

3DGen-Bench: Comprehensive Benchmark Suite for 3D Generative Models

Yuhan Zhang^{1,3*} **Mengchen Zhang^{2,3*}** **Tong Wu^{4†}** **Tengfei Wang³**
Gordon Wetzstein⁴ **Dahua Lin⁵** **Ziwei Liu^{6†}**

¹Fudan University ²Zhejiang University ³Shanghai Artificial Intelligence Laboratory

⁴Stanford University ⁵The Chinese University of Hong Kong

⁶S-Lab, Nanyang Technological University .

*These authors contributed equally to this work. †Corresponding author.

Abstract

3D generation is experiencing rapid advancements, while the development of 3D evaluation has not kept pace. How to keep automatic evaluation equitably aligned with human perception has become a well-recognized challenge. Recent advances in the field of language and image generation have explored human preferences and showcased respectable fitting ability. However, the 3D domain still lacks such a comprehensive preference dataset over generative models. To mitigate this absence, we develop *3DGen-Arena*, an integrated platform in a battle manner. Then, we carefully design diverse text and image prompts and leverage the arena platform to gather human preferences from both public users and expert annotators, resulting in a large-scale multi-dimension human preference dataset *3DGen-Bench*. Using this dataset, we further train a CLIP-based scoring model, *3DGen-Score*, and a MLLM-based automatic evaluator, *3DGen-Eval*. These two models innovatively unify the quality evaluation of text-to-3D and image-to-3D generation, and jointly form our automated evaluation system with their respective strengths. Extensive experiments demonstrate the efficacy of our scoring model in predicting human preferences, exhibiting a superior correlation with human ranks compared to existing metrics. We believe that our *3DGen-Bench* dataset and automated evaluation system will foster a more equitable evaluation in the field of 3D generation, further promoting the development of 3D generative models and their downstream applications.

Keywords: 3D Evaluation, 3D Generation, Human Preference, Leaderboard

1 Introduction

3D content generation has emerged as a pivotal research area with diverse applications across virtual reality, video games, films, and robotics. In recent years, the field has undergone a transformative revolution fueled by the extraordinary generative capabilities of diffusion models and the availability of large-scale 3D datasets. This has

facilitated an explosion of 3D generative models (R. Chen, Chen, Jiao, & Jia, 2023; Y. Chen et al., 2024; Z. Chen, Wang, & Liu, 2023; Cheng, Lee, Tuyakov, Schwing, & Gui, 2023; Di et al., 2024; Gao* et al., 2024; Gupta, Xiong, Nie, Jones, & Oğuz, 2023; F. Hong et al., 2024; Jun & Nichol, 2023; Kong et al., 2024; W. Li, Chen, Chen, & Tan, 2023; Lin et al., 2023; F. Liu, Wu, Wei, Rao, & Duan, 2023; M. Liu et al., 2023; M. Liu, Xu, et al., 2024; R. Liu et al., 2023; Y. Liu et

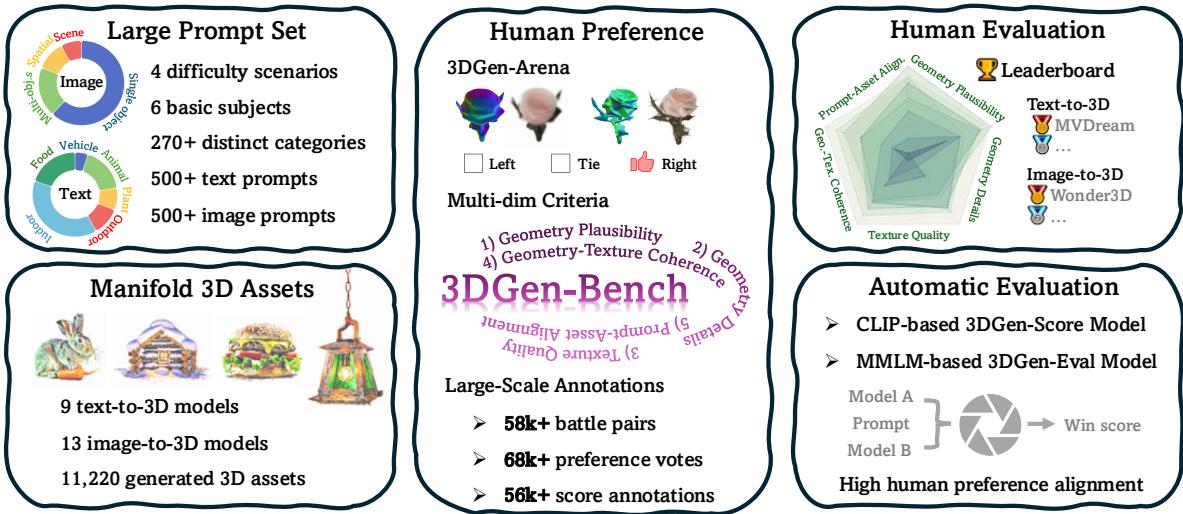


Fig. 1 Overview of 3DGen-Bench, which is the first comprehensive human preference dataset for 3D models. For efficient data collection, we build the **3DGen-Arena** platform in a pairwise battle manner. Based on the annotated data, we perform a comprehensive evaluation for state-of-the-art 3D generative models, and propose two scoring model, **3DGen-Score** and **3DGen-Eval** models, regarded as automated 3D evaluators that align well with human judgments.

al., 2023; Long et al., 2023; Ma et al., 2023; Melas-Kyriazi et al., 2024; Metzer, Richardson, Patashnik, Giryes, & Cohen-Or, 2023; Nichol, Jun, Dhariwal, Mishkin, & Chen, 2022; Poole, Jain, Barron, & Mildenhall, 2022; Qian et al., 2024; Raj et al., 2023; R. Shi et al., 2023; Y. Shi et al., 2023; J. Sun et al., 2023; Tang et al., 2024; Tang, Ren, Zhou, Liu, & Zeng, 2023; Tang, Wang, et al., 2023; Voleti et al., 2024; H. Wang, Du, Li, Yeh, & Shakhnarovich, 2023; Z. Wang, Lu, et al., 2024; Z. Wang, Wang, Hancke, Liu, & Lau, 2024; T. Wu, Wang, et al., 2023; Y. Xu et al., 2024; C. Zheng & Vedaldi, 2023; Zhu & Zhuang, 2023; Zou et al., 2023) that can craft high-fidelity 3D assets from specific textual or visual prompts.

However, the advancement in 3D evaluation techniques has not kept pace with the increase of 3D generative models. While CLIP similarity (Hessel, Holtzman, Forbes, Bras, & Choi, 2022; Radford et al., 2021) has been widely adopted for measuring text-image alignment, its application in 3D often proves inadequate and inaccurate. This is primarily due to the lack of 3D-specific prior knowledge and its inability to accommodate the versatile evaluation requirements of 3D assets. User-Study provides more reliable assessments, but the absence of a standardized prompt set makes fair comparisons across different works challenging. Recognizing this, T^3 Bench (Y. He

et al., 2023) attempted to establish a more suitable benchmark by manually curating a prompt set designed for text-to-3D generation. However, its metrics for the *Quality Assessment* dimension still rely on the original CLIP (Radford et al., 2021), leaving its inherent limitations unaddressed. Meanwhile, using BLIP (J. Li, Li, Xiong, & Hoi, 2022) as the caption model introduces unnecessary errors in the *Alignment Assessment* dimension. GPTEval3D (T. Wu et al., 2024) explored the potential of prompting GPT-4V(ision), a powerful multimodal large language model, as a versatile and human-aligned 3D evaluator. While it demonstrated notable robustness, the closed-source and black-box nature of GPT-4V raises concerns about potential biases, which remain challenging to interpret.

Existing metrics for 3D assessment lack grounding in real-world human preference data, a principle well-established in 2D image generation benchmarks. Numerous works in image evaluation (Kirstain et al., 2023; Ku et al., 2023; Schuhmann et al., 2022; X. Wu et al., 2023; J. Xu, Liu, et al., 2024) have developed metrics that closely align with human preferences. In contrast, a similar, transparent, user-focused metric for 3D assets is still absent. This highlights the pressing need

for a large-scale dataset that captures human preferences for 3D generation, a crucial step toward aligning 3D assessments with human judgment.

2 Related Work

2.1 3D Generative Models.

Conditional 3D generation creates 3D models from text or image prompts. Empowered by advances in text-to-image diffusion models (Rombach, Blattmann, Lorenz, Esser, & Ommer, 2022; Saharia et al., 2022), optimization-based methods (R. Chen et al., 2023; Y. Chen et al., 2024; W. Li et al., 2023; Lin et al., 2023; F. Liu et al., 2023; Ma et al., 2023; Metzer et al., 2023; Poole et al., 2022; Qian et al., 2024; Y. Shi et al., 2023; Tang, Wang, et al., 2023; H. Wang et al., 2023; Zhu & Zhuang, 2023) have applied Score Distillation Sampling (SDS) (Poole et al., 2022) for conditional 3D generation. (Z. Wang, Lu, et al., 2024) further presented Variational Score Distillation (VSD), which optimized the distribution of 3D scenes instead. Some works (Raj et al., 2023; J. Sun et al., 2023; Z. Wang, Wang, et al., 2024) employ a second-stage DreamBooth-like model fine-tuning for enhanced texture modeling. Recent works (Z. Chen et al., 2023; Di et al., 2024; Tang, Ren, et al., 2023) have incorporated SDS with 3D Gaussian Splatting (Kerbl, Kopanas, Leimkühler, & Drettakis, 2023). In the meantime, feed-forward diffusion models (Cheng et al., 2023; Gupta et al., 2023; F. Hong et al., 2024; Jun & Nichol, 2023; Nichol et al., 2022; T. Wang et al., 2023) also show respectable performance with fast inference. Transformer-based regression models (Z. He & Wang, 2023; Y. Hong et al., 2023; J. Li et al., 2023) train a regression model to predict triplanes directly from images. Another line of work (Gao* et al., 2024; Kong et al., 2024; M. Liu et al., 2023; M. Liu, Xu, et al., 2024; R. Liu et al., 2023; Y. Liu et al., 2023; Long et al., 2023; R. Shi et al., 2023; Voleti et al., 2024; C. Zheng & Vedaldi, 2023) fine-tunes image or video diffusion models to generate novel views for stronger 3D guidance. Some works (Melas-Kyriazi et al., 2024; Tang et al., 2024; Y. Xu et al., 2024; Zou et al., 2023) use 3D Gaussian Splatting (Kerbl et al., 2023) for multi-view reconstruction. With rapid advancements in 3D generative technologies, the need for an objective and comprehensive evaluation system has

become increasingly urgent. In this paper, we propose a comprehensive evaluation framework that encompasses five distinct dimensions informed by the collection of human preference data. Building on this foundation, we further train automated evaluators, achieving exceptional performance in simulating human judgments.

2.2 3D Evaluation Metrics.

Evaluating conditional 3D generative models is challenging due to the need to understand both 3D worlds and given conditions. For text-conditioned generative tasks, CLIP (Radford et al., 2021) has been widely used to assess text-to-3D alignment. For image-conditioned generative tasks, multi-view datasets such as GSO (Downs et al., 2022) and RTMV (Tremblay et al., 2022) are usually used to evaluate single-view reconstruction quality. However, these metrics are constrained by the challenge of acquiring a comprehensive reference set. User studies are adopted by many works (Lin et al., 2023; Y. Shi et al., 2023; Voleti et al., 2024; Z. Wang, Lu, et al., 2024; Y. Xu et al., 2024) but are costly and difficult to scale. Recent efforts have focused on automated text-to-3D evaluation. T^3 Bench (Y. He et al., 2023) introduced two automatic metrics based on multi-view images, but they rely on a hand-designed regional convolution mechanism, raising concerns about their credibility and validity. Wu et al. (T. Wu et al., 2024) proposed leveraging GPT-4V’s visual question-answering capability to evaluate 3D models, but this approach may introduce uninterpretable biases in answer selection. Our framework stands out by incorporating human preference data into a comprehensive five-dimensional evaluation scheme, effectively capturing the subtleties of generated 3D content missed by prior metrics. Leveraging these insights to train automated evaluators, we enhance alignment with human preferences. This hybrid approach bridges the gap between subjective evaluations and automated metrics, providing a more reliable and scalable solution for 3D model assessment.

2.3 Human Preference.

Recent advances in language and image generation have placed significant emphasis on human preference. Learning from human preference not only enables adaptation to complex tasks that may

Table 1 Comparison between 3DGen-Bench dataset and others. Our dataset is the first to incorporate both text and image prompts, as well as the first to implement absolute value scoring annotations, showcasing a substantial advantage in model scale. Note that dimension “public” is for annotations only, and symbol “ \mathbb{N} ” represents “partly public available” as 3DRewardDB (Ye et al., 2024) only release partial annotations of 1000 prompts.

Dataset	Condition		Prompt	Dim	Model	Assert	Annotation		
	text	image					pair	score	public
T^3 Bench	✓	-	300	2	6	-	-	-	-
GPTEval3D	✓	-	110	5	13	1,430	234	-	-
3DRewardDB	✓	-	2,530	4	5	-	25.3k	-	\mathbb{N}
3DGen-Bench(Ours)	✓	✓	1,020	5	22	11,220	13.8k	11.2k	✓

challenge automatic evaluation but also guides models to align with human expectations. Specifically, there are two major downstream scenarios: human evaluation for model ranking (Kirstain et al., 2023; Ku et al., 2023; Lu et al., 2024; mrfakename et al., n.d.; L. Zheng et al., 2023) and human feedback for model tuning (Bai et al., 2022; Ethayarajh, Choi, & Swayamdipta, 2022; Köpf et al., 2024; Nakano et al., 2021; Stiennon et al., 2020; Taori et al., 2023; Y. Wang et al., 2022). LMSYS (L. Zheng et al., 2023) introduced a new preference collection method to mitigate individual bias, making it accessible to the public through an open arena platform. In this paper, we extend human preference learning to the field of 3D evaluation by establishing the first 3D arena. Additionally, we integrate the two key application aspects: on one hand, we establish a 3D leaderboard driven by human evaluations; on the other hand, we develop a reward model for automatic evaluation and generative model optimization.

