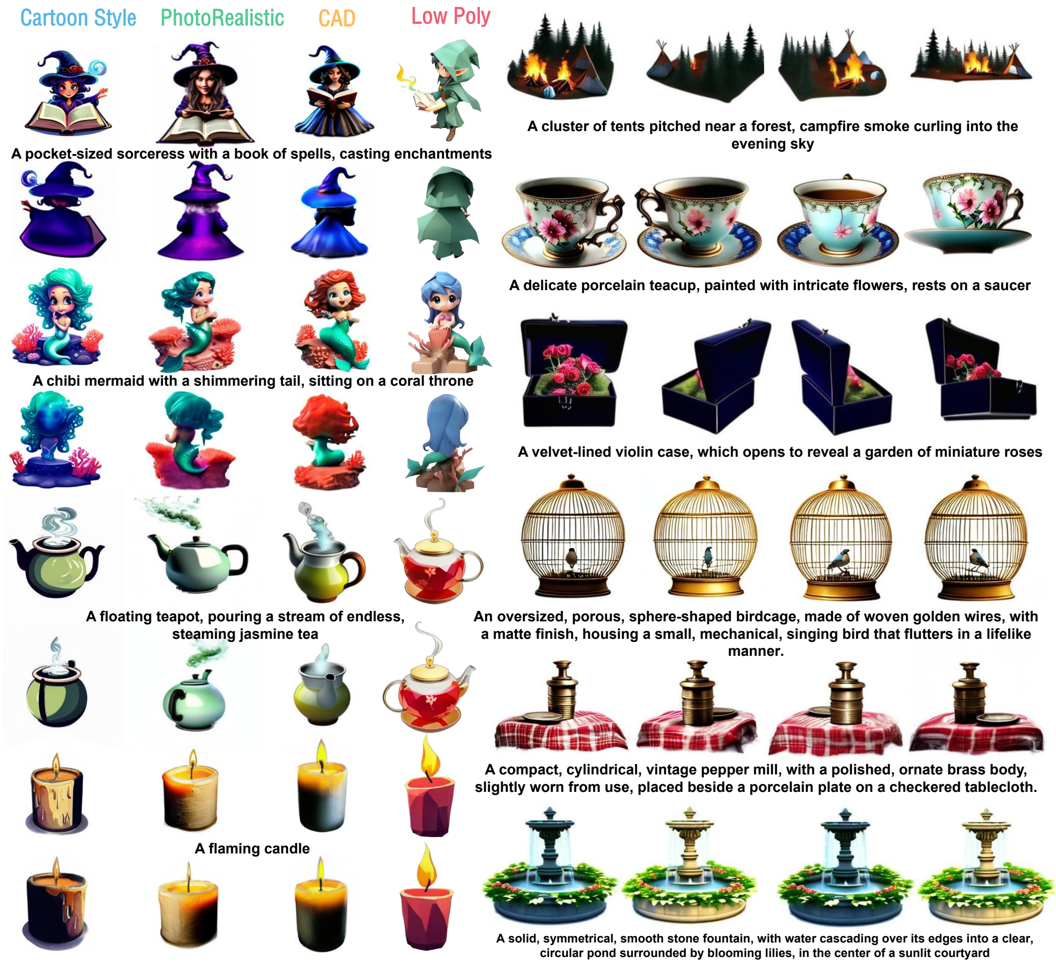


Figure 1: **Bootstrap3D** can generate high quality multi-view images with precise long text control and style customization while maintaining view consistency.

Moreover, in direct text-to-multi-view image generation, the pursuit of enhancing view consistency compromises the aesthetic and photo-realistic quality. For instance, Intant3D (Li et al., 2023a) fine-tunes SDXL (Podell et al., 2023) using only 10K high-quality Objaverse (Deitke et al., 2023) data with a small learning rate for 10K steps, which does not fundamentally prevent the catastrophic forgetting problem of losing 2D diffusion prior, leading to compromised image quality. Recent endeavors have predominantly focused on alleviating data scarcity and improving view consistency from a model-centric perspective (Kant et al., 2024; Shi et al., 2023a; Tang et al., 2024b), with limited exploration into the improvement of training data and training method itself.

Recent Multimodal Large Language Models (MLLMs) (Liu et al., 2024a; Chen et al., 2023b; Li et al., 2023b; Alayrac et al., 2022; Anil et al., 2023) like GPT-4V (OpenAI, 2023a) and Gemini (Team et al., 2023), possess image understanding capabilities and rudimentary 3D world awareness, has enabled automatic quality assessment of multi-view images and dense caption generation. Furthermore, notable advancements in video diffusion (Brooks et al., 2024; Voleti et al., 2024) have improved the generalizability of novel view synthesis (Voleti et al., 2024; Chen et al., 2024b; Kwak et al., 2023). Employing these advancements, we propose Bootstrap3D to generate synthetic data to counteract the data deficiencies inherent in training multi-view diffusion models. To be specific, we introduce the Bootstrap3D data generation pipeline for producing high-quality multi-view images with dense descriptive captions. Subsequently, we fine-tune a multi-view-aware MLLM model, dubbed as MV-LLaVA, to achieve fully automated high-quality data annotation with both efficiency and accuracy. To mitigate catastrophic forgetting of 2D diffusion prior, we introduce a training timestep reschedule (TTR) strategy when fine-tuning multi-view diffusion models. Specifically, we use the phased nature of the denoising process (Ho et al., 2020) and limit different training time steps for synthetic data to achieve enhanced image quality with maintained view consistency.

Through extensive experiments, we demonstrate that our method significantly enhances the adherence of the multi-view diffusion model to text prompts and image quality while ensuring view consistency. Integrated with the reconstruction model, our approach facilitates the creation of 3D models with superior quality. We show some of the qualitative results in Fig. 1, where our model can achieve high quality multi-view images with precise text control and style customization. Our contributions are summarized into the following points:

1) We present an automated **Bootstrap3D** data generation pipeline that uses the video diffusion model and our fine-tuned 3D-aware MV-LLaVA to synthesize an arbitrary number of high-quality multi-view image text pairs.

2) We propose a Training Time-step Reschedule (TTR) strategy for fine-tuning the multi-view diffusion model that employs both synthetic data and real data to enhance image quality and image-text alignment while maintaining view consistency.

3) We generate 1 million multi-view images with dense descriptive captions suitable for training the multi-view diffusion model and provide dense descriptive captions on Objaverse Deitke et al. (2023), which mitigates the gap with the 2D diffusion model from a data perspective.

2 Related Work

Existing 3D datasets and data pre-processing. Existing object level 3D datasets, sourced either from CAD (Chang et al., 2015; Wu et al., 2015; Deitke et al., 2023; 2024) or scan from real objects (Aanæs et al., 2016; Yao et al., 2020; Downs et al., 2022; Wu et al., 2023), are still small in size. Most state-of-the-art open-sourced 3D content creation models are trained on Objaverse (Deitke et al., 2023). However, there still exists a huge gap compared to data used for training 2D diffusion models (Schuhmann et al., 2022). In addition to quantity, quality is also an important problem remains to be solved as many methods (Shi et al., 2023b; Li et al., 2023a; Qiu et al., 2023; Tang et al., 2024a) trained on Objaverse rely on filtering out low-quality data, making the precious 3D data even less. Another critical gap that requires attention is the quality of the 3D object's caption. Previous work Cap3D (Luo et al., 2024) propose to apply BLIP-2 (Li et al., 2023b) and GPT-4 (OpenAI, 2023b) to generate caption based on multi-view images. However, this approach, without direct input image into GPT, can lead to severe hallucination. Given recent breakthroughs in improving text-image alignment through caption rewriting (Betker et al., 2023; Chen et al., 2023a; 2024a; Esser et al., 2024), there is a pressing need to rewrite denser and more accurate captions for 3D objects with the assistance of advanced Multimodal Large Language Models (MLLMs). In this work, we propose a new data generation pipeline to synthesize multi-view images and rewrite captions for 3D objects incorporating additional quality scoring mechanisms to address the aforementioned issues.

Text-to-3D content creation. The field of 3D content creation has been a vibrant area of research over the past years. One prominent research direction explores the use of Score Distillation Sampling (SDS) (Poole et al., 2022) and its variants (Chen et al., 2023c; Chung et al., 2023; Hertz et al., 2023; Liang et al., 2023; Lin et al., 2023; Liu et al., 2023b; Shi et al., 2023b; Liu et al., 2023c; Long et al., 2023; Wang et al., 2024a; Tang et al., 2023; Wang et al., 2023a; Yang et al., 2024; Qi et al., 2024), using the priors of 2D diffusion models to optimize 3D representations. While these methods have demonstrated success in producing high-quality 3D generations, they often require prolonged optimization time to converge. In contrast, recent studies (Hong et al., 2023; Wang et al., 2023b; Li et al., 2023a; Tang et al., 2024a; Tochilkin et al., 2024; Xu et al., 2024b; Wei et al., 2024) have proposed the direct inference of 3D representations (Mildenhall et al., 2021; Chan et al., 2022; Kerbl et al., 2023; Zhang et al., 2023a) conditioned by images. Among these approaches, Instant3D (Li et al., 2023a) stands out by utilizing multi-view images of the same object to directly deduce the Triplane (Chan et al., 2022) representation. This approach effectively addresses the issue of ambiguous unseen areas inherent in the single image to 3D conversions, as encountered in LRM (Hong et al., 2023) and TripoSR (Tochilkin et al., 2024). Instant3D, along with subsequent works (Xu et al., 2024b; Zheng et al., 2024; Wang et al., 2024b; Xu et al., 2024a), efficiently decomposes 3D generation into two processes: the generation of multi-view images using multi-view diffusion model (Liu et al., 2023b;c;a; Shi et al., 2023b; Liu et al., 2024b; Shi et al., 2023a; Long et al., 2023; Kant et al., 2024; Voleti et al., 2024) and large reconstruction model to generate 3D representations conditioned on these multi-view images. In this work, we introduce a method that significantly enhances the scalability of training and data generation for multi-view image generation.

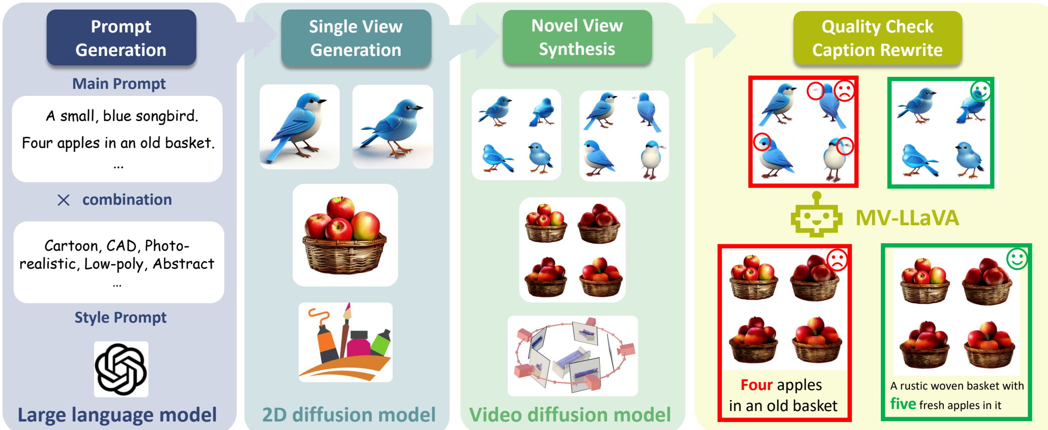


Figure 2: **Bootstrap3D data generation pipeline** that consists of 1) using LLM to generate diverse text prompts 2) employing the T2I model to generate single-view images 3) synthesizing arbitrary number of multi-view images by applying the video diffusion model, 4) employing MV-LLaVA to filter and select only high-quality data, and rewrite captions to be dense and descriptive.