3 3DGen-Bench Dataset

3.1 Comprehensive Model Zoo for 3D Generation

3.1.1 3D Generative Models Involved

Model Selection. To ensure comprehensive benchmark coverage, we collect a variety of 3D generative models, including both optimization-based and feed-forward approaches, and encompassing various representations such as point clouds, triplanes, NeRF, and 3D Gaussian models. Additionally, we include pioneering works that leverage multi-view diffusion models for novel sparse reconstruction. Ultimately, our benchmark includes 9 text-to-3D generative models (Jun & Nichol, 2023; Liang et al., 2023; Lin et al., 2023;

Metzer et al., 2023; Nichol et al., 2022; Poole et al., 2022; Y. Shi et al., 2023; H. Wang et al., 2023; Y. Xu et al., 2024) and 13 image-to-3D generative models (Z. He & Wang, 2023; Jun & Nichol, 2023; Kong et al., 2024; R. Liu et al., 2023; Y. Liu et al., 2023; Long et al., 2023; Nichol et al., 2022; Qian et al., 2024; Tang et al., 2024; Y. Xu et al., 2024; C. Zheng & Vedaldi, 2023; Zou et al., 2023), as shown in Figure 2. The complete list of the included models is provided in Appendix B.2. This diverse selection ensures that our benchmark encompasses a broad spectrum of 3D generative models, catering to varied research requirements and offering a comprehensive evaluation platform.

Model Implementation. We prioritize using the official code of each model when available; otherwise, we rely on Threestudio’s (Guo et al., 2023) implementation for consistency. For LRM (Y. Hong et al., 2023), we employ the open-sourced implementation OpenLRM (Z. He & Wang, 2023). All experiments were conducted using the default hyperparameters specified in each code. The entire generation process took place over 4 weeks, utilizing 8 NVIDIA A100-SXM4-80GB GPUs.

Render Implementation. For each 3D model, we attempt to obtain surrounding videos in three formats: RGB, normal, and geometry(texture-free geometry). As the 3D models are generated in PLY, OBJ or GLB formats, for models that can directly generate RGB and normal renderings, we capture these directly. Otherwise, we uniformly convert it into mesh representation, and then render it into required videos.

3.1.2 Prompt Generation

We carefully design 1,020 prompts, including 510 texts and 510 images, which is approximately 10× the previous prompt suites (Y. He et al., 2023;

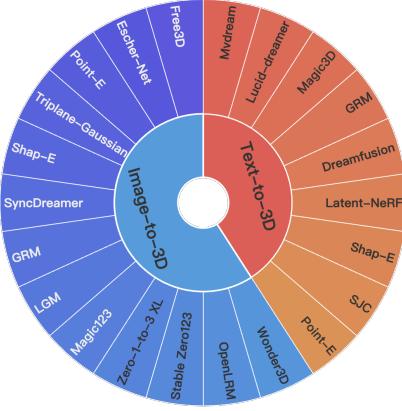


Fig. 2 3D Generative Models. The specific list of generative models we selected for our benchmark.

T. Wu et al., 2024). Our prompt set spans more than 270 distinct categories, derived from 6 basic subjects: “Vehicle”, “Animal”, “Plant”, “Food”, “Indoor objects”, “Outdoor objects”. Furthermore, to evaluate the models’ robustness for complex prompts, we expanded the prompt set to include scenarios with multiple objects and micro-scenes. Ensuring diversity and a balanced distribution of prompts across both text and image generation is crucial. During the prompt generation process, we use predefined proportions to sample categories, attributes, and scenario complexities. The raw prompts are then manually screened to ensure diversity and representativeness. The visualization of prompts can be found in Figure 3. Subsequently, we conducted stratified sampling on the cleaned prompts, splitting them into training, validation, and test sets in a 9 : 1 : 2 ratio. More details on this process can be found in Appendix B.1.

Text Prompt. We employ GPT-4 to generate text prompts heuristically. This process involves deconstructing sentences into four fundamental components: count, category, attribution, and spatial relation. For example, a sentence can be structured to include an explicit count ("three"), a category ("apples"), an attribution ("green"), and a spatial relation ("on the table"). Based on this, we provide GPT-4 with various sentence templates and manually filter the final prompts. More details can be found in Appendix B.1. Ultimately, our text set includes exactly 510 prompts, comprising 43 with explicit counts, 104 with implicit counts (e.g., "a pile of"), 73 featuring spatial relationships, and 290 individual objects.



Fig. 3 Visualization of prompts. Left are two fan diagrams respectively show the distribution of the 6 basic category subjects and 4 main difficulty scenarios (where the "implicit count" and "explicit count" are both grouped into the "Multi-obj.s"). Right is the visualization of word cloud of text prompts.

Image Prompt. We manually collect and curate a set of 510 images, ensuring their quality and diversity through a careful selection process. To prepare them as prompts, we use `rembg` (Qin et al., 2020) to remove backgrounds. As a result, the final image set includes 47 multi-object compositions featuring more than six objects, 49 compositions with a modest number of 2-6 objects, 61 instances of spatial relational compositions, 40 complex scenes, and 313 individual objects.

3.2 Multi-criteria Human Voting Collection

3.2.1 Evaluation Criteria

When evaluating the quality of 3D model generation, we typically consider the following five criteria as outlined in GPTEval3D ([T. Wu et al., 2024](#)), with some slight adjustments:

- 1) **Geometry Plausibility** assesses whether the general shape resemble the real object logically and physically. A plausible 3D asset is expected to have recognizable structures and avoid improbable features, such as distorted faces (Multi-face Janus problem), floating fragments or noisy geometry.
 - 2) **Geometry Details** assesses the fineness and intricacy of shapes. High-detail models exhibit well-defined and intentional surface features, carefully distinguishing them from noise, whereas low-detail models may appear rough and simplistic.
 - 3) **Texture Quality** underscores both aesthetic and consistency. The former includes aspects like realistic material representation and appropriate coloring, while the latter emphasizes uniformity and seamlessness across different viewing angles.

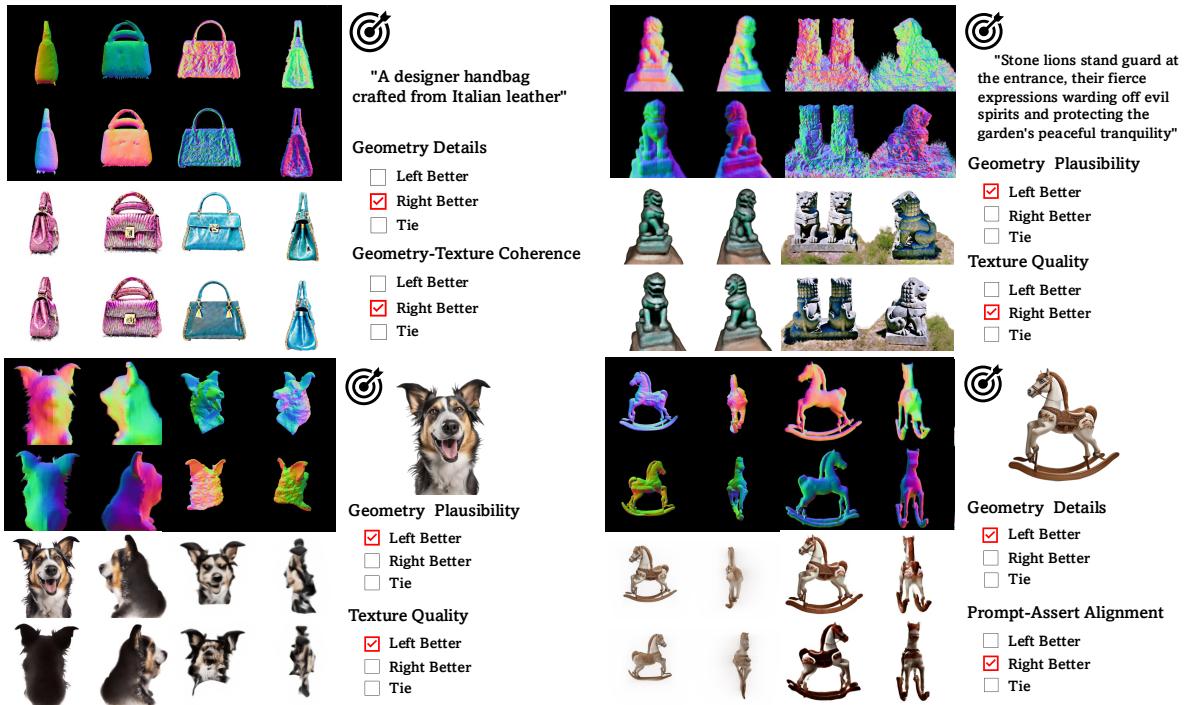


Fig. 4 Examples of evaluation criteria. We try to give an intuitive explanation by means of pairwise comparison, where the top row for text-to-3D generation and the bottom row for image-to-3D generaion. Notice that the five dimensions are mutually independent, each with its own emphasis.

4) Geometry-Texture Coherence evaluates the alignment between geometric and texture characteristics. Textures should conform naturally to the model's contours without distortion and maintain a consistent appearance without compensating for or obscuring geometric details.

5) Prompt-Asset Alignment ensures that generated 3D assets accurately reflect their prompts in terms of category, quantity, attributes, and spatial relationships. For image-to-3D generations, maintaining identity consistency is also essential. Figure 4 illustrates several examples of these criteria in a pairwise comparison style.

3.2.2 3DGen-Arena

Inspired by previous works (Ku et al., 2023; Lu et al., 2024; mrfakename et al., n.d.; L. Zheng et al., 2023), we introduce *3DGen-Arena*, the first arena-style benchmark platform in the 3D field. This platform employs an anonymous pairwise comparison manner, augmented with "named battles" and "direct chat" functionalities to encourage richer interactions. Specifically, visitors can sample prompts while the models in the battle

remain anonymous until all five voting dimensions have been completed. For each dimension, visitors can choose from the options: '*Model A is better*', '*Model B is better*', '*Tie*', '*Both bad*'. Moreover, to facilitate user understanding, instead of simply displaying the generated 3D assets, *3DGem-Arena* presents them through three distinct 360° panoramic videos: normal maps, textureless geometry, and fully textured renderings. The screenshots of user interfaces can be found in Appendix C.1. Additionally, we have received 8,045 anonymous votes from the public to date, with the corresponding leaderboard shown in Appendix D.

3.3 Large-scale Human Annotation

In addition to anonymous voting, we collect a substantial set of reliable labels from expert annotators. Specifically, we meticulously instruct 47 professional annotators with detailed guidelines and closely monitor their performance through regular feedback. Annotators are compensated at a rate of \$6 per hour, resulting in a total expenditure of approximately \$4.700. Additionally, during

the annotation process, they are required to complete two types of annotations, as detailed below:

1) Comparison votes in a battle format, where annotators are asked to choose the better one given two anonymous models generated from the same prompt. **2) Absolute scores** in numerical format, where annotators are asked to provide a reasonable value within a predefined range for each dimension given a single model.

3.3.1 Comparison Votes

Rules. We reuse the same annotation rules as 3DGen-Arena. Given a text/image prompt p_i , we sample two 3D models generated from different methods, (S_0, S_1) , and showcase their 360° panoramic videos $\{(V_{i_geo}, V_{i_normal}, V_{i_rgb}), i \in [0, 1]\}$. Annotators are expected to vote for the five dimensions outlined in Section 3.2.1 in a single route. More details about the annotation platform can be found in Appendix C.2.

Statistics. We randomly sampled 13,680 battle pairs and organized them into 456 packs, each containing 30 pairs. Finally, we collected 114,576 raw votes (the sum of 5 dimensions). Then, we employ three validation strategies simultaneously to clean raw votes: 1) We compute the strong-conflict ratio (when "left vote" meets "right vote") against valid pairs (287 pairs, 2%) labeled by us; 2) We set up cross-annotation packs (240 packs, 52.6%), which are labeled by two isolated annotators, and then compute the strong-conflict ratio between them; 3) We compute the "tie/both bad vote" ratio to identify potential lazy annotators. We revise our data with above three ratios, and ultimately obtain 68,400 unique votes.

3.3.2 Absolute Scores

Rules. Given a text/image prompt p_i , we want to assign a score to the quality of a generated 3D model in absolute terms. Even with highly detailed instruction documentation, it remains challenging to ensure scoring consistency. To mitigate this, we display all 3D models (9 for text, 13 for image) generated from the same prompt simultaneously. Annotators are required to rank these models based on different dimensions before providing scores. We apply three score ranges based on the complexity of distinct dimensions: [0, 9], [0, 4], and [0, 1]. The screenshot of the annotation platform can be found in Appendix C.2.

Statistics. We generated 11,220 3D models and assigned two independent annotators to each model. Finally, we collected 124,695 raw scores (the sum of 5 dimensions). Similarly to preference data, we also employ three validation strategies: 1) We check the consistency between labeled ranking and score order; 2) We compute the error rate against valid samples (336 models, 3%) labeled by us; 3) We compute the conflict ratio between two annotators. Noting that quality scoring is a very subjective task, we allow a certain degree of deviation when calculating the error rate and conflict ratio. In this paper, we set the absolute deviation threshold to 1. The final score for each dimension is weighted by two records, totaling 56,100 entries.

4 3DGen-Evaluator

An intuitive downstream application of our collected data is training automated evaluators. Given that 3D and video encoding have yet to reach the same level of maturity as image encoding, we propose utilizing multi-view image encoding as an alternative to direct 3D representation in this work. Building on this idea, we propose the 3DGen-Score model, a CLIP-based evaluator, which inherits the advantages of CLIP-like architectures, including open-domain adaptability, ease of expansion, and computational efficiency. However, since it primarily returns numerical scores, its interpretability remains somewhat limited. To address this, we further introduce the 3DGen-Eval model, an MLLM-based evaluator, which leverages the strong visual-text reasoning capabilities of large models, offering enhanced open-world generalization and an interpretable reasoning process. However, MLLM-based models often suffer from inherent irreproducibility, where variations in prompts or even input timing can lead to different outputs. Therefore, these two models complement each other in practice, jointly forming a comprehensive evaluation system.