Video diffusion for novel view synthesis. Recent advancements in video diffusion have marked a significant breakthrough, with models such as Sora (Brooks et al., 2024) and SVD (Blattmann et al., 2023) scaling up the direct generation process from images to videos. Following these developments, a series of works (Wang et al., 2023c; Kwak et al., 2023; Blattmann et al., 2023; Melas-Kyriazi et al., 2024; Han et al., 2024; Chen et al., 2024b) represented by SV3D (Voleti et al., 2024), have fine-tuned these video diffusion models for 3D content creation. Despite these groundbreaking developments, the new perspective images generated based on video priors still suffer from issues like motion blur. In this work, we propose to utilize SV3D (Voleti et al., 2024) as a data generator to produce novel views of given images with additional quality checks to leave only high-quality data.

Multimodal Large Language Models. With the development of large language models (Brown et al., 2020; OpenAI, 2023b; Chowdhery et al., 2022; Anil et al., 2023; Hoffmann et al., 2022; Touvron et al., 2023), multimodal large language models (MLLMs) (Zhang et al., 2023b; Alayrac et al., 2022; Li et al., 2023b; 2022; Huang et al., 2023; Driess et al., 2023; Awadalla et al., 2023; Liu et al., 2024a; Dong et al., 2024; Sun et al., 2023), such as GPT-4V (OpenAI, 2023a), have demonstrated groundbreaking 2D comprehension capabilities and open-world knowledge. As is discovered in GPTEval3D (Wu et al., 2024), GPT-4V can achieve human-aligned evaluation for multi-view images rendered from 3D objects. In this work, we fine-tune the LLaVA (Liu et al., 2024a) for quality judgment and descriptive caption generation based on multi-view images.

3 Methods

Due to the scarcity of high-quality 3D data, we develop the Bootstrap3D data generation pipeline to efficiently construct an arbitrary number of training data (Sec. 3.1). Subsequently, the quality of generated multi-view images is assessed using the powerful GPT-4V (OpenAI, 2023a) or our proposed MV-LLaVA (Liu et al., 2024a) model to generate dense descriptive captions efficiency and faithfully (Sec. 3.2). We also design a training timestep reschedule (Sec. 3.3) when fine-tuning the multi-view diffusion model with our synthetic and real data.

3.1 Bootstrap3D Data Generation Pipeline

As illustrated in Fig.2, our data generation pipeline initially employs GPT-4 (OpenAI, 2023a) to generate a multitude of imaginative and varied text prompts (Wu et al., 2024). Subsequently, to generate 2D images that closely align with the text prompts, we utilize the PixArt-Alpha (Chen et al., 2023a) model use FlanT5 (Chung et al., 2024) text encoder with DiT (Peebles & Xie, 2023) architecture for text-to-image (T2I) generation. Thereafter, we use SV3D (Voleti et al., 2024) for novel view synthesis. Given the significant motion blur and distortion often present in SV3D (Voleti et al., 2024) outputs, we further employ Multimodal Large Language Models(MLLM) to evaluate the quality of multi-view images. To rectify mismatches between multi-view images and the original

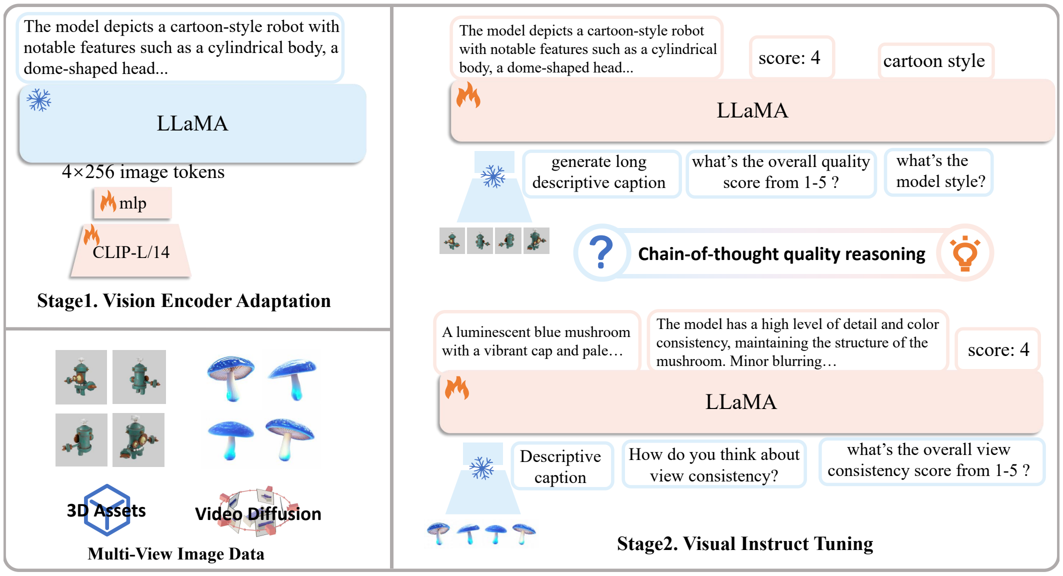


Figure 3: **MV-LLaVA.** We use GPT-4V (OpenAI, 2023a) to generate long descriptive captions, quality scoring, and reasoning processes for multi-view images to construct the instruction tuning dataset. Then we fine-tune our MV-LLaVA based on LLaVA (Liu et al., 2024a) to serve as the human-aligned quality checker and captioner for multi-view images.

text prompts induced by novel view synthesis and provide more precise captions, we further propose MV-LLaVA to generate dense descriptive captions for multi-view images.

3.2 Multi-View LLaVA (MV-LLaVA)

To efficiently generate captions and label quality scores for both generated multi-view images and 3D assets in Objaverse (Deitke et al., 2023), we propose the Multi-View LLaVA (MV-LLaVA) that fine-tune LLaVA (Liu et al., 2024a) based on our instructive conversation pairs generated by the powerful GPT-4V (OpenAI, 2023a).

Preparing the instruction tuning data. As shown in Fig.2, we use GPT-4 to generate 20k varied text prompts based on prompts designed in (Wu et al., 2024) and use PixArt-alpha (Chen et al., 2023a) to generate single view image and use SV3D (Voleti et al., 2024) or Zero123++ (Shi et al., 2023a) to generate multi-view images. For these 20k generated multi-view images, we prompt GPT-4V (OpenAI, 2023a) to generate comments on view consistency, image quality and generate dense descriptive captions. For the additional 10K rendered multi-view images from Objaverse (Deitke et al., 2023), we prompt GPT-4V (detailed prompts in Sup. A.5.1) to offer feedback on the quality and aesthetic appeal of 3D objects, along with style judgments. We utilize these 30K high-quality multi-view image text pairs (prompts detailed in Sup. A.5.2) as the instruction tuning data for LLaVA.

Instruction tuning. As presented in the left part of Fig. 3, due to the LLaVA's maximum training context length constraints of 2048, we input four images separately into CLIP-L/14 (Radford et al., 2021) and generate 4×256 image tokens. Inspired by ShareGPT-4V (Chen et al., 2023b), we freeze only a portion of layers of CLIP (Radford et al., 2021) in the first stage of pre-training to enhance multi-view awareness and texture perception of vision encoder (detailed in Sup. A.4.1). As shown in the right part of Fig. 3, we first ask the model to generate descriptions, then let the model score the quality based on multi-view images and captions. Our approach encourages LLM to deduct more reasonable quality scores through chain-of-thought (Wei et al., 2022). A mixture of original training data of LLaVA is adopted to mitigate over-fitting. As a result, we obtain MV-LLaVA, which efficiently filters and re-captions both synthetic data and 3D assets. As detailed in Sup.A.4, MV-LLaVA can not only generate more accurate, less hallucinated dense captions that faithfully describe 3D objects compared to Cap3D (Luo et al., 2024) but also assign the human-aligned quality score on both synthetic data and Objaverse assets. The filtered high-quality multi-view images with rewritten dense captions served as training data for the diffusion model.

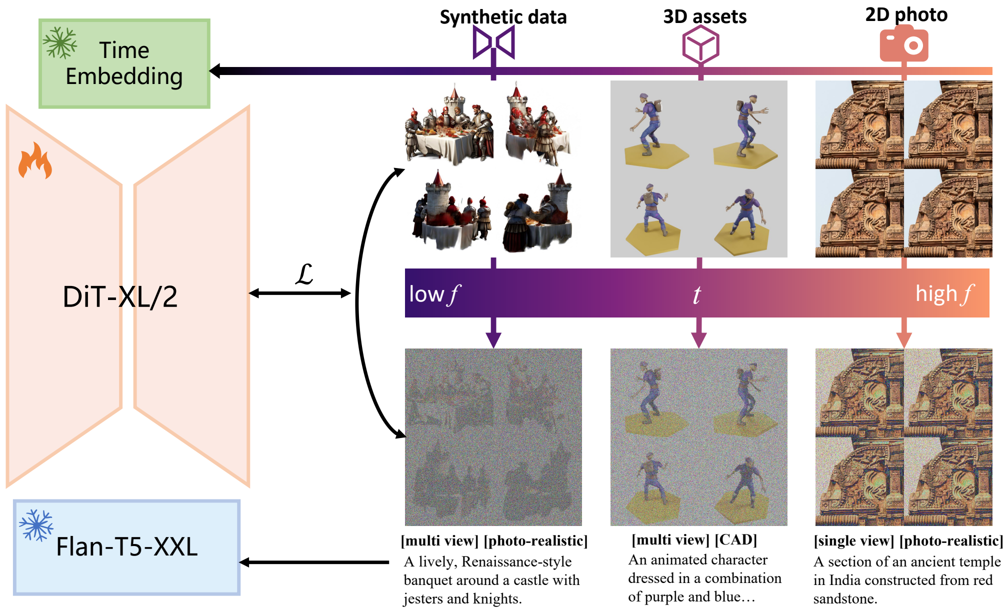


Figure 4: **Training Timestep Reschedule (TTR).** For different types of training data, we restrict the training time step t accordingly to achieve the balance between varied high aesthetic images that are better aligned with text prompt, photo-realistic texture, and view consistency for 3D generation.

3.3 Training Timestep Reschedule (TTR)

Despite retaining only relatively high-quality synthetic data with minimal motion blur from SV3D (Voleti et al., 2024) through MV-LLaVA, small areas of blurring persist, stemming from both motion and out-of-distribution scenarios for SV3D and SVD (Blattmann et al., 2023). These blurred data can potentially compromise the final performance of the multi-view diffusion model. To restrict the training time step for synthetic data, we proposed a simple yet effective Training Timestep Reschedule (TTR) method.