4.1 CLIP-based 3DGen-Score Model

Traditional CLIP-based metrics are designed for Image Generation tasks, primarily to evaluate the similarity between images and text. However, when applied to 3D generation, these metrics often

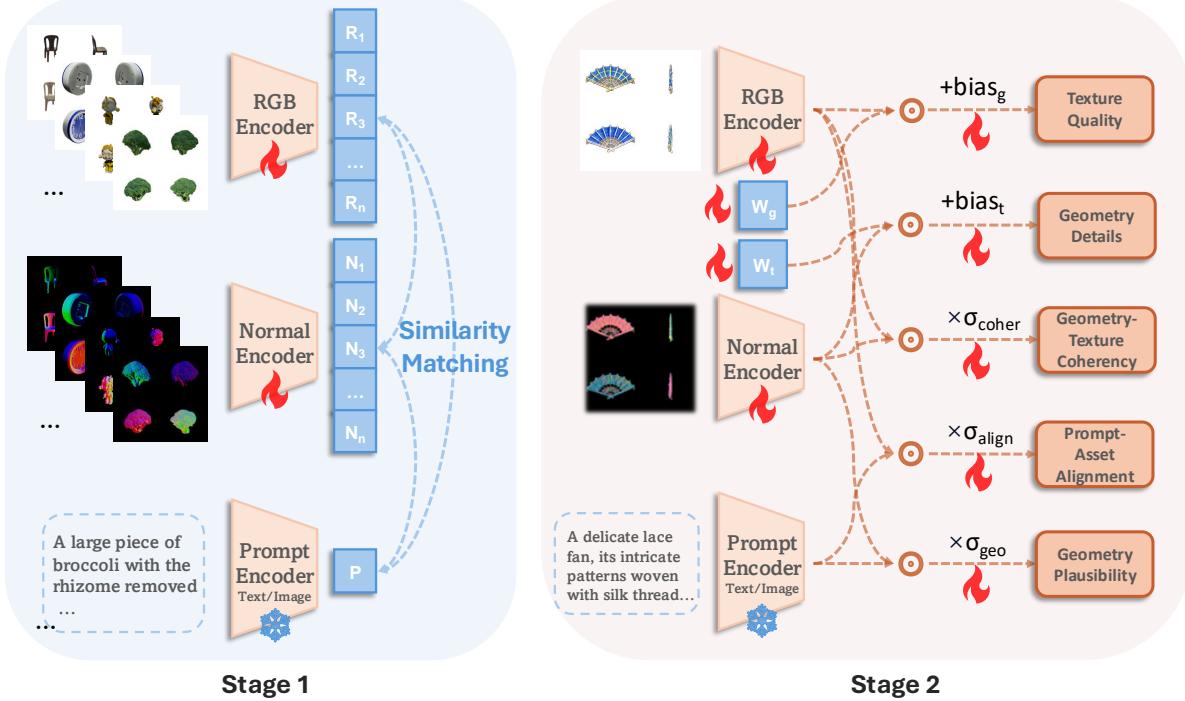


Fig. 5 Overview of 3DGen-Score Model. The model consists of three encoder modules, which can simultaneously process both multi-view normal and RGB images and support both text and image prompts. It works in a pairwise comparison style, and return a five dimension win-rate tuple for each input 3D model. And we apply a "two-stage" training stage. As shown in the left, we finetune visual encoders just like original CLIP to align embedding spaces in the first stage. Then in the second stage (right), we train the five predictors with the supervision of *Comparison votes* data.

overlook adverse aspects such as cross-view consistency and unrealistic repetition due to the lack of information exchange between views. To alleviate this, we modify the original CLIP framework to enable the simultaneous processing of multi-view images. Additionally, we incorporate normal maps, as they tend to capture geometric features more accurately than RGB images.

Model. Our *3DGen-Score* model S , as shown in Figure 5, is equipped with three encoder modules: a prompt encoder module E_p , a normal visual encoder E_n and a RGB visual encoder E_r . Specifically, to support text and image prompts simultaneously, the prompt encoder module incorporates two distinct modal encoders, just like the original CLIP. The addition of two extra visual encoder modules allows for the separation of multi-view encoding from prompt encoding, simplifying the training process and facilitating easier expansion.

Formula. To enhance the interpretability of our score model, we approximate each evaluation dimension in an abstract sense. For example,

the *Prompt-Asset Alignment* dimension is usually approximate as the similarity between RGB renderings and prompts. Similarly, we could calculate the similarity between normal renderings and prompts to estimate *Geometry Plausibility* and the similarity between normal renderings and RGB renderings to estimate *Geometry-Texture Coherence*. As for the *Geometry Details* and *Texture Quality*, illuminated by the LAION Aesthetic Predictor (Schuhmann et al., 2022), we simplify them into two linear predictors. To be specific, given the prompt p and the multi-view normal renderings I_n and RGB renderings I_r of a 3D model, the five dimensions can be formulated as follows:

$$S_{geo}(p, I_n) = \sigma_{geo} \cdot E_p(p) \cdot E_n(I_n), \quad (1)$$

$$S_{align}(p, I_r) = \sigma_{align} \cdot E_p(p) \cdot E_r(I_r), \quad (2)$$

$$S_{coher}(I_n, I_r) = \sigma_{coher} \cdot E_n(I_n) \cdot E_r(I_r), \quad (3)$$

$$S_{geo_detail}(I_n) = W_g \cdot E_n(I_n) + bias_g, \quad (4)$$

$$S_{texture}(I_r) = W_t \cdot E_r(I_r) + bias_t, \quad (5)$$

Stage1. Vision Encoder Adaptation

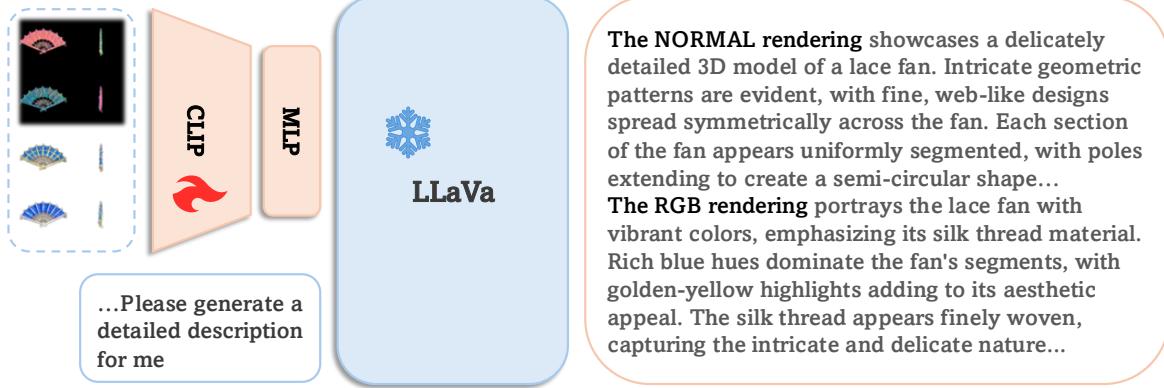


Fig. 6 Vision Encoder Adaption, the first training stage of **3DGen-Eval** model. In this stage, we serve the model as a captioner for multi-view images, supervised by captions generated by GPT-4V.

where σ_{geo} , σ_{align} , σ_{coher} are learnable scalar temperature parameters similar to CLIP, and W_g , $bias_g$, W_t , $bias_t$ are learnable parameters.

Training. We initialize our model from CLIP-ViT-H/14 (Ilharco et al., 2021; Radford et al., 2021), and keep the prompt encoders frozen during training. Noting that CLIP has merely seen four-view normal and RGB images during the pretraining process, we propose a "two-stage" training strategy, shown in Figure 5. In the first stage, we fine-tune vision encoders only with contrastive loss to align the embedding spaces of concatenated normal and RGB renderings with the prompt embeddings. However, images rendered from generated models are often imperfectly aligned with their prompts or even severely collapsed. To address this, we additionally incorporate 3D assets from OmniObject3D (T. Wu, Zhang, et al., 2023) and Cap3D (Luo, Johnson, & Lee, 2024) datasets, which not only exhibit high-quality geometry and texture but are also well aligned with their annotations. Specifically, for each asset, we take its caption annotation as the text prompt and its front-view rendering image as the image prompt. Then, in the second stage, we optimize the parameters of scoring functions supervised by human preference labels. The objective of the second training stage is to minimize the KL-divergence between the softmax-normalized scores and ground-truth labels, formulated as below. Notice that both training stages are trained on the *Comparison votes* data. Meanwhile, rather than unfreezing all layers of the vision encoders, we selectively unfreeze specific layers during training, illustrated by (L. Chen et al., 2023; Z. Sun et al., 2024). In

The **NORMAL** rendering showcases a delicately detailed 3D model of a lace fan. Intricate geometric patterns are evident, with fine, web-like designs spread symmetrically across the fan. Each section of the fan appears uniformly segmented, with poles extending to create a semi-circular shape...
The **RGB** rendering portrays the lace fan with vibrant colors, emphasizing its silk thread material. Rich blue hues dominate the fan's segments, with golden-yellow highlights adding to its aesthetic appeal. The silk thread appears finely woven, capturing the intricate and delicate nature...

this paper, we only unfreeze the last 4 layers. We conducted experiments on two A100 GPUs for 4,000 steps, requiring 5 hours for a single stage with a batch size of 32 and a learning rate that warms up to 3e-6 over the first 500 steps.

$$\min_{\theta} KL(\hat{P}||P) = \sum_{i \in \{0,1\}} p_i \cdot \log(p_i - \hat{p}_i), \quad (6)$$

$$\hat{p}_i = \frac{\exp(s_i)}{\exp(s_0) + \exp(s_1)}, i \in \{0,1\}, \quad (7)$$

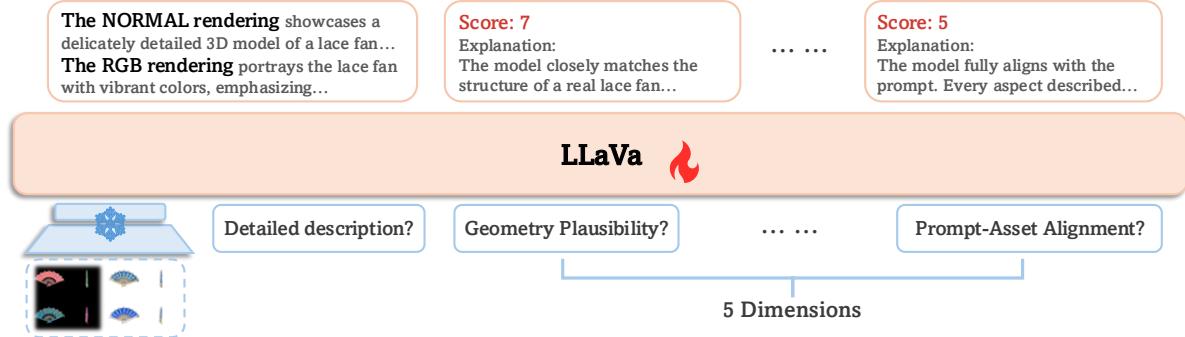
where the ground truth $p_i \in \{0, 0.5, 1\}^5$, representing "win vote", "tie vote" and "fail vote", respectively.

4.2 MLLM-based 3DGen-Eval Model

Building on the success of MLLMs, some explorations (Z. Sun et al., 2024; T. Wu et al., 2024) have advocated the use of MLLMs to simulate human experts for evaluation. However, since GPTEval3D (Qi et al., 2024) calls the black-box GPT-4V directly, the inevitable and unpredictable bias often causes it to deviate from human judgments. Although MV-LLaVa (Z. Sun et al., 2024) has been fine-tuned to preference data, potential biases remain, as pseudo-labels are generated by GPT-4V. To alleviate this, we propose fine-tuning MV-LLaVa with our human annotations.

Preliminary. MV-LLaVA (Z. Sun et al., 2024) is a caption generation and overall quality scoring model for 3D assets, fine-tuned from instructive conversation pairs generated by GPT-4V. It takes four-view images as input and processes visual

Stage2a. Instruct Tuning -- Absolute Score



Stage2b. Instruct Tuning -- Pairwise Comparison

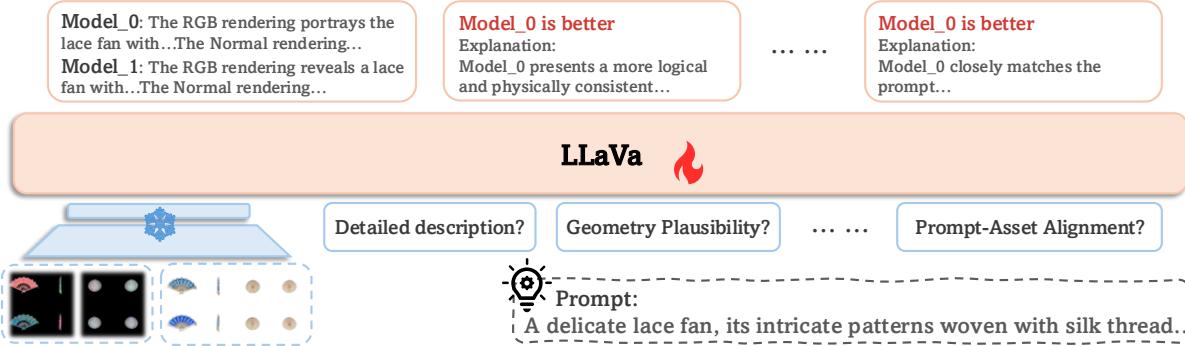


Fig. 7 Instruct Tuning, the second training stage of **3DGen-Eval** model, consisting of two sub-modules. The top sub-module is trained on *Absolute Score* data, to help the model establish an overall understanding of each evaluation dimension. In the subsequent sub-module, the model learns to identify subtle differences between instances within each dimension, using *Pairwise Comparison* data for supervision. Noticing that despite human annotations, we also encourage LLaVa to generate some explanations, guided by responses from GPT-4V, to enhance interpretability.

signals by feeding them into LLaVa (H. Liu, Li, Wu, & Lee, 2024) in the form of separate CLIP embeddings. Inspired by ShareGPT-4V (L. Chen et al., 2023), the model trains the vision encoder only in the first stage, using captions for supervision to enhance multi-view awareness and texture perception. Subsequently, the model evaluates quality based on the multi-view images and captions through a chain-of-thought reasoning process. While MV-LLaVA relies solely on RGB renderings as input and is supervised by GPT-4V, it tends to provide only a coarse evaluation of 3D models and frequently deviates from humans.

Model. We follow the same model structure as MV-LLaVa (Z. Sun et al., 2024), equipped with two primary modules: CLIP for visual embedding and LLaVa for evaluation in a style of Question&Answer(QA). Instead of separate four-view

RGB images, our *3DGen-Eval* model takes concatenated images arranged in a 2*2 grid format as input and additionally incorporates normal renderings to enhance the evaluation of geometry-related dimensions.

Training. We initialize our *3DGen-Eval* model with MV-LLaVa and perform a "two-stage" training process. In the first stage, we finetune the visual encoder while keeping LLaVA frozen, allowing the model to learn to understand images in both the concatenated format and the normal format, as illustrated in Figure 6. In the subsequent "Instruct Tuning" stage, we introduce two sub-modules designed to adapt to our data and distinct objectives, as shown in Figure 7, which separately learn to assign an absolute score for a single 3D model and to compare between two models over five dimensions. Moreover, aimed at enhancing the confidence of the evaluator, except

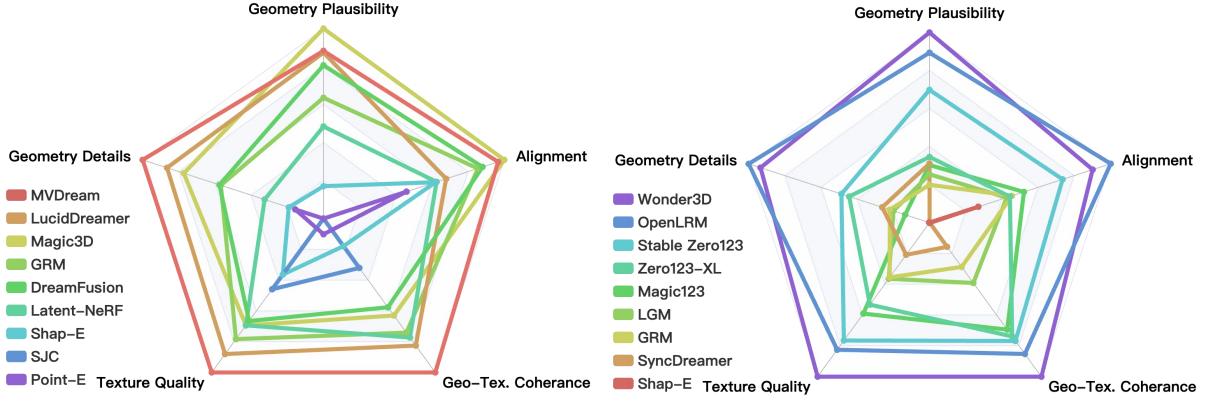


Fig. 8 Radar visualization of Elo scores based on human annotations. **Left:** text-to-3D generative models. **Right:** The top-9 image-to-3D generative models. The legend in the bottom-left corner of each image lists the model ranking based on the average Elo scores across five dimensions, from top to bottom. The detailed leaderboard can be checked in appendix D

solely supervised by the human annotations, the model is also encouraged to generate some reasonable explanations, supervised by responses from GPT-4V. More details of the prompts used for pseudo-labels and templates used in training can be found in appendix F.3

5 Evaluation

In this section, we conduct a preliminary evaluation to assess the alignment of our metric with human judgments. To be specific, we introduce the baseline metrics in Section 5.1, and present the human evaluation results based on the preference data in Section 5.2. Subsequently, we compare our models with baseline metrics in terms of the alignment of human preference in Section 5.3. Furthermore, we conduct validation experiments for the generalization of our models in Section 5.4 and ablation experiments for the training strategy in Section 5.5. Finally, we summarize the advantages of our two models and give some recommendations for automated evaluation in Section 5.6

5.1 Baseline Metrics

We selected 7 evaluation metrics based on various considerations. **1) CLIP Score** (Hessel et al., 2022) is widely used for *Text–Asset Alignment*. The compatibility of normal-caption pairs helps measure *Geometry Plausibility*. In this paper, we use CLIP-ViT-H/14. **2) CLIP Similarity** measures the cosine distance between the CLIP features of the multi-view renderings and the image

prompt, which is used to assess *Image–Asset Alignment*. **3) Aesthetic Score** (Schuhmann et al., 2022) is a CLIP-based linear estimator that predicts the aesthetic quality of images, which can be used to measure the *Texture Quality*. **4) HPSv2.1** (X. Wu et al., 2023) and **5) PickScore** (Kirstain et al., 2023) are human preference scoring models trained on user preference datasets. They are designed to predict preferences for generated images. **6) OpenShape** (M. Liu, Shi, et al., 2024) is a CLIP-based point cloud embedding. We calculate its similarity with the CLIP embedding of the input prompt. **7) GPTEval3D** (T. Wu et al., 2024) is a human-aligned evaluator for text-to-3D generation utilizing GPT-4V. For image-to-3D generation, we use BLIP (J. Li et al., 2022) to obtain text captions. All metrics are calculated on the annotated test dataset. For fairness, we average the four-view scores as the final value. Except for GPTEval3D (T. Wu et al., 2024), limited by API, we uniformly sample 40 objects as test data. Additionally, due to its formatting restriction and modality limitation, we cannot achieve a fair comparison with **T³Bench** (Y. He et al., 2023) in this section. However, we provide supplementary comparison experiments in Appendix F.2.

5.2 Human Preference Evaluations

In this section, we conduct a comprehensive assessment of each generative model based on *human expert annotations*. Specifically, we calculate the average Elo scores for each model

Table 2 Text-to-3D pairwise rating alignment with human judgment. We assess the average probability of alignment with human judgments for each comparison. Our method demonstrates strong alignment across all criteria, with higher values indicating better performance. **Bold** and underlined results indicate the best and second-best performers.

Methods	Plausibility	Geo. Details	Tex. Quality	Geo-Tex.	Alignment	Average
CLIP	0.666	0.640	0.684	0.702	0.612	0.661
Aesthetic score	0.524	0.536	0.557	0.549	0.508	0.535
HPSv2.1	0.505	0.505	0.507	0.507	0.503	0.505
PickScore	0.625	0.646	0.684	0.705	0.567	0.645
OpenShape	0.502	0.501	0.500	0.500	0.502	0.501
MV-LLaVa	0.493	0.514	0.495	0.490	0.511	0.501
GPTEval3D	0.660	<u>0.667</u>	<u>0.699</u>	0.767	0.592	<u>0.677</u>
3DGen-Eval(Ours)	<u>0.687</u>	0.665	0.689	0.696	<u>0.625</u>	0.672
3DGen-Score(Ours)	0.729	0.707	0.760	<u>0.764</u>	0.667	0.725

Table 3 Image-to-3D pairwise rating alignment with human judgment. We assess the average probability of alignment with human judgments for each comparison. Our method demonstrates strong alignment across all criteria, with higher values indicating better performance.

Methods	Plausibility	Geo. Details	Tex. Quality	Geo-Tex.	Alignment	Average
CLIP similarity	0.529	0.523	0.542	0.538	0.524	0.531
Aesthetic score	0.560	0.560	0.567	0.566	0.545	0.560
OpenShape	0.500	0.501	0.499	0.499	0.500	0.500
MV-LLaVa	0.463	0.447	0.441	0.45	0.485	0.457
GPTEval3D	0.703	0.676	0.714	0.701	<u>0.654</u>	0.670
3DGen-Eval(Ours)	<u>0.744</u>	<u>0.750</u>	<u>0.754</u>	<u>0.755</u>	0.651	<u>0.731</u>
3DGen-Score(Ours)	0.777	0.768	0.778	0.798	0.712	0.767

across the five criteria and present radar charts in Figure 8, which depict the top 9 models for Text-to-3D and Image-to-3D tracks.

Text-to-3D. According to our criteria, the top-3 methods in text-to-3D track are MVDream (Y. Shi et al., 2023), Lucid-Dreamer (Liang et al., 2023), and Magic3D (Lin et al., 2023), achieving average Elo scores of 1177.66, 1112.21, and 1088.93, respectively. The model ranking allows us to explore each model’s strengths and weaknesses and gain some insights for further improvement. The success of MVDream (Y. Shi et al., 2023) demonstrates the advancement of multi-view diffusion model, which not only achieves the effective combination of 2D prior and 3D prior but also facilitates the information exchange between different views. Building on the 3DGS representation, Lucid-Dreamer (Liang et al., 2023) performs well on geometry and texture quality. Magic3D (Lin et al., 2023) benefits from a “two-stage” optimization strategy but is somehow short on texture quality.

Image-to-3D. On the image-to-3D track, Wonder3D (Long et al., 2023), OpenLRM (Z. He & Wang, 2023), and Stable Zero123 (R. Liu et al., 2023) emerge as the frontrunners, achieving average Elo scores of 1304.05, 1279.67, and 1200.69, respectively. Wonder3D (Long et al., 2023) employs a multi-view cross-domain attention mechanism, which again verifies the significance of 2D&3D domain blending and multi-view consistency. Benefiting from Objaverse (Deitke et al., 2024, 2023) datasets, OpenLRM reinvigorates the transformer-based encoder-decoder framework. Meanwhile, Stable Zero123 (R. Liu et al., 2023) achieves a comprehensive improvement over Zero123-XL (R. Liu et al., 2023), but still leaves room for improvement in geometric details.

5.3 Comparisons on Human Preference Correlation

In this section, we assess the alignment between our proposed metric and human preferences. To

Table 4 Text-to-3D ranking alignment with human judgment. The table illustrates Kendall’s tau ranking correlation (Kendall, 1938) between Text-to-3D rankings produced by various metrics and those determined by human experts. A higher correlation suggests a stronger alignment with human judgment. **Bold** and underlined results indicate the best and second-best performers.

Methods	Plausibility	Geo. Details	Tex. Quality	Geo-Tex.	Alignment	Average
CLIP-Score	<u>0.667</u>	<u>0.611</u>	<u>0.722</u>	<u>0.833</u>	0.333	0.635
Aesthetic score	0.333	0.389	0.556	0.667	0.167	0.422
HPSv2.1	0.389	0.444	0.611	0.722	0.222	0.478
PickScore	0.333	0.389	0.556	0.667	0.167	0.422
OpenShape	0.071	0.071	0.000	-0.143	0.143	0.029
MV-LLaVa	-0.278	0.056	-0.056	-0.111	0.111	-0.056
GPTEval3D	0.500	0.500	0.889	<u>0.833</u>	<u>0.389</u>	0.622
3DGen-Eval(Ours)	0.722	0.722	<u>0.722</u>	<u>0.833</u>	0.333	<u>0.667</u>
3DGen-Score(Ours)	<u>0.667</u>	0.556	0.889	0.944	0.500	0.711

Table 5 Image-to-3D ranking alignment with human judgment. The table illustrates Kendall’s tau ranking correlation (Kendall, 1938) between Image-to-3D rankings from various metrics and human experts. A higher correlation suggests a stronger alignment with human judgment.

Methods	Plausibility	Geo. Details	Tex. Quality	Geo-Tex.	Alignment	Average
CLIP similarity	0.256	0.154	0.256	0.256	0.128	0.210
Aesthetic score	0.487	0.436	0.590	0.538	0.513	0.513
OpenShape	0.600	0.527	0.636	0.636	0.636	0.607
MV-LLaVa	-0.179	-0.231	-0.308	-0.282	-0.154	-0.231
GPTEval3D	0.564	0.590	0.667	0.667	0.513	0.600
3DGen-Eval(Ours)	<u>0.744</u>	<u>0.744</u>	<u>0.744</u>	<u>0.795</u>	<u>0.692</u>	<u>0.744</u>
3DGen-Score(Ours)	0.897	0.872	0.821	0.923	0.769	0.856

accomplish this, we employ the Softmax function over basic metrics to assign a win probability for each evaluation match. Experimental results demonstrate a strong correlation between our models and human judgment.

Pairwise Alignment. Let p_i represent the predicted win rating of the left model player in the i -th battle and q_i denote the human judgment of making the same choice. Thus, the pairwise rating alignment is given by $\mathbb{E}_i[p_i q_i + (1 - p_i)(1 - q_i)]$. A higher value signifies a stronger alignment with human experts. As demonstrated in Table 2 and Table 3, our *3DGen-Score* model consistently outperforms other metrics, highlighting its versatility across various evaluation criteria. Moreover, our *3DGen-Eval* model achieves significant improvements in all dimensions compared to MV-LLaVa (Z. Sun et al., 2024). In some dimensions, it even outperforms GPT-4V, demonstrating the effectiveness of the fine-tuning process.

Ranking Alignment. Table 4 and Table 5 display the ranking correlations between various

evaluation metrics and the reference Elo scores computed from the expert annotations. All metric scores, excluding GPTEval3D (Qi et al., 2024), are based on the average scores of prompts in the test set, while GPTEval3D and our models calculate scores using the Elo rating system, starting with an initial score of 1,000 and a K factor of 32. Table 5 shows that our methods achieves the highest ranking across all dimensions on the image-to-3D track, securing the 1st and 2nd positions. As for the text-to-3D ranking, our *3DGen-Eval* model demonstrates impressive performance in the geometric-related dimension, while our *3DGen-Score* model achieves the highest overall ranking on average, with a notable improvement in "Prompt-Asset Alignment" dimension.

Analysis Experimental results demonstrate that our models outperform all existing metrics across comprehensive evaluation dimensions, regardless of whether the assessment is conducted on pairwise alignment or ranking correlation and in both text-to-3D and image-to-3D tracks. Notably,

Table 6 Pairwise rating alignment with additional Instant-Mesh. We assess the average probability of alignment with human judgments for each comparison between Instant-Mesh and other 13 image-to-3D models.

Methods	Plausibility	Geo. Details	Tex. Quality	Geo-Tex.	Alignment	Average
CLIP similarity	0.667	0.654	0.703	0.714	0.751	0.698
Aesthetic score	0.529	0.544	0.558	0.547	0.559	0.547
OpenShape	0.544	0.580	0.603	0.587	0.553	0.573
MV-LLaVa	0.553	0.550	0.558	0.565	0.535	0.552
GPTEval3D	<u>0.711</u>	0.782	<u>0.789</u>	0.776	<u>0.757</u>	<u>0.763</u>
3DGen-Eval(Ours)	0.700	0.712	0.727	0.708	0.692	0.708
3DGen-Score(Ours)	0.754	<u>0.741</u>	0.826	<u>0.765</u>	0.815	0.780

Table 7 Pairwise rating alignment compared on 3DRewardDB. We assess the average alignment across considered dimensions for each method.

Methods	3DGen-Score(Ours)	3DGen-Eval(Ours)	GPTEval3D	PickScore	HPSv2.1	CLIP
Alignment	0.716	0.576	0.676	0.505	0.675	0.671

among our models, the CLIP-based 3DGen-Score outperforms the MLLM-based 3DGen-Eval model. We attribute this to CLIP’s simpler architecture and fewer parameters, which enable more efficient training and contribute to stronger performance. Additionally, model performance is generally stronger on the image-to-3D track compared to the text-to-3D track. We hypothesize that this discrepancy arises from two key factors: first, the image-to-3D track benefits from a larger volume of training data; second, the process of aligning text and images within encoders may introduce biases. Furthermore, the performance of 3DGen-Score on the text-to-3D track in the ranking setting highlights the necessity of training 3DGen-Eval to improve generalization to normal images.

5.4 Generalization Performance

In Section 5.3, we conduct evaluation experiments only on our proprietary dataset, which may introduce specific biases and limit our ability to assess the model’s performance and robustness. To alleviate it, we expand our testing to include external data, such as new 3d models and external dataset.

Evaluation on New Model To verify the robustness of our model on unseen data, we utilize Instant-Mesh (J. Xu, Cheng, et al., 2024), an advanced image-to-3D model, to generate 510 3D assets using our prompt set. We then randomly pair these assets with those from 13 other image-to-3D models, resulting in 130 comparison pairs.

Subsequently, we calculate pairwise rating alignment between our model and annotations from experts, shown in Table 6. The results indicate that our 3DGen-Score model exhibits strong generalization capabilities, allowing it to assess previously unseen models while surpassing existing methods across multiple dimensions. This underscores its effectiveness as a robust and reliable evaluation metric for 3D assessment tasks.