Background. Before delving into TTR, we briefly review some basic concepts needed to understand diffusion models (DDPMs) (Ho et al., 2020; Sohl-Dickstein et al., 2015; Salimans & Ho, 2022; Rombach et al., 2022; Chen et al., 2023a). Gaussian diffusion models assume a forward noising process which gradually applies noise to real data x_0

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{1}$$

here constants $\bar{\alpha}_t$ are hyperparameters. By applying the reparameterization trick, we can sample

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t \tag{2}$$

During training, t is randomly sampled in $[0, N]$ ($N = 1000$ in (Chen et al., 2023a; Rombach et al., 2022)) for the model to predict the added noise ϵ_t, where x_0 denotes for the clear nature image and x_N denotes for pure Gaussian noise. As depicted in Fig.4, when t is large, the denoising process primarily focuses on determining the global low frequency(f) content such as overall structure and shape. Conversely, when t is small, the denoising process is predominantly responsible for generating high f components such as texture.

When adapting Stable Diffusion (Rombach et al., 2022) for multi-view generation, the previous approach (Shi et al., 2023a) changes the default scaled linear schedule into the linear schedule to emphasize more on early denoising stage for structural variation and view consistency. Inspired by this, we propose restricting the denoising time step of synthetic data during training. As small yet observable blur still exists in synthetic data with novel view generated by SV3D (Voleti et al., 2024), we limit them to training diffusion model only with large t. This restricts the backpropagation of these synthetic data to focus on the low-frequency component of the image like the overall structure and shape that faithfully follow text prompts and consistency between different views. Small t values are only sampled on clear and physically consistent multi-view images rendered from Objaverse (Deitke et al., 2023) and supplemented high-quality 2D images from SA-1B (Kirillov et al.,

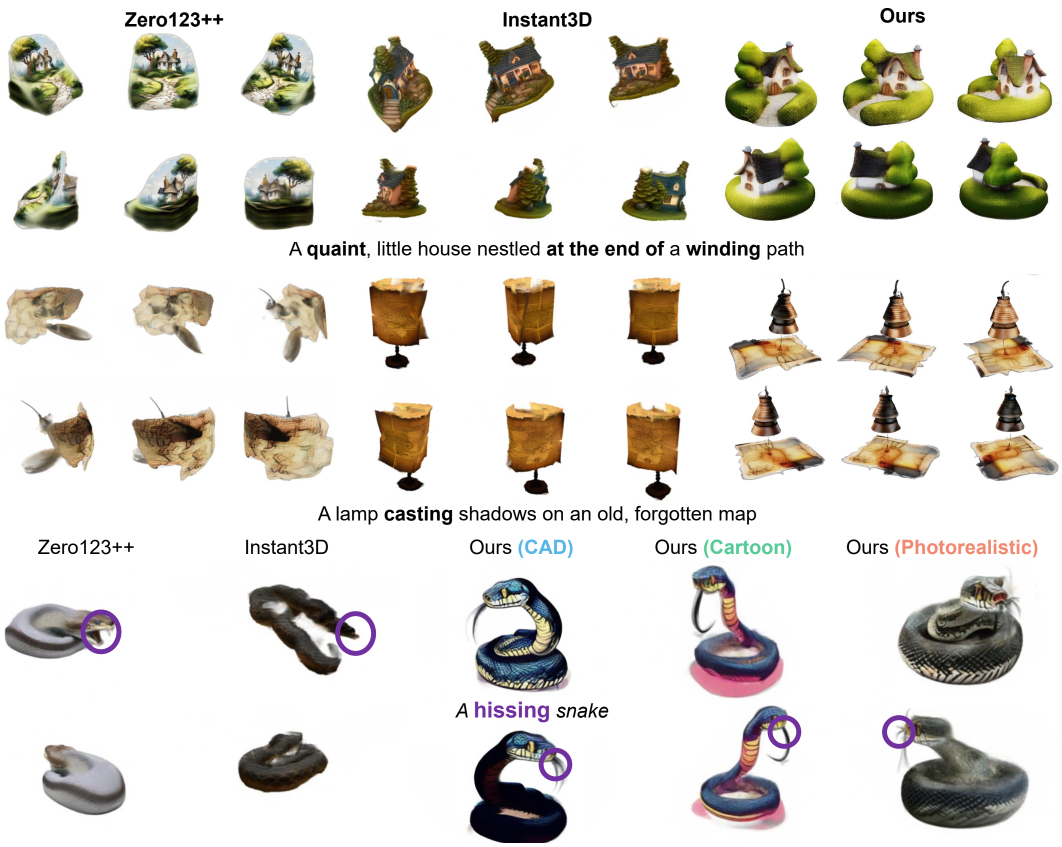


Figure 5: **Bootstrap3D generates 3D objects compared to other edge-cutting methods** given text prompt. More results with higher resolution are available in Sup.A.8.1.

2023), help model outcome high-quality images with more photo-realistic and varied texture details.

4 EXPERIMENTS

4.1 EXPERIMENT SETTINGS

Training data. For each set of 4-view images obtained from both Objaverse (Deitke et al., 2023) and generated by SV3D (Voleti et al., 2024) or Zero123++(Shi et al., 2023a), we use MV-LLaVA to generate long descriptive captions with predicted quality score. Detailed quality check of MV-LLaVA is supplied in Sup. A.4 and data analysis in Sup. A.3. In the end, we generate 200K 4-view image-text pairs on Objaverse (Deitke et al., 2023), 1000K 4-view image-text pairs from synthetic data from SV3D (Voleti et al., 2024) and Zero123++(Shi et al., 2023a). We also sample 35K HQ SA (Kirillov et al., 2023) data with captions from ShareGPT4V (Chen et al., 2023b).

Training details. We test our framework directly on the text-to-multi-view diffusion model. We fine-tune PixArt-α (Chen et al., 2023a) with backbone DiT-XL/2 (Peebles & Xie, 2023) model on the data as mentioned earlier. Similar to Instant3D (Li et al., 2023a), we train the diffusion model directly on 4-view images naturally arranged in a 2×2 grid. For 4 same view images from SA (Kirillov et al., 2023), we limit training time step $t \in [0, 50]$. We limit synthetic multi-view images $t \in [200, 1000]$. Regarding 3D object-rendered images, we do not limit t but sample more frequently in the range $[50, 200]$ as a complement. We set the total batch size to 1024 with the learning rate set to 8e-5 for 20K steps. Training is conducted on 32 NVIDIA A100-80G GPUs for 20 hours with Flan-T5-XXL (Chung et al., 2024) text features and VAE (Kingma & Welling, 2013) features pre-extracted.

Evaluation metrics. We primarily benchmark the quantitative results of our approach and other methods from two main dimensions: 1). **Image-text alignment** measured by CLIP score and CLIP-

Table 1: **Benchmark of CLIP and FID score of text-to-multi-view (T2MV) models** on generated 4 view images, CLIP score tests on 110 text prompts from GPTeval3D (Wu et al., 2024) while FID is measured with the distribution of 30K object-centric images generated by SOTA T2I models. For text-to-image-to-multi-view(T2I2MV), we input I2MV models with single view images generated by Pixart-α, which superior single view image quality is marked in green .

Domain	Method	CLIP-R Score ↑		CLIP Score ↑		FID ↓	
		CLIP-L/14	CLIP-bigG	CLIP-L/14	CLIP-bigG	PG2.5	PixArt-α
T2I	PixArt-α	96.1	94.7	25.9	41.5	20.7	5.4
T2I2MV	SV3D	78.8	81.3	24.7	37.3	55.7	54.2
	CRM	77.5	85.1	24.9	38.9	59.0	52.2
	Zero123++	78.0	84.5	24.2	36.9	53.2	49.3
T2MV	Instant3D (unofficial)	83.6	91.1	25.6	39.2	83.2	77.9
	MVDream	84.8	89.3	25.5	38.4	60.2	59.2
	Bootstrap3D	**88.8**	**92.5**	**25.8**	**40.1**	**42.4**	**31.0**

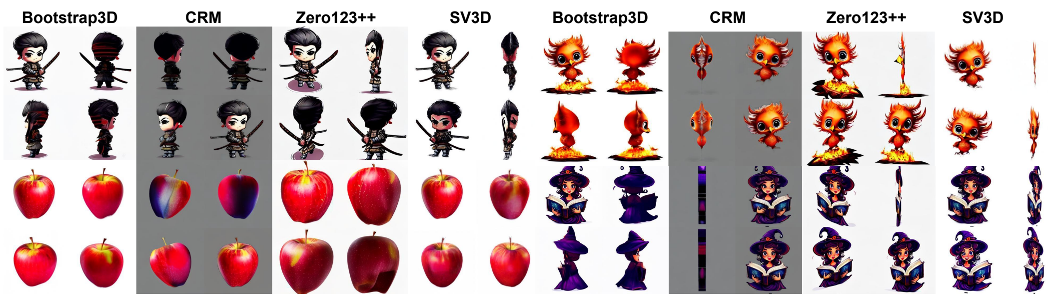


Figure 6: **Bootstrap3D can generate high quality multi-view images** in out of domain cases compare to other edge-cutting multi-view diffusion models trained on Objaverse only.

R score indicating the prompt follow ability of text-to-multi-view (T2MV) diffusion model. 2). **Quality of generated images** measured by FID (Heusel et al., 2017). Given the trend of decoupling multi-view image generation and sparse view reconstruction, we conduct tests separately on multi-view images by T2MV and rerendered images from generated 3D objects. To test the robustness and diversity of Bootstrap3D beyond prompts generated by GPT, we also collect real user prompts from public website, the details and test results are available in Sup. A.1.

Evaluation details. For CLIP-R Score and CLIP Score, we test on 110 text prompts from GPTeval3D (Wu et al., 2024) using different CLIP models (i.e., CLIP-L/14 (Radford et al., 2021) and CLIP-bigG (Ilharco et al., 2021)) following the same setting of Instant3D (Li et al., 2023a). Regarding the FID (Heusel et al., 2017) test, as there is no golden standard for HQ 3D objects, we follow the similar evaluation idea of PlayGround2.5 (Li et al., 2024) (PG2.5) to use powerful T2I model generated images to form ground truth (GT) distribution. We use curated prompts to guide powerful PixArt and PG2.5 to generate high-quality CAD-style images with a single object in the pure background. Rembg (etc, 2020) is adopted to create white background object-centric images. We use the method proposed in GPTeval3D (Wu et al., 2024) to generate 3K prompts. For both PG-2.5 and PixArt, we generate 10 images for each prompt with different seeds, resulting in 30K images to form the GT distribution of high-quality CAD-style objects.

Comparing methods. In addition to Instant3D (Li et al., 2023a) and MVDream (Shi et al., 2023b) as direct text-to-multi-view (T2MV) methods, we also adopt edge-cutting single image to multi-view (I2MV) methods CRM (Wang et al., 2024b), SV3D (Voleti et al., 2024) and Zero123++(Shi et al., 2023a). For these methods, we condition the diffusion model on the single view image generated by PixArt (prompted to generate CAD-style single object-centric image). The result of the CLIP score is 3 times averaged with different seeds. For FID, we use 3 different seeds for each of the 3K prompts to generate 9K images to test the distance with GT high-quality images.