Evaluation on External Dataset We conduct experiments using 3DRewardDB (Ye et al., 2024), a coarse preference dataset of generated text-to-3D models and, to our knowledge, the only available relative dataset. However, 3DRewardDB provides four-view perspectives only for each model, lacking normal maps or raw 3D assets. To address this limitation, we apply Metric3D (Yin et al., 2023), the state-of-the-art method for surface normal estimation, to predict normal maps. We then sample 120 comparison pairs as the test set. Notably, 3DRewardDB derives final scores based on alignment, consistency, and overall quality while omitting criteria such as Geometry Details and Geometry-Texture Alignment during its labeling process. To align its criteria with ours, we exclude those less-considered dimensions and aggregate scores for “3D Plausibility”, “Texture Quality”, and “Asset-Prompt Alignment” to compute a final score. The same approach is applied to GPTEval3D (T. Wu et al., 2024). The experimental results are presented in Table 7. Since the normal maps used here are estimated using

Table 8 Human alignment results of ablation experiments. The first 3 columns demonstrate training strategies, where **C.loss** means "contrastive loss" applied in stage-1, and **V.enc** means "finetune vision encoders" applied in stage-2. The rest columns demonstrate the **pairwise alignment** on test data, and the 'Average' column is weighted by the count of involved models. The higher values indicate better alignment with human judgments.

Strategy		Dimensions					Average
C.loss	V.enc.	Plausibility	Geo. Details	Tex. Quality	Geo-Tex.	Alignment	
✓		0.759	0.714	0.735	0.759	0.684	0.730
	✓	0.746	0.742	0.771	0.776	0.686	0.744
✓	✓	0.757	0.747	0.771	0.784	0.694	0.750

Table 9 Model robustness results of ablation experiments. We calculate the **pairwise alignment** on comparisons between Instant-Mesh and other 13 image-to-3D models. The higher values indicate better model robustness.

Strategy		Dimensions					Average
C.loss	V.enc.	Plausibility	Geo. Details	Tex. Quality	Geo-Tex.	Alignment	
✓		0.715	0.599	0.781	0.665	0.738	0.700
	✓	0.677	0.757	0.774	0.781	0.800	0.758
✓	✓	0.754	0.741	0.826	0.765	0.815	0.780

Metric3D rather than ground truth, this may introduce visual bias, potentially affecting the performance of our scoring models in geometry-related dimensions. Despite this limitation, our scoring models still achieve outstanding performance on external datasets, demonstrating not only strong generalization capabilities but also remarkable robustness in handling diverse data sources and inherent estimation noise.

5.5 Ablation Experiments

We introduce a "two-stage" training strategy in Section 4. To validate its effectiveness and necessity, we carefully design and conduct ablation experiments on the *3DGen-Score* model. Furthermore, considering the requirements of the evaluation task, we thoroughly assess two aspects: **1) alignment with human judgments** and **2) robustness to new models**. Specifically, we compute pairwise alignment on test pairs to measure the models' alignment with human judgments, and use Instant-Mesh pairs to assess their generalization performance. The experimental results, presented in Table 8 and Table 9, demonstrate that the two-stage training strategy generally outperforms single-stage training. In particular, the second-stage training enhances the model's alignment with human judgments, while incorporating the first stage further strengthens its robustness to new models.

5.6 Discussion

In summary, our two models each have distinct advantages. The 3DGen-Score model aligns well with human preferences while remaining computationally efficient and highly scalable. In contrast, the 3DGen-Eval model provides superior interpretability, matching or even surpassing GPT-4V in human preference alignment without incurring additional costs. Based on these strengths, we recommend the 3DGen-Eval model for detailed quality analysis of individual samples. For method-level comparisons, the 3DGen-Score model proves to be a suitable choice, particularly for the image-to-3D track. However, for those geometry-related dimensions of text-to-3D models, we recommend referring to the 3DGen-Eval model's results. Additionally, we also provide a CLIP-based absolute scoring predictor in Appendix E, which is more computationally efficient, allowing model ranking without the need for comparisons.

6 Application

Using human preference data to establish leaderboards or train automatic scoring models is a straightforward application, but its potential extends beyond these use cases. One of the most critical applications is leveraging it as a reward signal to optimize generative models. In practice, Reinforcement Learning from Human Feedback

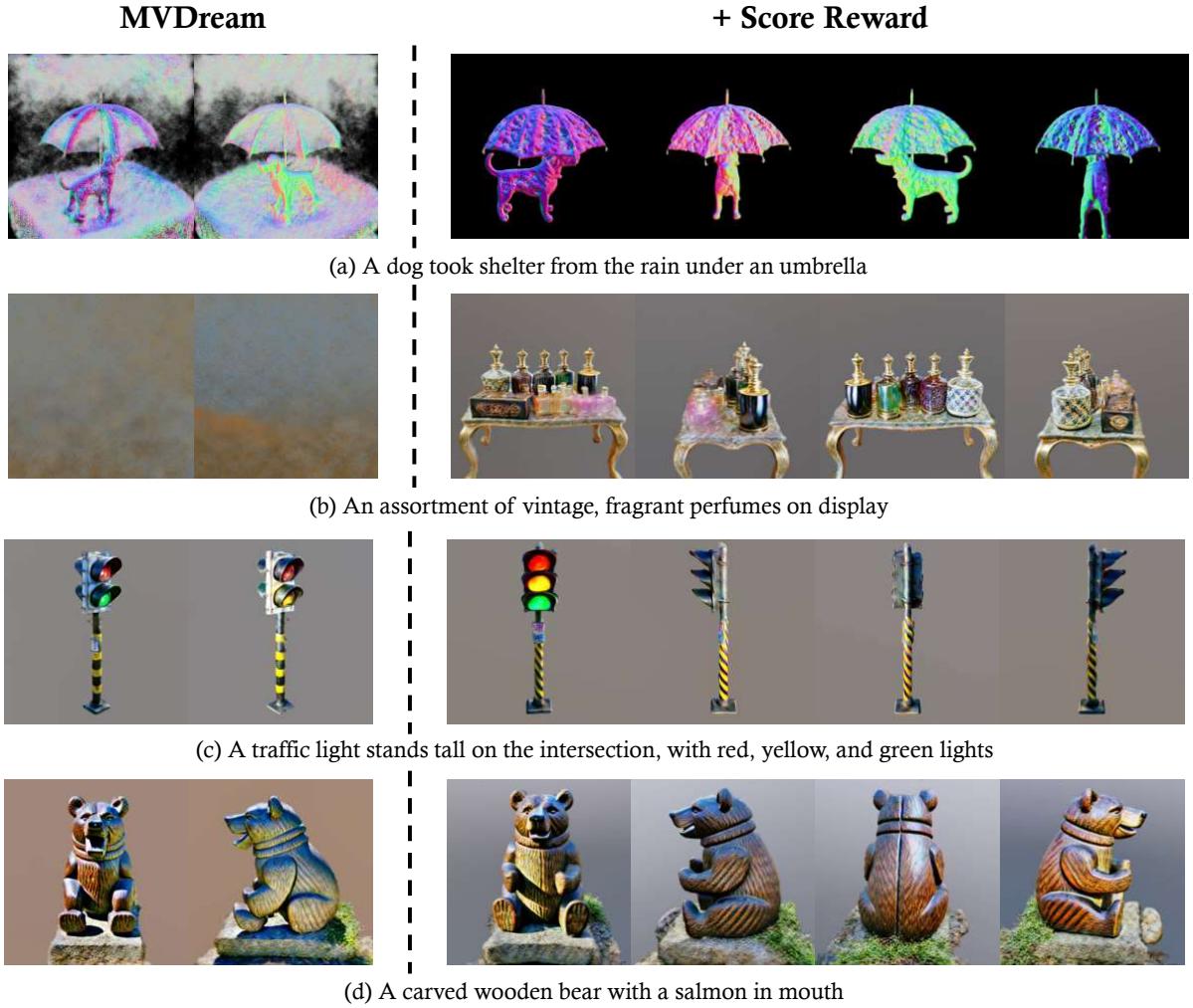


Fig. 9 Visual results of Reward Generation We operate experiments over MVDream (Y. Shi et al., 2023), with original results on **Left** and RM-optimized on **Right**.

(RLHF) has gained significant traction in the field of language and image generation (Bai et al., 2022; Nakano et al., 2021; Ouyang et al., 2022; Stien-non et al., 2020; J. Xu, Liu, et al., 2024). Inspired by recent works such as (J. Xu, Liu, et al., 2024; Ye et al., 2024), we similarly employ *3DGen-Score* as a Reward Model (RM) to simulate human feedback, allowing us to refine the behavior of generative models. To achieve this, we simplify the final objective function into the following formula:

$$\mathcal{L}_{Reward}(\theta) \approx \mathcal{L}_{SDS} - \lambda_r r(p, I_{normal}, I_{rgb}), \quad (8)$$

where r represents the *3DGen-Score* model, λ_r is a hyperparameter, p denotes the prompt, and

I_{normal} and I_{rgb} are rendering images arranged in four views. For a detailed derivation, please refer to DreamReward (Ye et al., 2024).

Experiments. In this paper, we conduct experiments on MVDream(Y. Shi et al., 2023), which ranks as the top text-to-3D generative model according to our criteria. Moreover, MVDream generates multi-view images in parallel during optimization, which is particularly compatible with our objective function. In practice, we derive the final reward for each asset by averaging the scores across five evaluation dimensions. As shown in Figure 9, our approach effectively enhances the quality and coherence of 3D generative outputs

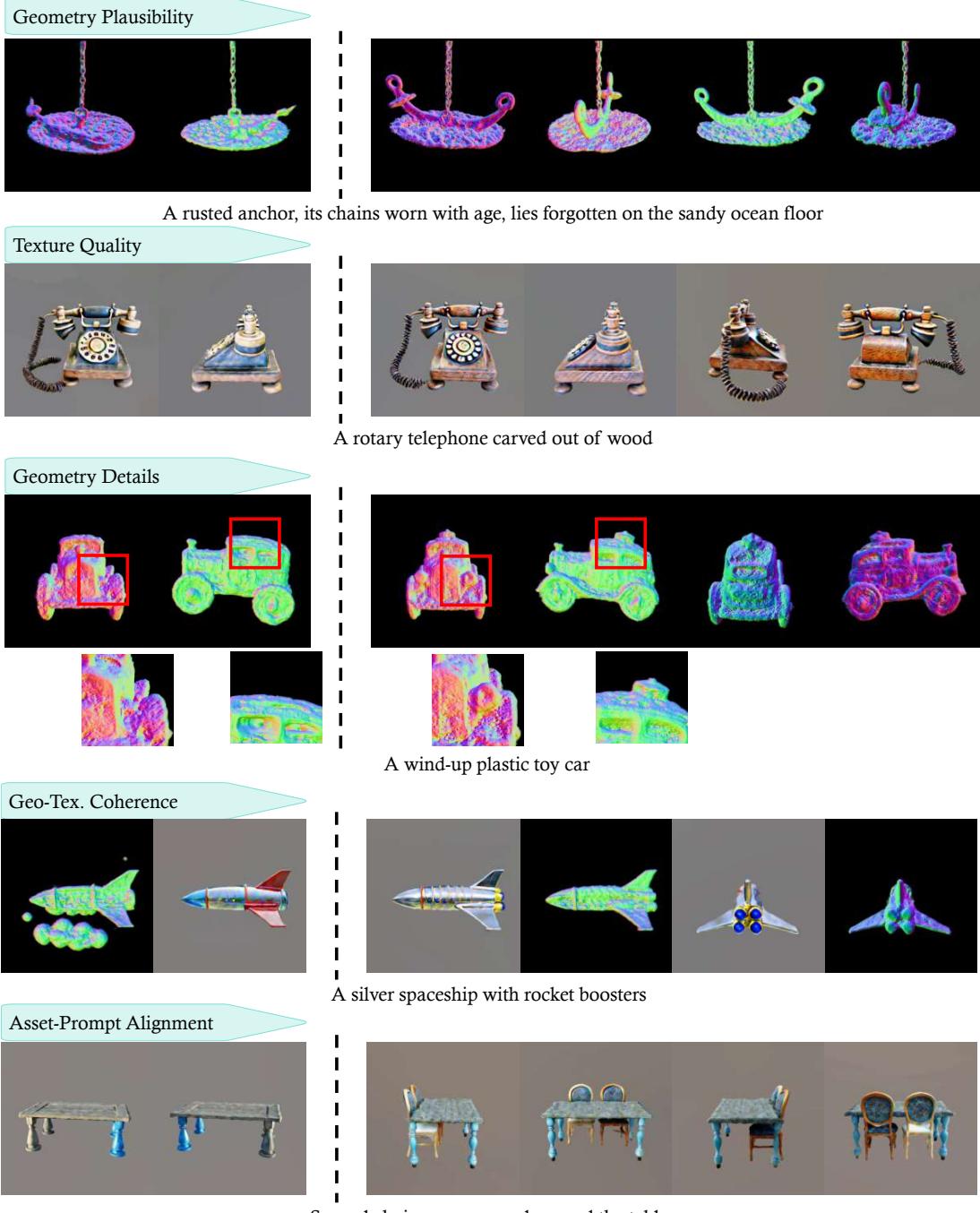


Fig. 10 Visual results of Reward Generation for each dim

across critical dimensions. For instance, examples (a) and (b) illustrate improvements in noise reduction, while examples (c) and (d) exhibit the enhanced capability of simulating 3D priors and refining textures. These results indicate that our

3DGen-Score model provides valuable insights for advancing generative performance.

Dimension Study. We explored the impacts on each evaluation dimension through taking single dimension scores as the final reward. Visual

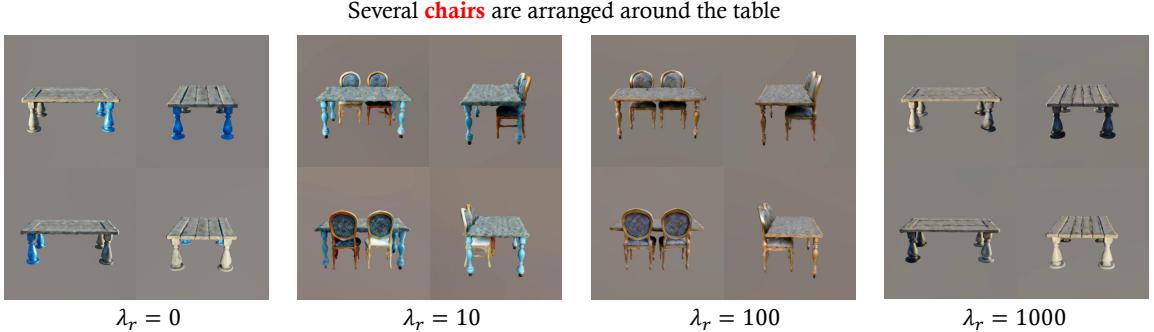


Fig. 11 Visual Comparison on various λ_r

results in Figure 10 highlight the potential for targeted improvements in specific dimensions. Interestingly, we find that the effect of the reward score is not entirely isolated to each dimension. Instead, there is notable interaction across different dimensions, as optimizing for one dimension can also enhance or impact performance in others.