4.2 Evaluation of Multi-view Images

As illustrated in Tab.1, compared to other methods, the T2MV diffusion model trained by our framework yields the best results both according to image-text alignment and image quality. For qualitative experiments, we visualize some of the comparisons with other edge-cutting multi-view diffusion model in Fig.6. For these image-to-multi-view models, we condition them on the top-left image generated by Bootstrap3D. Compared to these models trained solely on Objaverse Deitke et al. (2023), our model demonstrates superior generalizability when the image domain is beyond the domain of Objaverse. Since it is difficult to directly measure view consistency as there is no ground truth 3D object for text-to-3D generation. we evaluate the view consistency by synthesizing 3D objects through large reconstruction model in the following experiments. Qualitative results of real user cases are in Sup. A.1.

4.3 Evaluation of Generated 3D Objects

Table 2: **Benchmark of CLIP and FID score of generated 3D objects** based on rendered 9 view images. *MVDream is tested on 200 generated objects for FID test using SDS (Shi et al., 2023b), other methods are tested on 1000 objects using GRM (Xu et al., 2024b) and InstantMesh (Xu et al., 2024a) as sparse view reconstruction model.

Reconstruction	Method	CLIP-R Score ↑		CLIP Score ↑		FID ↓	
		CLIP-L/14	CLIP-bigG	CLIP-L/14	CLIP-bigG	PG2.5	PixArt
SDS	MVDream*	85.2	90.8	**26.1**	39.4	57.8	56.7
GRM	Instant3D (unofficial)	81.7	89.4	24.8	37.1	85.4	80.3
	SV3D	74.1	82.8	23.4	34.1	68.4	69.1
	Zero123++	71.2	80.3	22.3	34.5	69.3	72.4
	Bootstrap3D	86.3	91.6	25.9	**39.7**	**51.2**	**50.7**
InstantMesh	Zero123++	73.2	84.1	23.0	37.2	82.3	88.8
	Bootstrap3D	**87.1**	**92.0**	26.0	39.2	61.2	55.3

View consistency is another crucial factor in reconstructing reasonable 3D objects. Miss alignment between different views can lead to blurred areas in reconstructed objects by large reconstruction model (Hong et al., 2023; Wei et al., 2024). This misalignment causes a significant deterioration in quality, resulting in a notable increase in metrics like FID. To assess the view consistency directly on 3D object, we employ GRM (Xu et al., 2024b) and InstantMesh (Xu et al., 2024a) to reconstruct the object given sparse view images generated in Sec. 4.2. We render 9 view images evenly in orbit for each object and evaluate the image-text alignment and image quality. As reported in Tab. 2. Bootstrap3D, after conditioning GRM or InstantMesh on 4 view images, can generate the best 3D objects both according to image-text alignment and image quality. GPT-4V based human-aligned evaluation based on GPTeval3D (Wu et al., 2024) is supplied in Sup. A.6.

We also present visualizations of some results in Fig.5. Bootstrap3D can generate objects with higher quality and prompt following ability. For other methods, as shown in the first column of Fig.5, although the first image may be well aligned with the given text prompt, the final 3D object may be compromised due to the limitations of its poor generalizability as they are also fine-tuned on Objaverse (Deitke et al., 2023) only. More visualizations and discussions of this are in Sup. A.2

4.4 Ablation Study

Training Timestep Reschedule (TTR) is proposed in 3.3 to better integrate different types of data. The training time step of synthetic data is restricted in $[T, 1000]$, where T is a hyper-parameter to be set in training. We demonstrate the effect of the time-step limit in Fig.7, where the bar in the middle is the value of T. When T is large, namely synthetic data won't affect more time-step at the end of the denoising process, Synthetic data has less influence on the denoising process towards the end, which leads to better view consistency but lower prompt-following ability. Conversely, if T is small, the denoised result better follows the given text prompt but blurring becomes much more severe. In summary, there is a trade-off in injecting synthetic data into the training process: better image-text alignment comes at the cost of worse view consistency and increased blurring. Ultimately, we set $T = 200$ based on empirical study.

Synthetic data and dense captioning are proposed in our work to achieve high-quality images and better image-text alignment. We ablate their effects and the importance of data quantity in Tab. 3. Direct use of synthetic data without Training Timestep Reschedule (TTR) can cause severe blurs and deformation in final outcome. With the help of TTR, the mixture of data can not only improve image-text alignment but also maintain view consistency. Replacing Cap3D (Luo et al., 2024)'s caption

Table 3: **Ablation study of proposed components and quantity of synthetic data.** with CLIP-R Score represents image-text alignment and FID represents image quality.

Methods	Multi-view Image		Generated Object	
	CLIP-R Score	FID PG-2.5	CLIP-R Score	FID PG-2.5
Instant3D (unofficial)	83.6	83.2	81.7	85.4
Cap3D only	77.9	101.3	74.6	120.4
Cap3D + Synthetic Image (100k) w/o TTR	81.5	92.0	71.2	134.6
Cap3D + Synthetic Image (100k) w/ TTR	83.3	60.8	80.2	70.6
Dense recaption + Synthetic Image (100k)	87.4	50.2	85.1	50.9
Dense recaption + Synthetic Image (500k)	88.8	42.4	86.3	51.2

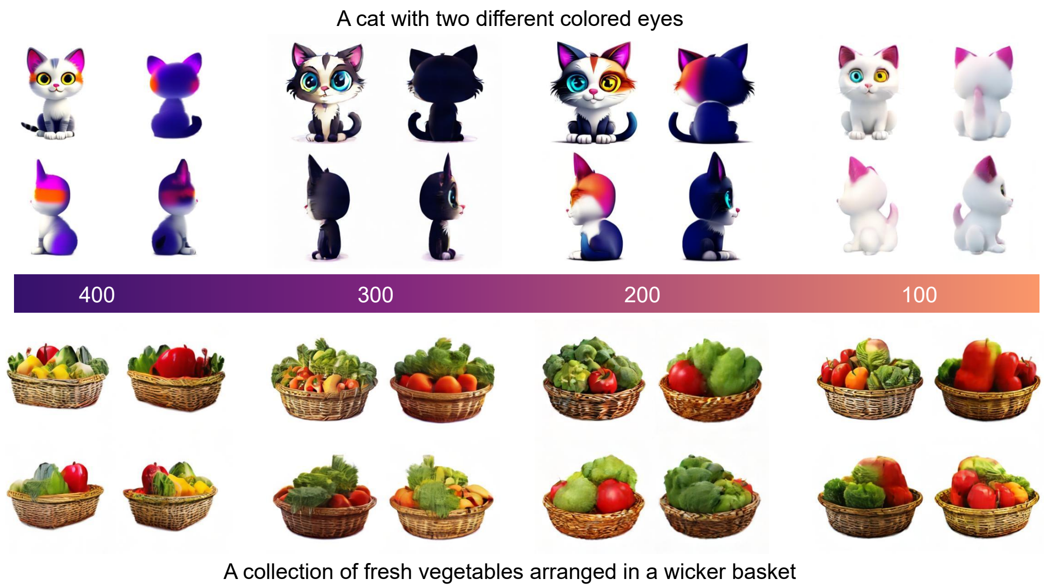


Figure 7: **Ablation study of training time reschedule (TTR)** demonstrates a trade-off between image-text alignment and image quality with different t.

with MV-LLaVA's dense descriptive caption further improves the model's capability of following prompts faithfully. Improvement through increasing volume of data also proves the scalability of our framework.

5 Conclusion and Discussion

In this work, we introduce a novel framework that employs MLLMs and diffusion models to synthesize high-quality data for bootstrapping multi-view diffusion models. With a powerful fine-tuned 3D-aware MLLM serving as the dense captioner and quality filter, the generated synthetic data addresses the issue of insufficient high-quality 3D data. The proposed strategy of injecting different data at different training time steps uses the property of the denoising process to further achieve higher image quality while maintaining view consistency. We believe this work will contribute to the goal of achieving 3D content creation with each rendered view comparable with the single view diffusion model, with more advanced MLLMs and diffusion models on the horizon.

Limitations and future work. Despite its promise, our work still faces several unresolved challenges. Firstly, the multi-view diffusion model is only the first step of the 3D content creation pipeline. Sparse view reconstruction models also need improvement as most edge-cutting sparse view reconstruction models are also trained on Objaverse Deitke et al. (2023) only. Secondly, Although MLLMs can estimate general quality and view consistency, subtle view inconsistency is hard to detect until ambiguity leads to blurred areas in reconstructed 3D object. While the proposed Training Timestep Reschedule can mitigate this problem, it cannot solve the problem fundamentally. Using synthetic data to train sparse view reconstruction models and quality estimation directly based on the reconstructed object are thus interesting future directions for improving 3D content creation.

6 Ethics Statement

Our training code is modified based on public available repository https://github.com/PixArt-alpha/PixArt-alpha. Part of training data are synthesized by our proposed data generation pipeline. For other part of original Objaverse Deitke et al. (2023) data, we only use Cap3D Luo et al. (2024) filtered assets (Objects with CC BY-NC-SA and CC BY-NC licenses are removed, while we retain those with CC-BY 4.0, CC BY-SA, and CC0 licenses) and with face recognizable objects filtered through MSFW classifier and face detector. The ethical filtering in Cap3D make our work using only data without ethics problem. For our synthetic new data, We will launch both the generated captions for Objaverse Deitke et al. (2023) and high-quality synthetic data, model checkpoints and codes with CC-BY 4.0 license for the research community.

7 Reproducibility Statement

Main experimental setting/details (training data, hyperparameters, optimizer, evaluation settings, etc) are clearly presents in Sec. 4.1. For main results, we detail the full test settings in Sec. 4.1. For GPT-4V OpenAI (2023a) based preference study, we provide detailed test prompts and test settings in Sup. A.6. Readers can easily follow the same settings and reproduce all of our experiment results. We provide code for generating synthetic data. Both codes for training the model and testing are also available in supplementary material. The full data and model checkpoints are too large to provide public link without violation of double-blinding. We will release full data and model checkpoints after review.

References

Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120: 153–168, 2016.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198, 2022. URL https://api.semanticscholar.org/CorpusID:248476411.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Tachard Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Z. Chen, Eric Chu, J. Clark, Laurent El Shafey, Yanping Huang, Kathleen S. Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Michael Brooks, Michele Catasta, Yongzhou Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, C Crépy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, M. C. D'iaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fan Feng, Vlad Fienber, Markus Freitag, Xavier García, Sebastian Gehrmann, Lucas González, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, An Ren Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wen Hao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Mu-Li Li, Wei Li, Yaguang Li, Jun Yu Li, Hyeontaek Lim, Han Lin, Zhong-Zhong Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alexandra Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Marie Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniela Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Ke Xu, Yunhan Xu, Lin Wu Xue, Pengcheng Yin, Jiahui Yu, Qiaoling Zhang, Steven Zheng, Ce Zheng, Wei Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report. *ArXiv*, abs/2305.10403, 2023. URL https://api.semanticscholar.org/CorpusID:258740735.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *ArXiv*, abs/2308.01390, 2023. URL https://api.semanticscholar.org/CorpusID:261043320.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3), 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. URL https://api.semanticscholar.org/CorpusID:218971783.

Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16123–16133, 2022.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023a.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024a.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023b.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22246–22256, 2023c.

Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738*, 2024b.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113, 2022. URL `https://api.semanticscholar.org/CorpusID:247951931`.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13142–13153, 2023.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model, 2024.

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2553–2560. IEEE, 2022.

Danny Driess, F. Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Ho Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Peter R. Florence. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, 2023. URL `https://api.semanticscholar.org/CorpusID:257364842`.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.

Daniel Gatis etc. Rembg. `https://github.com/danielgatis/rembg`, 2020.

Ye Fang, Zeyi Sun, Tong Wu, Jiaqi Wang, Ziwei Liu, Gordon Wetzstein, and Dahua Lin. Make-it-real: Unleashing large multimodal model's ability for painting 3d objects with realistic materials, 2024.

Junlin Han, Filippos Kokkinos, and Philip Torr. Vfusion3d: Learning scalable 3d generative models from video diffusion models. *arXiv preprint arXiv:2403.12034*, 2024.

Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2328–2337, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and L. Sifre. Training compute-optimal large language models. *ArXiv*, abs/2203.15556, 2022. URL https://api.semanticscholar.org/CorpusID:247778764.

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 preprint arXiv:2311.04400*, 2023.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models. *ArXiv*, abs/2302.14045, 2023. URL https://api.semanticscholar.org/CorpusID:257219775.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip. https://github.com/mlfoundations/open_clip, 2021.

Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023.

Yash Kant, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, Igor Gilitschenski, and Aliaksandr Siarohin. Spad: Spatially aware multiview diffusers. *arXiv preprint arXiv:2402.05235*, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Jeong-gi Kwak, Erqun Dong, Yuhe Jin, Hanseok Ko, Shweta Mahajan, and Kwang Moo Yi. Vivid-1-to-3: Novel view synthesis with video diffusion models. *arXiv preprint arXiv:2312.01305*, 2023.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024.

Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:246411402.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023b. URL https://api.semanticscholar.org/CorpusID:256390509.

Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. *arXiv preprint arXiv:2311.11284*, 2023.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 300–309, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.

Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. *arXiv preprint arXiv:2311.07885*, 2023a.

Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024b.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9298–9309, 2023b.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023c.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, 2023.

Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pre-trained models. *Advances in Neural Information Processing Systems*, 36, 2024.

Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, Natalia Neverova, Andrea Vedaldi, Oran Gafni, and Filippos Kokkinos. Im-3d: Iterative multiview diffusion and reconstruction for high-quality 3d generation. *arXiv preprint arXiv:2402.08682*, 2024.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

OpenAI. Gpt-4v(ision) system card. *OpenAI*, 2023a. URL https://api.semanticscholar.org/CorpusID:263218031.

R OpenAI. Gpt-4 technical report. *arXiv*, pp. 2303–08774, 2023b.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Zhangyang Qi, Yunhan Yang, Mengchen Zhang, Long Xing, Xiaoyang Wu, Tong Wu, Dahua Lin, Xihui Liu, Jiaqi Wang, and Hengshuang Zhao. Tailor3d: Customized 3d assets editing and generation with dual-side images. *arXiv preprint arXiv:2407.06191*, 2024.

Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. *arXiv preprint arXiv:2311.16918*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Prafull Sharma, Varun Jampani, Yuanzhen Li, Xuhui Jia, Dmitry Lagun, Fredo Durand, William T Freeman, and Mark Matthews. Alchemist: Parametric control of material properties with diffusion models. *arXiv preprint arXiv:2312.02970*, 2023.

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023a.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023b.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want, 2023.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024a.

Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. Mvdiffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. *arXiv preprint arXiv:2402.12712*, 2024b.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023. URL https://api.semanticscholar.org/CorpusID:257219404.

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*, 2024.

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12619–12629, 2023a.

Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023.

Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction, 2023b.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024a.

Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. *arXiv preprint arXiv:2403.05034*, 2024b.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. *arXiv preprint arXiv:2312.03641*, 2023c.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Meshlrm: Large reconstruction model for high-quality mesh, 2024.

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 803–814, 2023.

Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. *arXiv preprint arXiv:2401.04092*, 2024.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1912–1920, 2015.

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 preprint arXiv:2404.07191*, 2024a.

Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*, 2024b.

Fan Yang, Jianfeng Zhang, Yichun Shi, Bowen Chen, Chenxu Zhang, Huichao Zhang, Xiaofeng Yang, Jiashi Feng, and Guosheng Lin. Magic-boost: Boost 3d generation with mutli-view conditioned diffusion. *arXiv preprint arXiv:2404.06429*, 2024.

Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1790–1799, 2020.

Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–16, 2023a.

Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition, 2023b.

Xin-Yang Zheng, Hao Pan, Yu-Xiao Guo, Xin Tong, and Yang Liu. Mvd^2: Efficient multiview 3d reconstruction for multiview diffusion. *arXiv preprint arXiv:2402.14253*, 2024.

A APPENDIX

A.1 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 (Li et al., 2023a). 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. A.6). Results and some qualitative cases are shown in Tab. 4 and Fig. 8. We highlight that our Bootstrap3D excels Instant3D (Li et al., 2023a) when tested on real user prompts through training on synthetic data.

Figure 8: **Real user prompt cases** visualization compared to Instant3D (Li et al., 2023a)

A.2 MORE VISUALIZATION COMPARED TO OTHER METHODS.

We show more visualization of the quantitative experiments shown in the main paper in Fig.9

Table 4: **Test results of in the wild cases.** Bootstrap3D also excels Instant3D (Li et al., 2023a) in generating high quality images according to real user prompts.

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%

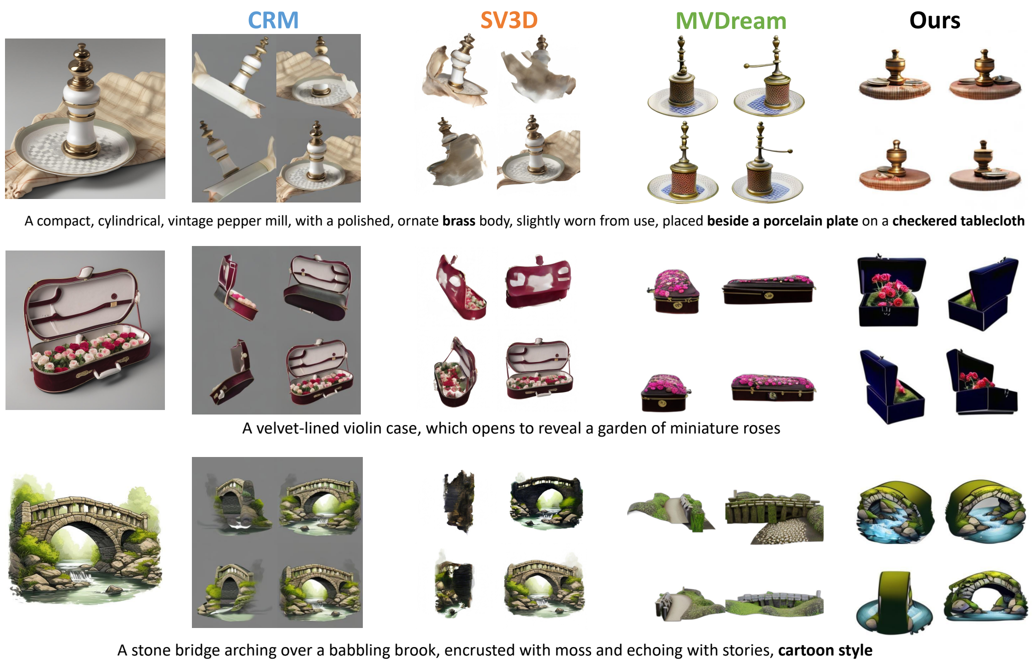


Figure 9: **Generated multivew images compare to other methods.** Our method can generate multi-view images with long text control without encountering blurring effect from data generated by SV3D thanks to TTR and quality filtering.

For Image-to-3D methods, they can sometimes produces significant motion blurring and fails when the input image is out-of-distribution (like the 3rd cartoon style case). We resample 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 clear results. In addition, we use a caption rewriting method, enabling finer prompt control for the generated multi-view images.

A.3 Data Statistics

A.3.1 Caption Analysis

Fig. 10 and 11 provide a visualization of the root noun-verb pairs for the captions generated by GPT-4V (OpenAI, 2023a) 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 (Fang et al., 2024) and stylization (Sharma et al., 2023) 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. 12 visualizes the histogram of caption length compared with Cap3D (Luo et al., 2024). We fine-tune MV-LLaVA to generate two different lengths suitable for different diffusion architecture, namely CLIP-based text encoding (Blattmann et al., 2023; Podell et al., 2023) with 77 token length and T5 based text encoding (Chen et al., 2023a; 2024a) with 120 token length. Both excel the length of Cap3D with less hallucinations.

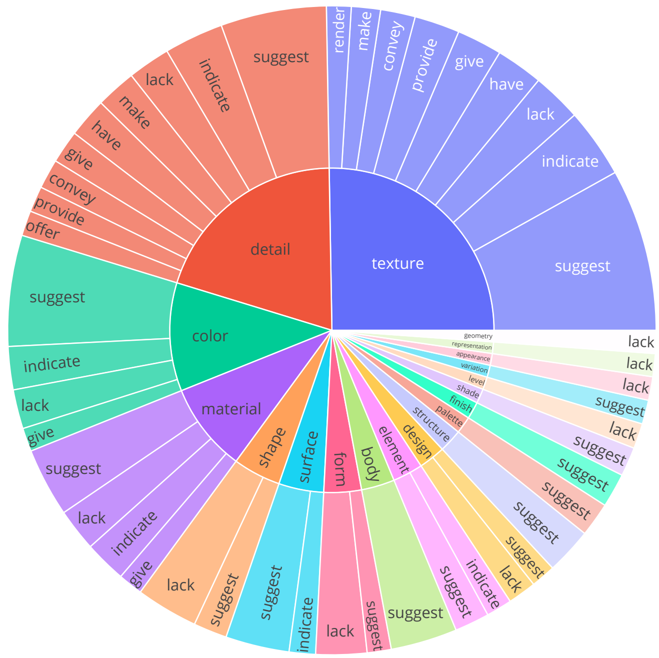


Figure 10: **Visualized analysis of dense reasoning descriptions generated by GPT4-Vision (OpenAI, 2023a)** of the root noun-verb pairs (occurring over 1%) of the descriptions