Ablation Study. We investigate the impact of varying the value of λ_r . As shown in Figure 11, the effects of λ_r are not purely linear. Positive influence is observed only within an optimal range. When λ_r is set too low, the improving potential remains underutilized. Conversely, if λ_r is too large, the benefits diminish, and the generation process even begins to show signs of collapse.

7 Conclusions

In this work, we present *3DGen-Bench*, the first large-scale human preference dataset for 3D models. To construct this dataset, we developed *3DGen-Arena*, a data platform that integrates 11,200 3D models generated by 19 different 3D generative models. By collecting votes from both public users and expert annotators, *3DGen-Bench* eventually compiles over 68,000 preference votes and more than 56,000 score labels. Leveraging this data, we conduct a comprehensive evaluation and analysis of 3D generative models across five criteria, providing valuable insights for the advancement of 3D generation research. Additionally, we train two scoring models, *3DGen-Score* and *3DGen-Eval*, which serve as automated 3D evaluators, demonstrating superior alignment with human judgment and greater robustness to new models compared to existing metrics.

Limitations. Our collection now includes 19 open-source generative models, but many works in this area still remain closed-source. Our goal is

to involve state-of-the-art models and consistently update our leaderboard with the latest advancements. Additionally, due to the lack of a robust 3D embedding model, we opt for 2D CLIP embedding as an alternative. Developing more advanced 3D embedding techniques remains a priority to fully leverage the naive 3D data for evaluation.

Future Works. Our 3DGen-Bench dataset innovatively realizes the subdivision of the annotation dimension, rescuing it from a holistic and rough score. While our annotations currently operate at the instance level, evolving evaluation standards may encourage finer-grained labeling, such as region-level annotations. Additionally, there remains a need for a more advanced 3DGen-Evaluator. Future iterations should explore more sophisticated model architectures or improved embedding techniques to enhance the evaluator’s effectiveness and alignment.

Data Availability. The dataset used in this study is publicly available in the 3DGen-Bench repository: <https://huggingface.co/datasets/3DGen/3DGen-Bench>. Noting that the 3D assets adhere to the licensing terms of the respective methods from which they are derived, any use of the 3D data must comply with the original licensing terms and copyright policies.

Appendix

In Appendix, we provide the basic information and licensing terms of our dataset in Section A, along with additional details regarding its construction in Section B. We then present further instructions on the annotation platform and human leaderboard in Section C and Section D, respectively. Finally, we introduce a new scoring model in absolute form in Section E and provide more experimental details in Section F.

A Dataset Information

A.1 Dataset Link and Documentation

The dataset we established in this study, together with its metadata and license, is now publicly available in the 3DGen-Bench repository: <https://huggingface.co/datasets/3DGen/3DGen-Bench>. Specifically, we release the curated prompts, generated 3D models along with their renderings, and annotated human preferences. Detailed documentation of its structure and usage can be found in the dataset card, and the Croissant metadata record can be viewed and downloaded by reviewers at <https://huggingface.co/api/datasets/3DGen/3DGen-Bench/croissant>. Additionally, we host the 3DGen-Arena app on Huggingface Space <https://huggingface.co/spaces/ZhangYuhan/3DGen-Arena>, enabling online voting among our diverse 3D assets.

A.2 Licensing Terms

Depending on the data source, the prompts and annotations are released under the [MIT License](#), while the 3D assets follow the original licenses of the respective methods. Any use of the data must comply with the corresponding licensing terms and copyright policies.

B Additional Details on Dataset Construction

B.1 Prompt Generation

We provide 510 text prompts in TXT format and 510 image prompts in RGBA format, generated respectively from ChatGPT and Stable Diffusion. Specifically, for text prompts, we first identified 6 basic category domain: **"Vehicle"**, **"Plant"**, **"Animal"**, **"Food"**, **"Indoor Object"** and **"Outdoor Object"**. Then, aiming to obtain as many specific categories as possible, we prompted ChatGPT with *"We want to generate a prompt suite for text-to-3D generation evaluation. We hope the suite is comprehensive and challenging. The first step is to define category domain. We totally spit into 6 prime domain: vehicle, animal, plant, food, indoor, outdoor. Now we need you to list specific categories for each domain as many as possible"*. Subsequently, equipped with diverse

categories, we started to generate descriptive sentences. In order to generate prompts at different difficulty levels, we employed the following templates in turn.

- **Noun phrase:**
{a/an}{attribution}{category}{with ...}
- **Simple sentence:**
{count}{attribution}{category}{adv./verb-phrase /with ...}
- **Composite sentence:**
{count}{attribution}{category}{verb-phrase}{count}{attribution}{category}
- **Other templates:**
allow ChatGPT to generate freely

where {attribution} and {adv./verb-phrase/with ...} are optional, {count} should smaller than six if defined explicitly, and prompts are encouraged to contain descriptive part to describe either geometric features or appearance features concretely. Through heuristic generation in different domains, we obtained 600 raw prompts. We then do manual screening and targeted regeneration to ensure the balance of categories and difficulties. Eventually, we obtained 510 text prompts.

B.2 3D Generative Models

In our benchmark, we include 9 text-to-3D generative models: Mvdream ([Y. Shi et al., 2023](#)), Lucid-dreamer ([Liang et al., 2023](#)), Magic3D ([Lin et al., 2023](#)), GRM ([Y. Xu et al., 2024](#)), Dream-fusion ([Poole et al., 2022](#)), Latent-NeRF ([Metzger et al., 2023](#)), Shap-E ([Jun & Nichol, 2023](#)), SJC ([H. Wang et al., 2023](#)), and Point-E ([Nichol et al., 2022](#)), and 13 image-to-3D generative models: Wonder3D ([Long et al., 2023](#)), OpenLRM ([Z. He & Wang, 2023](#)), Stable Zero123 ([R. Liu et al., 2023](#)), Zero-1-to-3 XL ([R. Liu et al., 2023](#)), Magic123 ([Qian et al., 2024](#)), LGM ([Tang et al., 2024](#)), GRM ([Y. Xu et al., 2024](#)), Sync-Dreamer ([Y. Liu et al., 2023](#)), Shap-E ([Jun & Nichol, 2023](#)), Triplane-Gaussian ([Zou et al., 2023](#)), Point-E ([Nichol et al., 2022](#)), Escher-Net ([Kong et al., 2024](#)), and Free3D ([C. Zheng & Vedaldi, 2023](#)).

B.3 Annotation Cost

Throughout the annotation process, we employed 47 expert annotators at a rate of \$6 per hour,

Table R1 Leaderboard of Text-to-3D generative models. Ranked by the average Elo score computed from comparison annotations.

Method	Plausibility	Geo. Details	Tex. Quality	Geo-Tex.	Alignment	Average
MvDream	1107.90	1182.60	1229.61	1270.78	1095.39	1177.66
LucidDreamer	1105.05	1133.85	1158.18	1163.19	1000.79	1122.21
Magic3D	1143.94	1100.30	1046.93	1042.77	1106.71	1088.93
GRM	1032.35	1026.28	1100.23	1111.33	1058.49	1065.74
Dreamfusion	1084.69	1029.05	1030.02	1009.73	1066.99	1044.10
Latent-NeRF	986.09	938.78	1046.76	1130.76	983.47	1017.97
Shap-E	889.87	890.49	849.57	764.86	981.75	875.71
SJC	837.64	821.15	906.59	852.02	777.70	839.71
Point-E	812.47	877.50	632.11	654.56	928.72	781.71

Table R2 Leaderboard of Image-to-3D generative models. Ranked by the average Elo score computed from comparison annotations.

Method	Plausibility	Geo. Details	Tex. Quality	Geo-Tex.	Alignment	Average
Wonder3D	1333.95	1308.03	1347.60	1321.71	1210.94	1304.05
OpenLRM	1294.69	1330.10	1270.31	1260.36	1244.88	1279.67
Stable Zero123	1222.58	1157.20	1243.73	1225.31	1154.64	1200.69
Zero123-XL	1092.45	1142.73	1140.74	1213.14	1055.74	1128.96
Magic123	1076.61	1038.36	1166.43	1194.64	1082.10	1111.23
LGM	1058.78	1060.23	1066.24	1068.82	1049.68	1060.35
GRM	1038.48	1067.10	1062.29	1025.53	1059.53	1042.99
SyncDreamer	1079.61	1081.16	997.50	971.29	905.29	1006.97
Shap-E	965.11	993.74	905.02	906.95	996.59	953.48
TriplaneGaussian	840.28	819.21	880.53	893.25	836.79	854.01
Point-E	740.57	769.02	707.68	704.43	916.69	767.68
EscherNet	721.99	690.57	689.32	728.21	843.21	734.66
Free3D	534.90	542.55	522.61	486.36	643.91	546.47

spending approximately \$4,700 in total. The cleaned results are published on Huggingface¹.

individual model interactions. The corresponding screenshots are listed in Figure S8, S9, S10.

C Additional Details about Annotation Platform

C.1 3DGen-Arena

We run the 3DGen-Arena app on Huggingface Space². It supports online voting across our manifold 3D assets, offering three interaction modes: ‘Anonymous battle models’, ‘Named battle models’ and ‘Single model chat’. The first two modes enable model comparisons, differing in whether the voter is aware of the participant models’ identities, while the third mode allows for

C.2 More Platforms

Comparison Votes. This process follows the same rules as the anonymous comparison in 3DGen-Arena, but given fixed battle pairs sampled by us to maintain balanced data distribution. In addition, to maintain the integrity and authenticity of the voting process, participants are required to vote using their real names, adding a layer of accountability to the evaluation. The screenshot of voting platform is in Figure S11.

Absolute Scores. To enhance the credibility of the scoring results, we command annotators to rank all 3D models generated from the same prompt firstly, ensuring a consistent and fair assessment across similar generations. Subsequently, annotators proceed to label each

¹<https://huggingface.co/datasets/3DGen/3DGen-Bench>

²<https://huggingface.co/spaces/ZhangYuhan/3DGen-Arena>

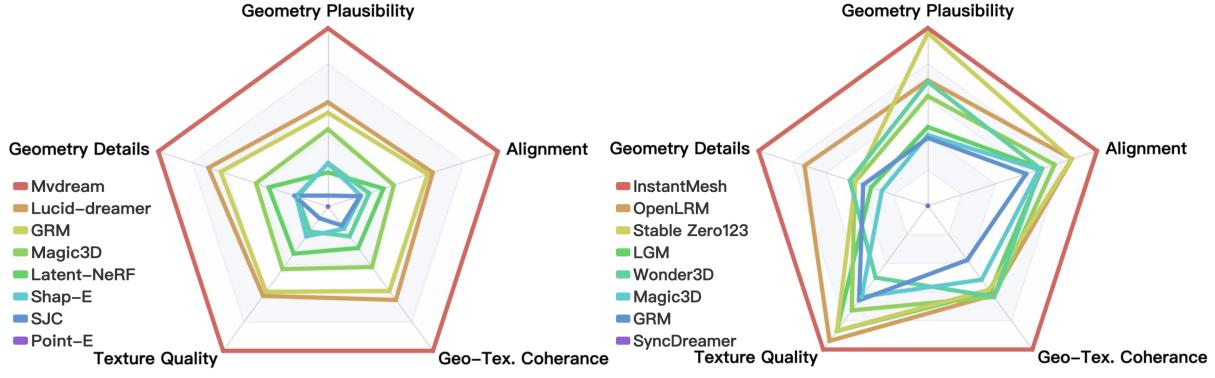


Fig. S1 Elo scores of Generative models calculated by anonymous votes gathered from 3DGen-Arena. **Left:** Text-to-3D generative models. **Right:** The top-9 Image-to-3D generative models. The legend in the down left corner of each image lists the ranking of each method based on the average across five dimensions of Elo scores, from top to bottom. The accurate scores can be checked in the leaderboard page running on huggingface ².

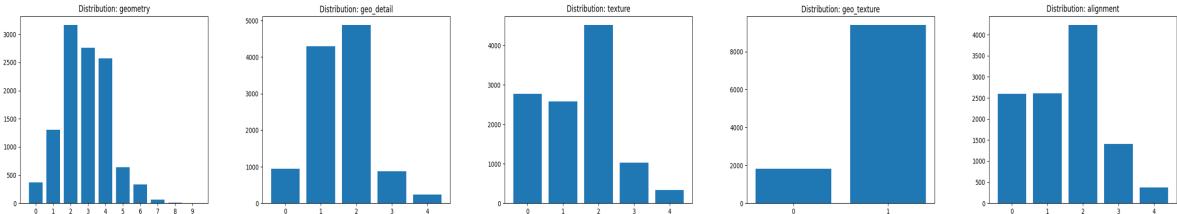


Fig. S2 Distribution of Absolute Score data. It shows that high-quality 3D models are few in the collected dataset, annotated only over generated models.

model systematically, maintaining uniform criteria throughout the evaluation process. This dual-step approach helps standardize scoring, reduce bias, and improve the reliability of our results. The scoring platform screenshot is in Figure S12.

D Human Leaderboard

Expert Annotations We employ the Elo rating algorithm (Elo, 1967) to compute the human leaderboard on our collected 13.8k comparison annotations. This approach enables a dynamic and adaptive ranking system, where models are evaluated based on their relative performance against each other. The accurate elo scores are listed in Table R1 and Table R2.

Anonymous Votes After hosting the 3DGen-Arena app for several months, we have collected 8,045 votes from anonymous visitors until now, with 6,351 votes for text-to-3D track and 1,694 for image-to-3D track. Based on these anonymous votes, we establish the leaderboard similarly by calculating Elo rankings, as shown in Figure S1.

Additionally, we also collected some votes from 'Named battle models' mode. And the detailed leaderboard that aggregates these two types of data is presented in huggingface ².

E Absolute Scoring Predictor

E.1 3DGen-Score Predictor

We proposed the *3DGen-Score* model in Section 4.1 to predict the win rate between two candidate models, learned from *Comparison Votes* data, while *Absolute Scores* data remain under exploration. To fully leverage the potential of these data, we introduce *3DGen-SP*, short for 3DGen-Score Predictor. Inspired by LAION Aesthetic Predictor V1 (Schuhmann et al., 2022), we extend vision encoders by incorporating five additional predictor heads to predict absolute scores for each dimension of the criterion.

Data Preparation As shown in Figure S2, the distribution of absolute score annotations for the 11.2k generated 3D models is imbalanced, with high-quality samples being underrepresented in

Table R3 Ranking alignment with human judgment. we compute the Kendall’s tau ranking correlation (Kendall, 1938), where a higher value suggests a stronger alignment with human judgment.

Task	Method	Plausibility	Geo. Details	Tex. Quality	Geo-Tex.	Alignment	Average
Text-to-3D	CLIP	0.833	0.722	0.833	0.2778	0.889	0.711
	3DGen-SP(ours)	0.833	0.833	0.889	0.389	0.889	0.767
Image-to-3D	CLIP	0.154	0.00	0.256	0.462	0.180	0.210
	3DGen-SP(ours)	0.897	0.821	0.821	0.692	0.780	0.802

Table R4 Ranking alignment with additional Instant-Mesh. we compute the Kendall’s tau ranking correlation (Kendall, 1938), where a higher value suggests a stronger alignment with human judgment.

Method	Plausibility	Geo. Details	Tex. Quality	Geo-Tex.	Alignment	Average
CLIP	0.209	0.077	0.297	0.473	0.231	0.257
3DGen-SP(Ours)	0.802	0.736	0.846	0.692	0.780	0.771

some dimensions. To address this, we expand our score dataset by incorporating models from OmniObject3D (T. Wu, Zhang, et al., 2023) and Cap3D (Luo et al., 2024). Specifically, we assign full scores for 6k scanned models from OmniObject3D and filter 3.4k high-quality models rated as “good” or higher by GPT-4 from Cap3D (Z. Sun et al., 2024). For each model, we treat its caption as the text prompt and its rendering image as the image prompt. As a result, we obtain approximately 19k high-quality models and 11.2k human-annotated models.

Training Strategy We follow a similar model structure to *3DGen-Score* model and employ a “two-stage” training strategy. In the first stage, called the Encoder-Adaptation stage, we minimize contrastive loss using only the 19k high-quality samples. In the subsequent stage, we minimize MSE loss with supervision from a total of 30k training samples. For implementation, we initialize the encoders from CLIP-ViT-H/14 (Radford et al., 2021) and initialize the learnable predictor heads using Xavier Uniform. And we only unfreeze selected final layers of vision encoders during training, specifically the last 4 layers in this paper. The model is trained with a batch size of 32 and a learning rate of 3e-6, using the Adam optimizer to optimize the training process.

E.2 Evaluation

Alignment with Human As an evaluation task, we prioritize consistency across distributions rather than precise score values of individual models. Therefore, we assess the performance of

the scoring model by computing ranking consistency with human annotations. Specifically, we rank models by calculating the average score over test dataset predicted by CLIP and *3DGen-SP*, respectively, and calculate the Kendall’s tau ranking correlation (Kendall, 1938) to assess alignment with human annotations. As shown in Table R3, our model consistently outperforms CLIP in all tested scenarios.

Robustness to New Model In a similar setting, we further conduct experiments on Instant-Mesh pairs to evaluate the generalization capability of *3DGen-SP* model. As shown in Table R4, our *3DGen-SP* model demonstrates strong robustness against novel models, consistently outperforming CLIP across all dimensions. These results further validate the effectiveness of our training strategy.

F Additional Details about 3DGen-Evaluator

F.1 3D Prior Exploration

We employ 3D point-based MLLMs (Qi et al., 2024; R. Xu et al., 2023) to explore the feasibility of assessing quality directly using 3D representations. The primary process involves evaluating the alignment between ground-truth prompts and generated captions, as shown in Figure S3. First, we extract a colored point cloud from the generated 3D models, typically represented in formats such as Mesh, NeRF, or Gaussian, and organize them into the format (N_points, 6), consisting of xyz coordinates and RGB colors. Next, we prompt

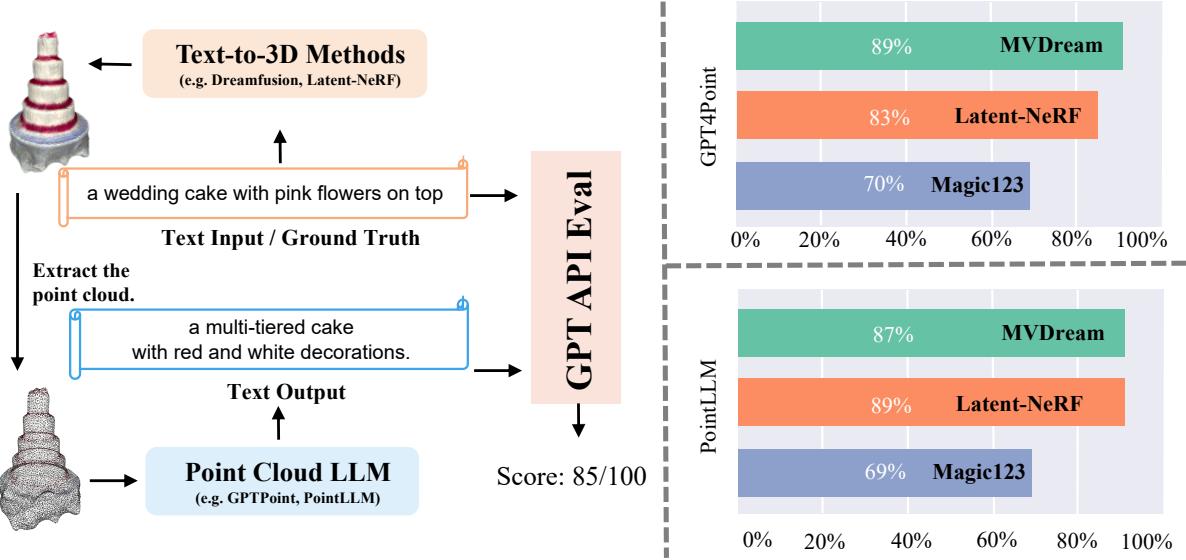


Fig. S3 Evaluation with point-based MLLMs **Left:** An overview of the evaluation pipeline. **Right:** Model rankings on the test data.

Table R5 Pairwise rating alignment compared with $T^3\text{Bench}$. We assess the average probability that our model's decisions align with human judgments for each comparison.

Methods	Plausibility	Geo. Details	Tex. Quality	Geo-Tex.	Alignment	Average
$T^3\text{Bench}$	0.686	0.675	0.693	0.665	0.608	0.665
3DGen-Score(Ours)	0.789	0.714	0.742	0.716	0.747	0.742

the 3D MLLMs with ”*What is the object?*” to generate corresponding captions. These text pairs are then submitted to ChatGPT for human-like assessment. Finally, we sample 100 battle pairs from the test dataset, and the ranking results are presented in Figure S3.

F.2 Comparison with $T^3\text{Bench}$

Since $T^3\text{Bench}$ (Y. He et al., 2023) has stricter requirements on the input format, converting the format for some models in the 3DGS representation proved challenging. So we only conduct experiments on five models that are supported: Dreamfusion (Poole et al., 2022), Latent-NeRF (Metzger et al., 2023), Mvdream (Y. Shi et al., 2023), SJC (H. Wang et al., 2023) and Magic3D (Lin et al., 2023). We compare pairwise alignment with human on our test dataset, as $T^3\text{Bench}$ has not provided or released validation data. The results, shown in Table R5, reveal that our scoring model consistently outperforms T3Bench across

all dimensions, demonstrating its impressive ability to simulate human preferences.

F.3 Prompt Design for 3DGen-Eval

Prompts for GPT-4V To generate pseudo labels for training the *3DGen-Eval* model, we prompt GPT-4V to produce detailed descriptions for each 3D model and well-reasoned explanations for each dimension to justify the assigned scores. The template used for *Absolute Scores* and *Comparison Votes* are shown in Figure S4 and Figure S5, respectively.

Templates for Instruct Tuning We then structure the responses generated from GPT-4V into templates compatible with LLaVA (H. Liu et al., 2024) for better understanding and processing. Example diagrams for *Abosolute Score* and *Comparison Votes* are shown in Figure S6 and Figure S7, respectively.

There are two quad images in the NORMAL and RGB format generated from rendering a SINGLE 3D model FROM FOUR DIFFERENT views, where the NORMAL rendering can show geometric features better. If there is an expert 3D model evaluation system and the score ranges from 1 to 5, representing the overall quality of the model from low to high.

Known that the current model is scored as {int(avg_score)}, I would like you to generate a description of the current model.

After description generation, I hope you to give some reasoning explanations for following refined dimensions.

- **Geometry Plausibility:** Score ranging from 1 to 10, assess whether the 3D model's overall shape and structure match the real object logically and physically. A plausible 3D asset should avoid improbable features, such as distorted faces (Janus problem) or floating, noisy geometry.
- **Geometry Details:** Score ranging from 1 to 5, assess shape's fineness and exquisiteness. High-detail models feature complex surface features, while low-detail models may look rough and simplistic. Pay attention to the distinction between details and noises.
- **Texture Quality:** Score ranging from 1 to 5, underscore both aesthetic quality and view-consistency. As the former encompasses aspects such as realistic representation of materials and appropriate coloring, the latter prioritizes uniformity and seamlessness across various viewing angles.
- **Geometry-Texture Coherency:** Score ranging from 1 to 2, assess the alignment between geometric features and texture features. Texture maps should neither make up nor cover up geometric details.
- **Prompt-Asset Alignment:** Score ranging from 1 to 5, assess the alignment between 3D model and its prompt, involving matches about category, quantity, attributes or spatial relationships possibly described in prompt.

For all above dimensions, the higher the score, the higher the quality on that dimension. And for each 3D model, multi-dimension scores are provided in the LIST format, written as `score_list`. Known that the `score_list` of the current 3D model is `[{score for score in score_dict.values()}]` and its visual prompt is `{prompt}`, please generate some reasoning explanations for each dimensions.

Fig. S4 Prompts for GPT-4V to generate descriptions and explanations of *Absolute score* for single 3D model, where the `{score_dict}` is score annotations of 5 dimensions, and the `{avg_score}` is averaged from `{score_dict}`.

There are two sets of quad images rendered from two 3D models generated from the SAME prompt. For each image set, there are two images respectively in the NORMAL and RGB format rendered from a SINGLE 3D model FROM FOUR DIFFERENT views, where the NORMAL rendering displays geometry structure and details clearly, and the RGB rendering shows color, material, overall style and other texture details.

If there is an expert 3D model evaluation system, it compares these two 3D models from the following dimensions.

- **Geometry Plausibility:** Assess whether the 3D model's overall shape and structure match the real object logically and physically. A plausible 3D asset should avoid improbable features, such as distorted faces (Janus problem) or floating, noisy geometry.
- **Geometry Details:** Assess shape's fineness and exquisiteness. High-detail models feature complex surface features, while low-detail models may look rough and simplistic.
- **Texture Quality:** Underscore both aesthetic quality and view-consistency. As the former encompasses aspects such as realistic representation of materials and appropriate coloring, the latter prioritizes uniformity and seamlessness across various viewing angles.
- **Geometry-Texture Coherency:** Assess the alignment between geometric features and texture features. Texture maps should neither make up nor cover up geometric details.
- **Prompt-Asset Alignment:** Assess the alignment between 3D model and its prompt, involving matches about category, quantity, attributes or spatial relationships possibly described in prompt.

The comparison results are selected from [0, 1, 0.5] and collected in a LIST format, where "0" represents "model_0 is better", "1" represents "model_1 is better" and "0.5" represents "tie(equally well or poorly)".

Now, known that the [task] prompt is {prompt} and the comparison results are `[float(winner_dict[dim]) for dim in COMP_DIMS]`, please generate some reasoning explanations for each dimension.

Fig. S5 Prompts for GPT-4V to generate explanations of *Pairwise comparison* for two 3D models generated from the same prompt, where the `{prompt}` is the corresponding prompt input and the `{winner_dict}` represents comparison annotations of five dimensions.



Fig. S6 Prompts for 3DGen-Eval model's Instruct-Tuning stage(a) trained with *Absolute scores*, where the `{score}` is supervised by the human annotations, and the `{explanation}` is supervised by GPT-4V's responses.



Fig. S7 Prompts for 3DGen-Eval model's Instruct-Tuning stage(b) trained with *Comparison votes*.