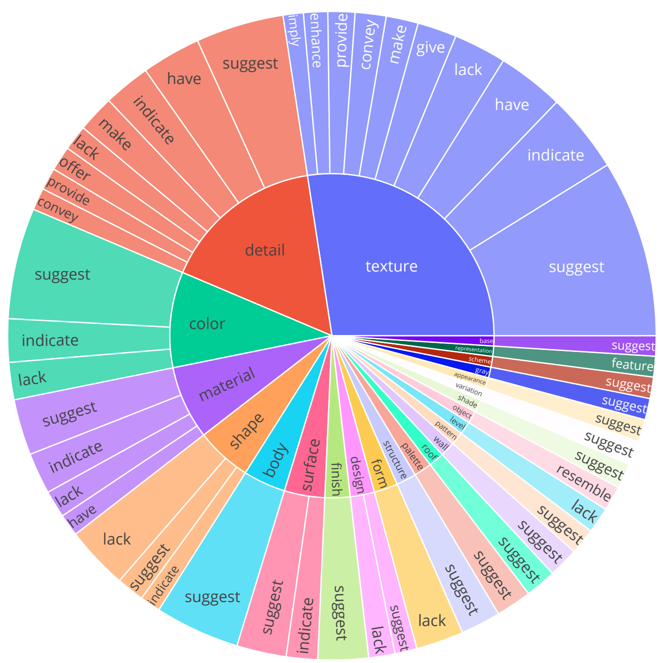


Figure 11: **Visualized analysis of dense reasoning descriptions generated by our MV-LLaVA** of the root noun-verb pairs (occurring over 1%) of the descriptions

A.3.2 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

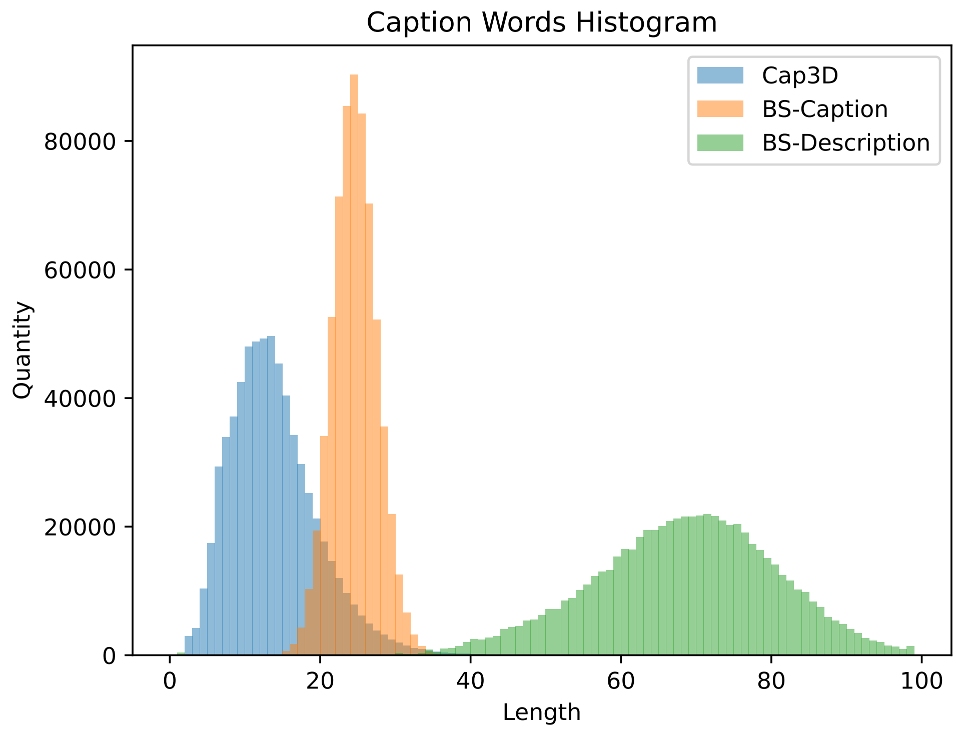


Figure 12: **Histogram Visualization of the Caption Length** compared with Cap3D (Luo et al., 2024)

Table 5: **Comparison of lexical composition of the captions** generated by GPT4-Vision and Share-Captioner.

Lexical	n.	adj.	adv.	v.	num.	prep.
GPT-4V (OpenAI, 2023a)	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%

Fig. 16, Fig. 17 and Fig. 18. We use high quality data with score 4 and 5 for the training of multi-view diffusion model.

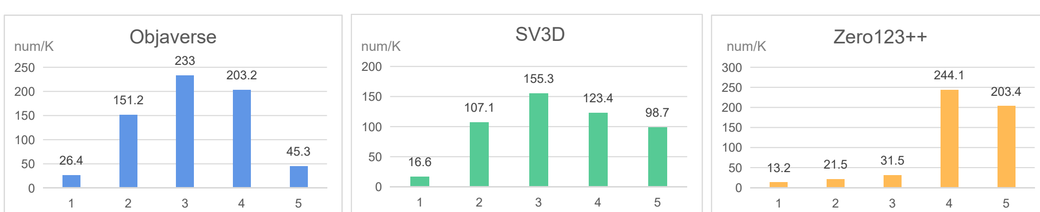


Figure 13: **Quality score statistics of different data source.**

We count the number of multi-view images from different data sources, namely 660K from Objaverse, 500K from SV3D (Voleti et al., 2024) and 500K from Zero123++ (Shi et al., 2023a) generated by our Bootstrap3D pipeline. Result are shown in Fig.13. 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.

A.4 QUALITY OF MV-LLAVA

A.4.1 CHOICE OF NUMBER OF UNFROZEN LAYERS OF VISION ENCODER.

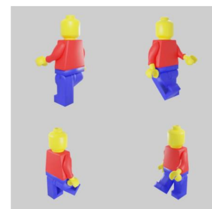
input image				
unfreeze 0 layers	The model displays a basic medieval **helmet** in matte olivegreen with a visor pattern and nasal guard, suitable forsimple visualizations or game assets with a fantasy style. Tag: [single object] [Cartoon]	A cartoonish **turtle** model with a smooth, creamy yellow bodylarge eyes, and a friendly expression, featuring softshading and a glossy finish for a playful appearance. Tag: [single object] [Cartoon]	A detailed classic wingback chair model, with **realistic darkcharcoal fabric** and weathered brown wooden legs and frameconveying an authentic and elegant aesthetic. Tag: [single object] [Photo realistic]	Stylized humanoid figure with a playful design, featuring ayellow head and hands, red torso, blue legs, **and a smilingface**, with a glossy finish for animation or gaming. Tag: [single object][Cartoon]
unfreeze 8 layers	Stylized **humanoid head** with a green, matte finish, featuringa 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 auniform cream color and simple black eyes, suggesting abasic 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 ashiny, curved wooden frame, suitable for realistic interiorvisualizations. Tag: [single object][Photo realistic]	A stylized humanoid figure with a glossy yellow head, redtorso, and blue legs, featuring a minimalistic face and smooth surfaces, ideal for animation or game usage. Tag: [single object][Cartoon]

Figure 14: **Qualitative results of unfreeze final layers of CLIP (Radford et al., 2021) vision encoder** compared to original fixed vision encoder setting in LLaVA (Liu et al., 2024a).

Inspired by ShareGPT-4V (Chen et al., 2023b), we unfreeze selected final layers of the CLIP (Liu et al., 2024a) vision encoder during the initial phase of vision language alignment. The CLIP-L/14 model used for LLaVA (Liu et al., 2024a) 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. 14 illustrates the differences between unfreezing eight layers and not unfreezing any (the original training setting of LLaVA (Liu et al., 2024a)). 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.

A.4.2 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 (Deitke et al., 2023) and exclude training data of MV-LLaVA. We use GPT4-V (OpenAI, 2023a) 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. 6, where MV-LLaVA shows comparable captioning ability with powerful GPT4-V (OpenAI, 2023a), 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 (Deitke et al., 2023) or generated by SV3D (Voleti et al., 2024). 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.7. 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 (Voleti et al., 2024) generated multi-view images. This result align with the observation of inevitable motion blurring of SV3D (Voleti et al., 2024). 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.

Table 6: **Human evaluation** on the quality of generated captions from MV-LLaVA vs. GPT4-Vision (OpenAI, 2023a) over 200 validation samples from Objaverse (Deitke et al., 2023).

Preference	GPT4-Vision (OpenAI, 2023a)	MV-LLaVA	Comparable
Percentage	39.5%	34.5%	26.0%

Table 7: **Confusion matrix** of mutli-view images quality estimation.

Objaverse quality check			Synthetic quality check		
	HQ-gt	LQ-gt		HQ-gt	LQ-gt
HQ by model	31.0%	4.5%	HQ by model	34.5%	11.5%
LQ by model	11.0%	53.5%	LQ by model	17.0%	37.0%

A.4.3 QUALITATIVE CAPTION QUALITY STUDY

We selective compare some of the captions generated by Cap3D (Luo et al., 2024) and MV-LLaVA in Fig. 15. Our MV-LLaVA can generate more detailed descriptive captions with less hallucinations.

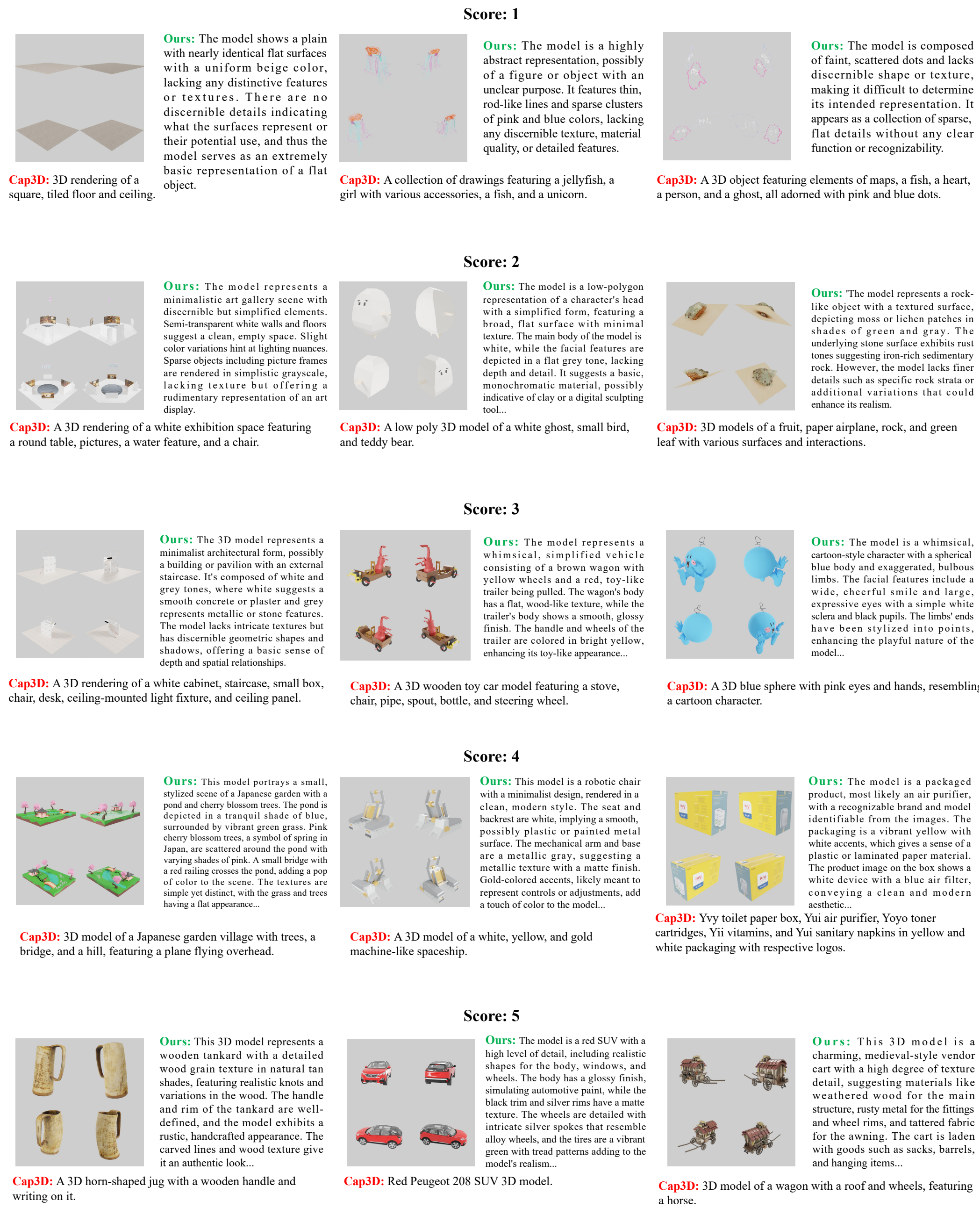


Figure 15: **Caption comparison with Cap3D (Luo et al., 2024).** Our MV-LLaVA can generate long captions that faithfully describing 3D assets from different perspectives like color, geometry and texture.

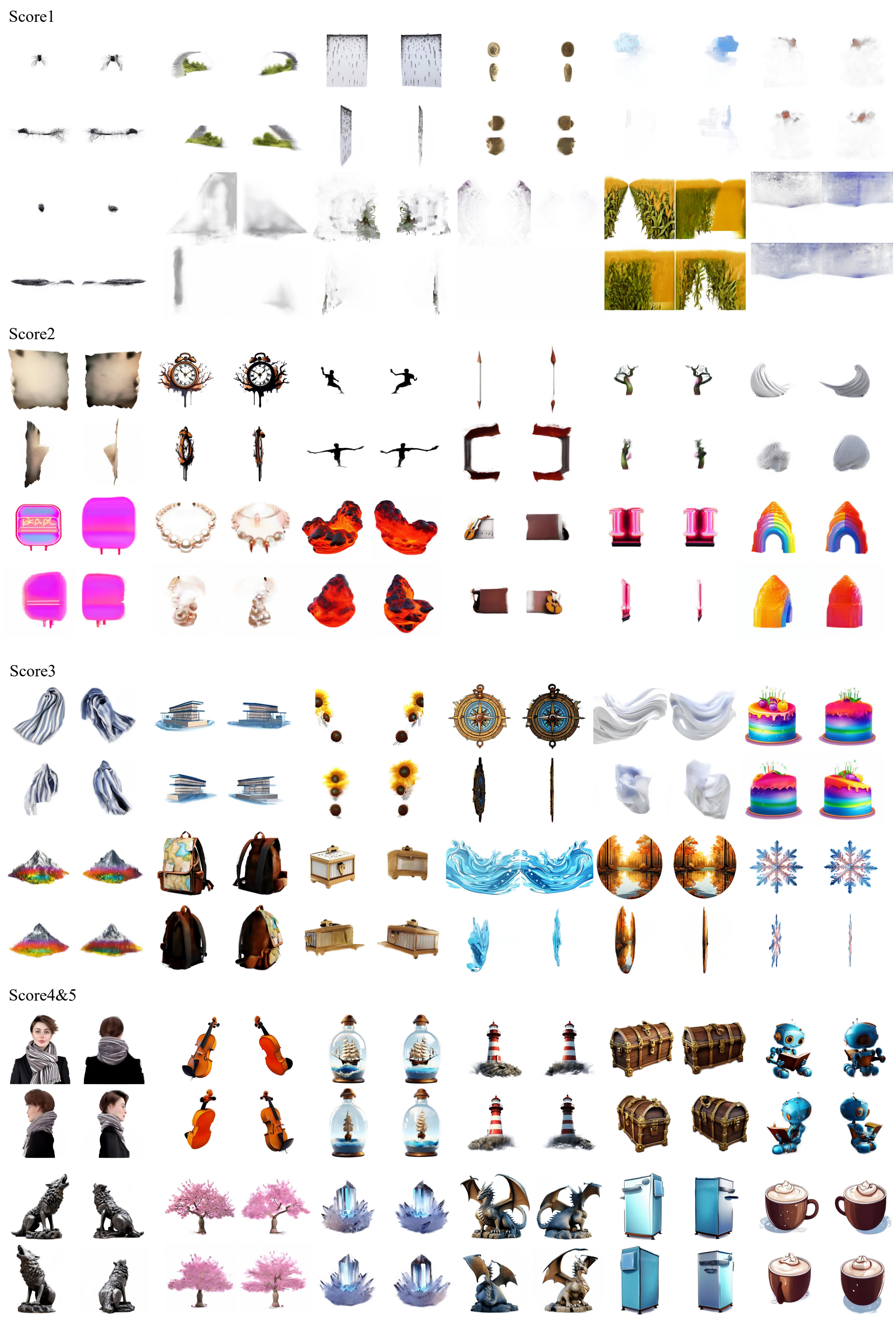


Figure 16: **Randomly picked multi-view images with different scores from 500k synthetic data generated by SV3D (Voleti et al., 2024).**

Figure 17: **Randomly picked multi-view images with different scores from 500k synthetic data generated by Zero123++ (Shi et al., 2023a).**

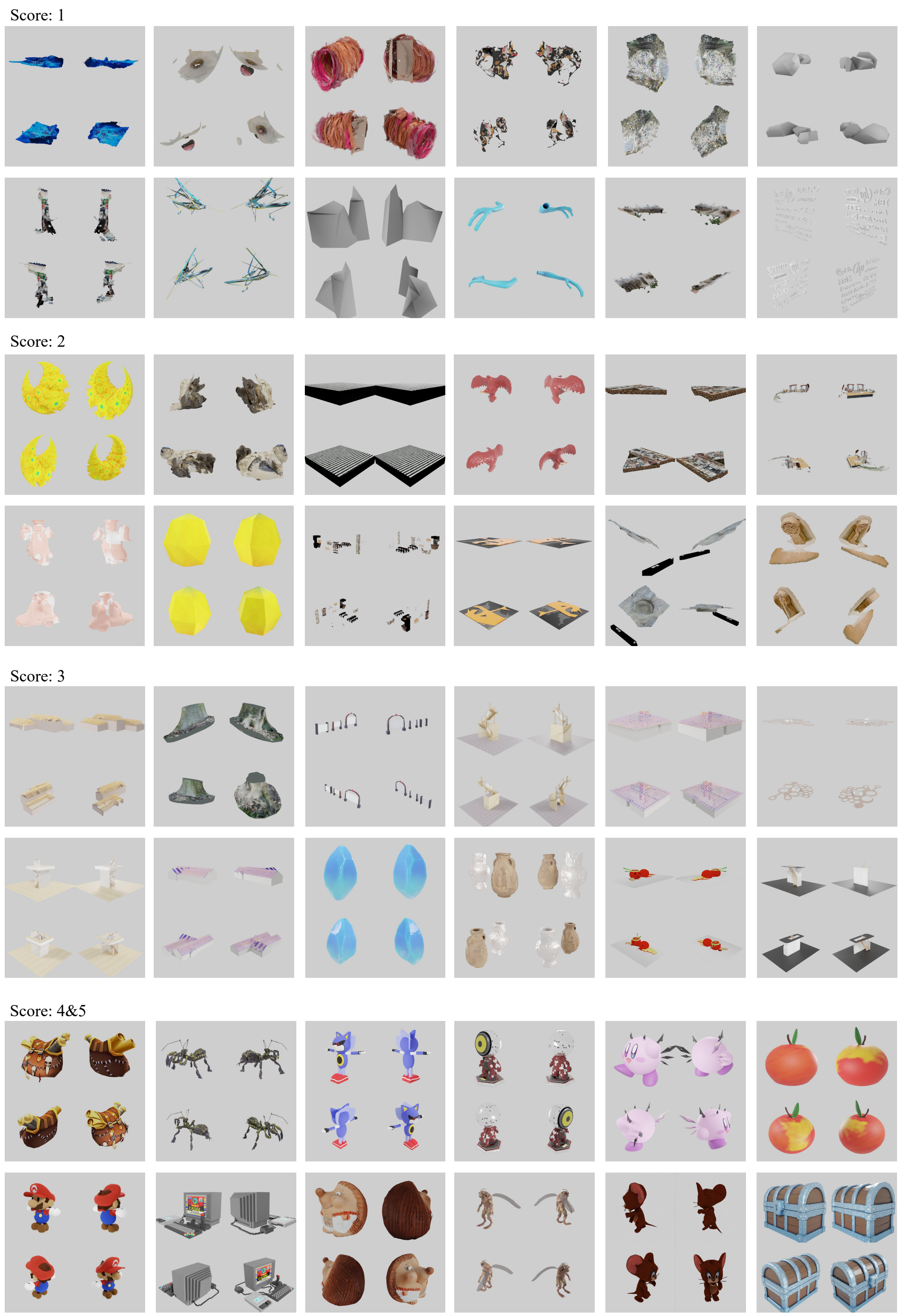


Figure 18: **Randomly picked multi-view images with different scores from 660k Objaverse (Deitke et al., 2023) 3D assets.**

A.5 Details of Prompt Design

A.5.1 Prompts for GPT-4V for Quality Check

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 groud-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 groud-truth image is not obvious. The blurring area indicating rotation or uncerntainty is acceceptable for usage.
4. The novel views are pretty good, although the might be blurring areas or less resolution. the view consistancy 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 shoud 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 discriptive 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 MENSHION 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 groud-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 discribe color and meterial 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. Tage 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 19: **Prompt for GPT-4V to generate caption and estimate quality of multi-view images from SV3D (Voleti et al., 2024), zero123++ (Shi et al., 2023a) and Objaverse (Deitke et al., 2023).**

Detailed prompts are shown in Fig.19.

A.5.2 Prompts for MV-LLaVA Instruct Tuning

Table 8: **Instruct tuning prompt for SV3D (Voleti et al., 2024) and Zero123++ (Shi et al., 2023a) multi-view images**

prompt type	prompt
generate 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?
reasoning	How about the view consistency of this synthesized multi-view image? Do some comments about the view consistency of this synthesized multi-view image. What do you think about the view consistency of this synthesized multi-view image?
quality estimation	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”.

Table 9: **Instruct tuning prompt for Objaverse (Deitke et al., 2023) rendered multi-view images**

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?
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”.

A.6 GPT-4V based 3D Object Generation Evaluation.

We adopt method proposed in GPTeval3D (Wu et al., 2024) for more thorough and human-aligned evaluation of the quality of generated object by different methods. A full test case is shown in Fig. 20. Left 9-view image is rendered from object generated by Bootstrap3D and the right one generated by Instant3D (Li et al., 2023a). 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 given reasonable comparison well

Our task here is the compare two 3D objects, both generated from the same text description. I will provide you with 9 specific view 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 imagin 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 20: **A test conversation with GPT-4V (OpenAI, 2023a)** of evaluating generated objects.

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 (Wu et al., 2024) to test Bootstrap3D generated object comparing with Instant3D (Li et al., 2023a), Zero123++ (Shi et al., 2023a) and MV-Dream (Shi et al., 2023b). 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. 10. Except MVDream (Shi et al., 2023b) (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.

Table 10: **GPT-4V based evaluation result.** the result is in format of ”number of objects preferred geneated by Bootstrap3D/ that of other methods”. Cases when GPT cannot answer the question or generates ”cannot decide” answer are excluded.

	Image-text alignment	Texture quality	3D plausibility
Compared to Instant3D (Li et al., 2023a) (unofficial)	247 / 116	202 / 162	259 / 110
Compared to Zero123++ (Shi et al., 2023a)	192 / 143	210 / 161	231 / 139
Compared to MVDream (Shi et al., 2023b) (GRM)	290 / 71	245 / 131	284 / 102
Compared to MVDream (Shi et al., 2023b) (SDS)	188 / 155	173 / 190	192 / 150

A.7 Improving Direct 3D Generative Models

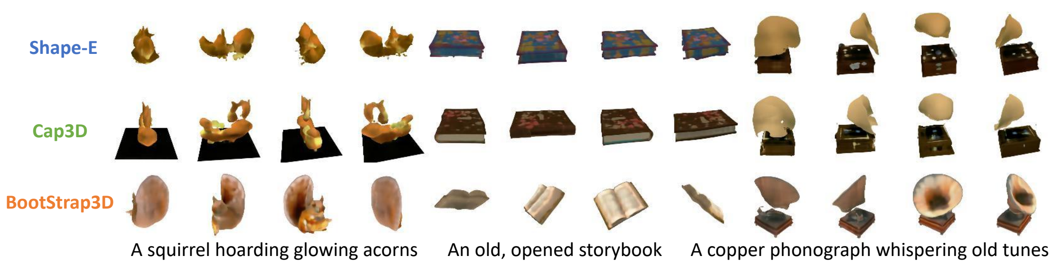


Figure 21: **Fine tuned Shape-E generation results** that shows better object-text alignment than original Shape-E (Jun & Nichol, 2023) and finetuned version in Cap3D (Luo et al., 2024).

Table 11: **Test results on Shape-E.** More accurate and descriptive 3D caption help model to achieve better object-text alignment.

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

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 (Jun & Nichol, 2023) model for experiment and assess the outcomes following the testing method the same to Cap3D (Luo et al., 2024). Specifically, we fine-tune Shape-E using 250K BS-Objaverse data, ensuring that all entries scored greater than 3, accompanied by more precise and descriptive captions. The metrics for training and testing are consistent with those employed in Cap3D (Luo et al., 2024). Some qualitative results are presented in Fig.21, where our finetuned verson can generate object that follow text prompt more precisely. Quantitative results are detailed in Tab.11, where more accurate and desciptive captions than Cap3D can significantly improve metrics like CLIP score. 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.

Figure 22: **Visualization of generated objects compared to other edge-cutting methods**

A.8 More Results Visualization

A.8.1 Comparison with Other Methods

A.8.2 Visualization of Generated objects with Different Styles

A.9 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

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

in the field. The model and data can serve as educational tools for students 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.

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

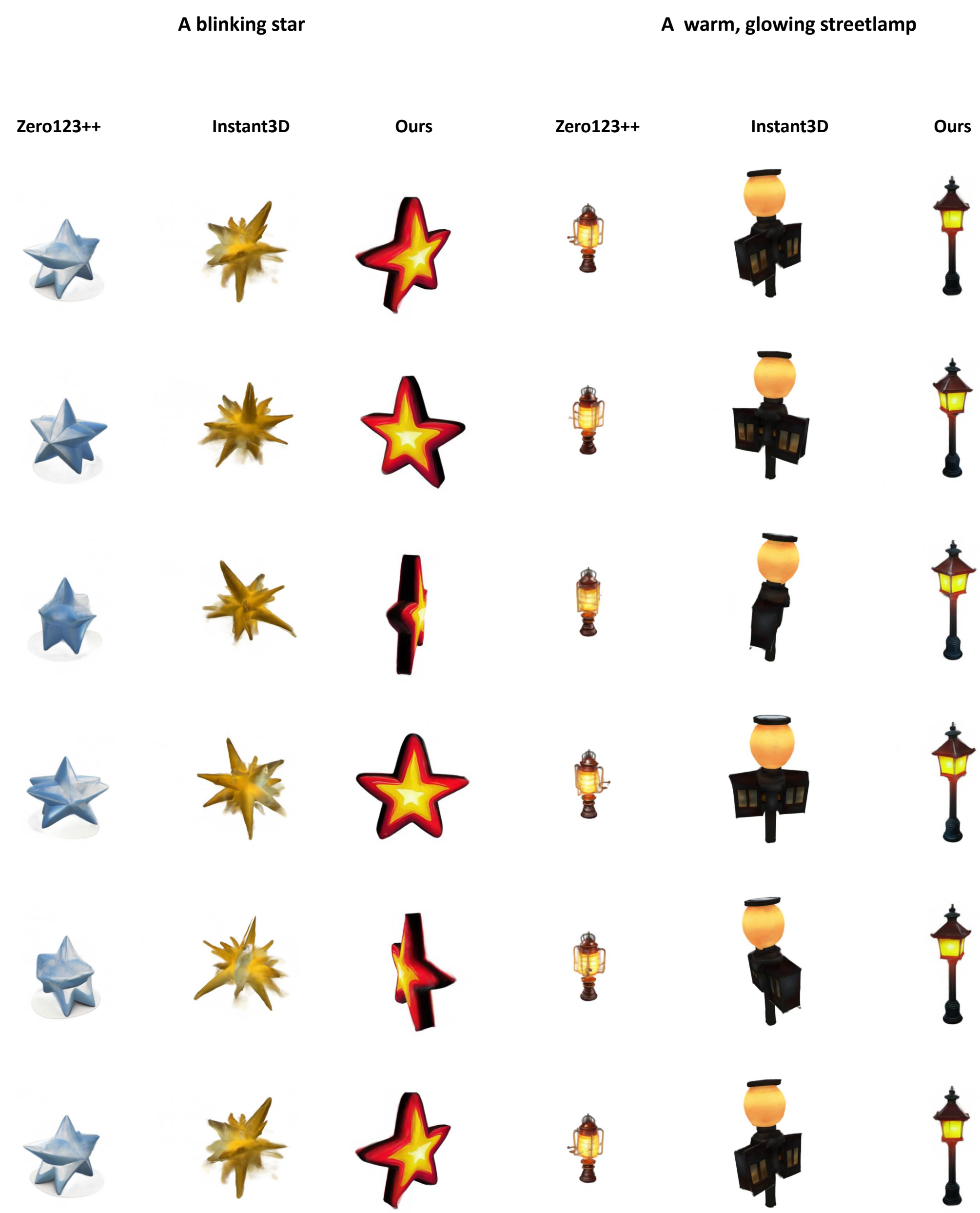


Figure 25: **Visualization of generated objects compared to other edge-cutting methods**

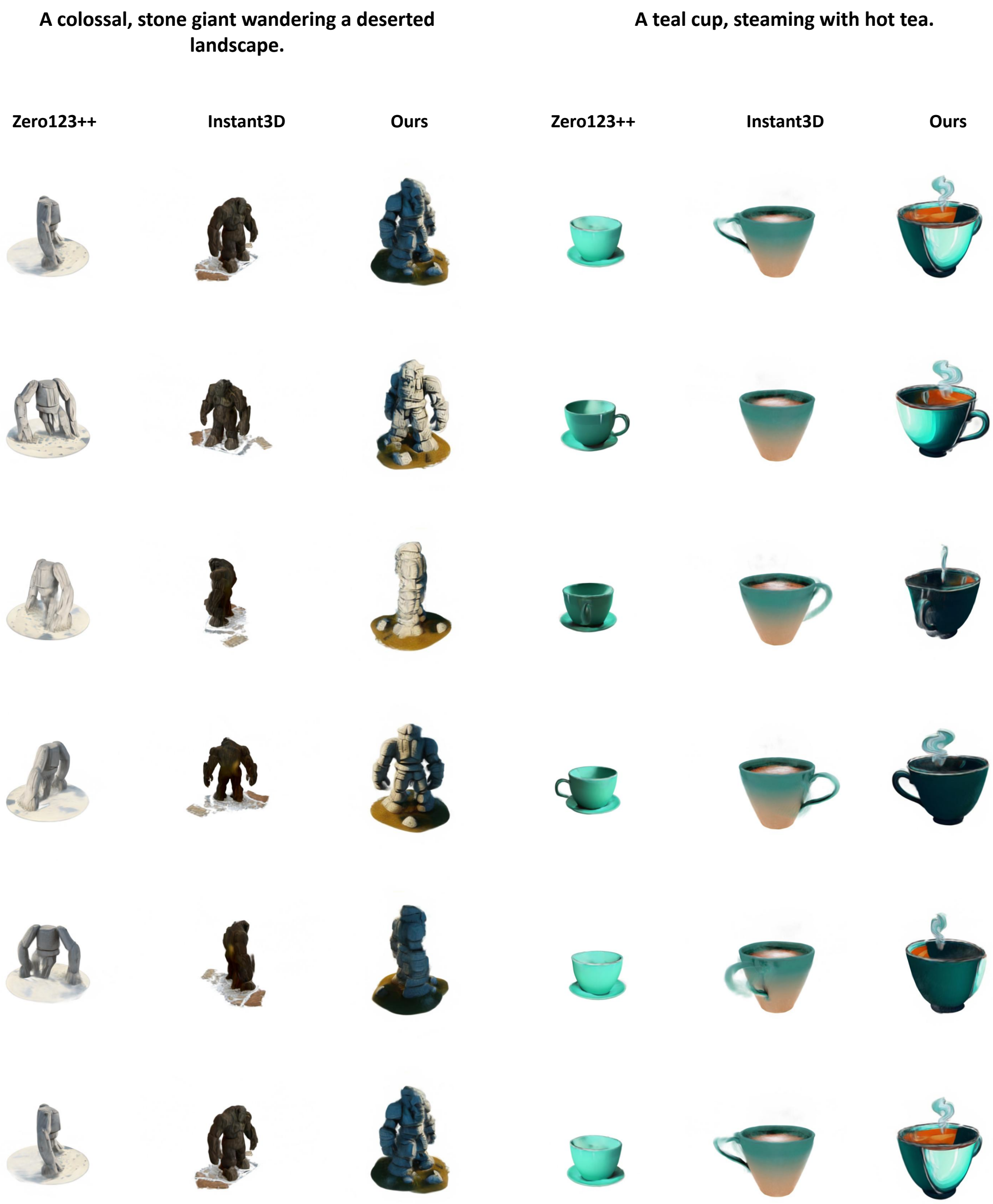


Figure 26: **Visualization of generated objects compared to other edge-cutting methods**

Figure 27: **Visualization of generated objects compared to other edge-cutting methods with different style control.**

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.**