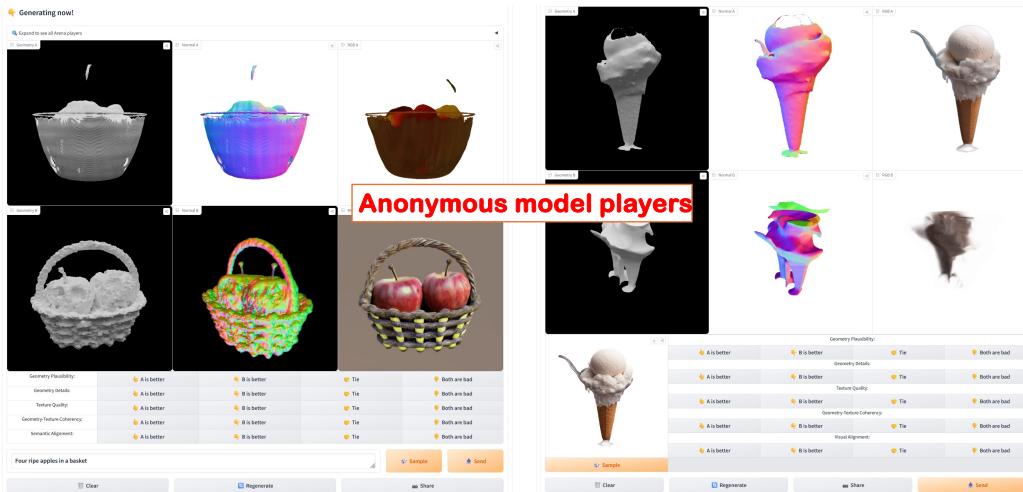


Fig. S8 3DGen-Arena Page1: Anonymous battle models

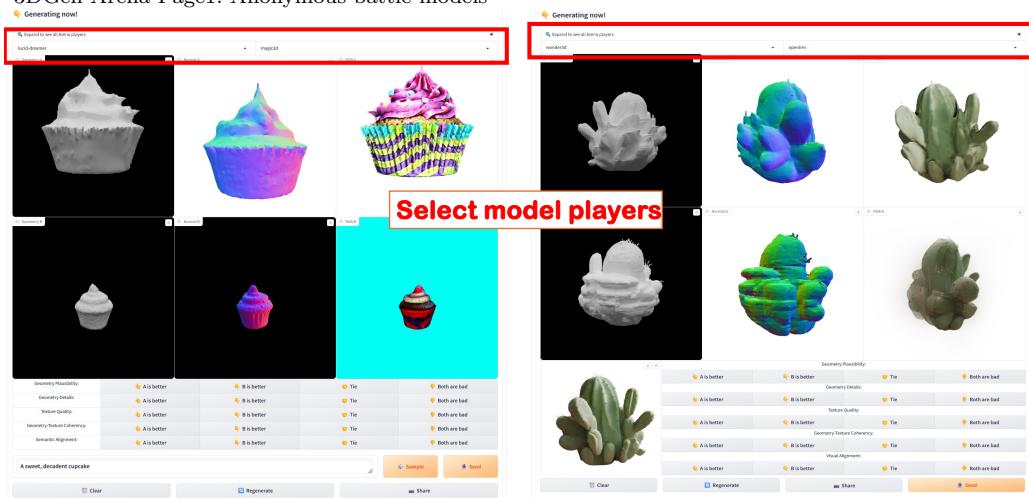


Fig. S9 3DGen-Arena Page2: Named battle models

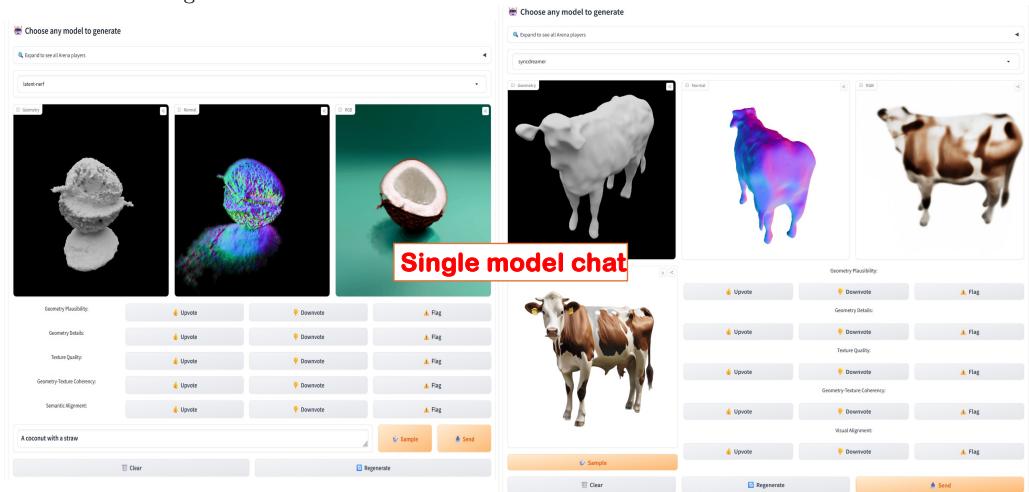


Fig. S10 3DGen-Arena Page3: Single model chat

“Comparison votes” Annotation platform

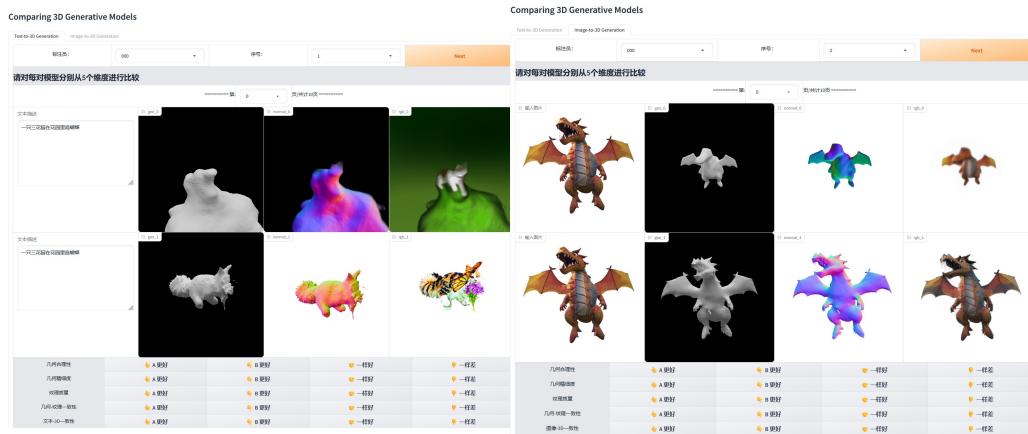


Fig. S11 Screenshots of annotation platform for Comparison Votes

“Absolute scores” Annotation platform



Fig. S12 Screenshots of annotation platform for Absolute Scores

References

- Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., ... others (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, ,
- Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., ... Lin, D. (2023). Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, ,
- Chen, R., Chen, Y., Jiao, N., Jia, K. (2023). Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *Proceedings of the ieee/cvf international conference on computer vision* (pp. 22246–22256).
- Chen, Y., Wang, T., Wu, T., Pan, X., Jia, K., Liu, Z. (2024). Comboverse: Compositional 3d assets creation using spatially-aware diffusion guidance.
- Chen, Z., Wang, F., Liu, H. (2023). Text-to-3d using gaussian splatting. *CoRR, abs/2309.16585*, ,
- Cheng, Y.-C., Lee, H.-Y., Tuyakov, S., Schwing, A., Gui, L. (2023). SDFusion: Multimodal 3d shape completion, reconstruction, and generation. *Cvpr*.
- Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., ... others (2024). Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems, 36*, ,
- Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., ... Farhadi, A. (2023). Objaverse: A universe of annotated 3d objects. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 13142–13153).
- Di, D., Yang, J., Luo, C., Xue, Z., Chen, W., Yang, X., Gao, Y. (2024). Hyper-3dg: Text-to-3d gaussian generation via hypergraph. *CoRR, abs/2403.09236*, ,
- Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., ... Vanhoucke, V. (2022). Google scanned objects: A high-quality dataset of 3d scanned household items.
- Elo, A.E. (1967). The proposed uscf rating system, its development, theory, and applications. *Chess life*, 22(8), 242–247,
- Ethayarajh, K., Choi, Y., Swayamdipta, S. (2022). Understanding dataset difficulty with \mathcal{V} -usable information. *International conference on machine learning* (pp. 5988–6008).
- Gao*, R., Holynski*, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P.P., ... Poole*, B. (2024). Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv*, ,
- Guo, Y.-C., Liu, Y.-T., Shao, R., Laforte, C., Voleti, V., Luo, G., ... Zhang, S.-H. (2023). threestudio: A unified framework for 3d content generation. <https://github.com/threestudio-project/threestudio>.
- Gupta, A., Xiong, W., Nie, Y., Jones, I., Oğuz, B. (2023). 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*, ,
- He, Y., Bai, Y., Lin, M., Zhao, W., Hu, Y., Sheng, J., ... Liu, Y.-J. (2023). T³ bench: Benchmarking current progress in text-to-3d generation. *arXiv preprint arXiv:2310.02977*, ,
- He, Z., & Wang, T. (2023). Openlrm: Open-source large reconstruction models. <https://github.com/3DTopia/OpenLRM>.

- Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y. (2022). *Clipscore: A reference-free evaluation metric for image captioning*.
- Hong, F., Tang, J., Cao, Z., Shi, M., Wu, T., Chen, Z., ... Liu, Z. (2024). 3dtopia: Large text-to-3d generation model with hybrid diffusion priors. *arXiv preprint arXiv:2403.02234*, ,
- Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., ... Tan, H. (2023). Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, ,
- Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., ... Schmidt, L. (2021, July). *Openclip*. Zenodo. Retrieved from <https://doi.org/10.5281/zenodo.5143773> (If you use this software, please cite it as below.)
- Jun, H., & Nichol, A. (2023). Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, ,
- Kendall, M.G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2), 81–93,
- Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G. (2023). 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 1–14,
- Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O. (2023). Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 36652–36663,
- Kong, X., Liu, S., Lyu, X., Taher, M., Qi, X., Davison, A.J. (2024). Eschernet: A generative model for scalable view synthesis. *arXiv preprint arXiv:2402.03908*, ,
- Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z.R., Stevens, K., ... others (2024). Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36, ,
- Ku, M., Li, T., Zhang, K., Lu, Y., Fu, X., Zhuang, W., Chen, W. (2023). Imagenhub: Standardizing the evaluation of conditional image generation models. *arXiv preprint arXiv:2310.01596*, ,
- Li, J., Li, D., Xiong, C., Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Icml*.
- Li, J., Tan, H., Zhang, K., Xu, Z., Luan, F., Xu, Y., ... Bi, S. (2023). Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, ,
- Li, W., Chen, R., Chen, X., Tan, P. (2023). Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. *arxiv:2310.02596*, ,
- Liang, Y., Yang, X., Lin, J., Li, H., Xu, X., Chen, Y. (2023). Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. *arXiv preprint arXiv:2311.11284*, ,
- Lin, C.-H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., ... Lin, T.-Y. (2023). Magic3d: High-resolution text-to-3d content creation. *Proceedings of the ieee/cvpr conference on computer vision and pattern recognition* (pp. 300–309).
- Liu, F., Wu, D., Wei, Y., Rao, Y., Duan, Y. (2023). *Sherpa3d: Boosting high-fidelity text-to-3d generation via coarse 3d prior*.

- Liu, H., Li, C., Wu, Q., Lee, Y.J. (2024). Visual instruction tuning. *Advances in neural information processing systems*, 36, , *arXiv preprint arXiv:2404.07984*, ,
- Liu, M., Shi, R., Chen, L., Zhang, Z., Xu, C., Wei, X., ... Su, H. (2023). One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. *arXiv preprint arXiv:2311.07885*, ,
- Liu, M., Shi, R., Kuang, K., Zhu, Y., Li, X., Han, S., ... Su, H. (2024). Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in neural information processing systems*, 36, ,
- Liu, M., Xu, C., Jin, H., Chen, L., Varma T, M., Xu, Z., Su, H. (2024). One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, ,
- Liu, R., Wu, R., Hoorick, B.V., Tokmakov, P., Zakharov, S., Vondrick, C. (2023). Zero-1-to-3: Zero-shot one image to 3d object.
- Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W. (2023). Sync-dreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, ,
- Long, X., Guo, Y.-C., Lin, C., Liu, Y., Dou, Z., Liu, L., ... others (2023). Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, ,
- Lu, Y., Jiang, D., Chen, W., Wang, W., Choi, Y., Lin, B.Y. (2024, February). *Wildvision arena: Benchmarking multimodal llms in the wild*. Retrieved from <https://huggingface.co/spaces/WildVision/vision-area/>
- Luo, T., Johnson, J., Lee, H. (2024). View selection for 3d captioning via diffusion ranking.
- Ma, B., Deng, H., Zhou, J., Liu, Y.-S., Huang, T., Wang, X. (2023). Geodream: Disentangling 2d and geometric priors for high-fidelity and consistent 3d generation. *arXiv preprint arXiv:2311.17971*, ,
- Melas-Kyriazi, L., Laina, I., Rupprecht, C., Neverova, N., Vedaldi, A., Gafni, O., Kokkinos, F. (2024). IM-3D: iterative multiview diffusion and reconstruction for high-quality 3d generation. *CoRR*, *abs/2402.08682*, ,
- Metzler, G., Richardson, E., Patashnik, O., Giryes, R., Cohen-Or, D. (2023). Latent-nerf for shape-guided generation of 3d shapes and textures. *Proceedings of the ieee/cvpr conference on computer vision and pattern recognition* (pp. 12663–12673).
- mrfakename, Srivastav, V., Fourrier, C., Pouget, L., Lacombe, Y., main, Gandhi, S. (n.d.). *Text to speech arena*. <https://huggingface.co/spaces/TTS-AGI/TTS-Arena>. Hugging Face.
- Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., ... Schulman, J. (2021). Webgpt: Browser-assisted question-answering with human feedback. *arxiv*.
- Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M. (2022). Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, ,
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... others (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35, 27730–27744,
- Poole, B., Jain, A., Barron, J.T., Mildenhall, B. (2022). Dreamfusion: Text-to-3d using 2d

- diffusion. *arXiv preprint arXiv:2209.14988*, ,
- Qi, Z., Fang, Y., Sun, Z., Wu, X., Wu, T., Wang, J., ... Zhao, H. (2024). Gpt4point: A unified framework for point-language understanding and generation. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 26417–26427).
- Qian, G., Mai, J., Hamdi, A., Ren, J., Siarohin, A., Li, B., ... Ghanem, B. (2024). Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *The twelfth international conference on learning representations (iclr)*. Retrieved from <https://openreview.net/forum?id=0jHkUDyEO9>
- Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jägersand, M. (2020). U²-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognit.*, *106*, 107404,
- Radford, A., Kim, J., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021, Feb.). Learning transferable visual models from natural language supervision. *Cornell University - arXiv, Cornell University - arXiv*, ,
- Raj, A., Kaza, S., Poole, B., Niemeyer, M., Ruiz, N., Mildenhall, B., ... others (2023). Dreambooth3d: Subject-driven text-to-3d generation. *Proceedings of the ieee/cvf international conference on computer vision* (pp. 2349–2359).
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B. (2022, Jun.). High-resolution image synthesis with latent diffusion models. *2022 ieee/cvf conference on computer vision and pattern recognition (cvpr)*. Retrieved from <http://dx.doi.org/10.1109/cvpr52688.2022.01042>
- Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., ... others (2022). Photo-realistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, *35*, 36479–36494,
- Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., ... Jitsev, J. (2022). LAION-5b: An open large-scale dataset for training next generation image-text models. *Thirty-sixth conference on neural information processing systems datasets and benchmarks track*. Retrieved from <https://openreview.net/forum?id=M3Y74vmsMcY>
- Shi, R., Chen, H., Zhang, Z., Liu, M., Xu, C., Wei, X., ... Su, H. (2023). Zero123++: a single image to consistent multi-view diffusion base model. *CoRR*, *abs/2310.15110*, ,
- Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X. (2023). Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, ,
- Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., ... Christiano, P. (2020, Sep.). Learning to summarize from human feedback. *arXiv: Computation and Language, arXiv: Computation and Language*, ,
- Sun, J., Zhang, B., Shao, R., Wang, L., Liu, W., Xie, Z., Liu, Y. (2023). Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. *arXiv preprint arXiv:2310.16818*, ,
- Sun, Z., Wu, T., Zhang, P., Zang, Y., Dong, X., Xiong, Y., ... Wang, J. (2024). Bootstrap3d: Improving 3d content creation with synthetic data. *arXiv preprint arXiv:2406.00093*, ,

- Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z. (2024). Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, ,
- Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G. (2023). Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, ,
- Tang, J., Wang, T., Zhang, B., Zhang, T., Yi, R., Ma, L., Chen, D. (2023). Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *International conference on computer vision ICCV*.
- Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., ... Hashimoto, T.B. (2023). Stanford alpaca: An instruction-following llama model. <https://github.com/tatsu-lab/stanford.alpaca>. GitHub.
- Tremblay, J., Meshry, M., Evans, A., Kautz, J., Keller, A., Khamis, S., ... Birchfield, S. (2022). Rtmv: A ray-traced multi-view synthetic dataset for novel view synthesis. *IEEE/CVF European Conference on Computer Vision Workshop (Learn3DG ECCVW), 2022*, ,
- Voleti, V., Yao, C., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., ... Jampani, V. (2024). SV3D: novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *CoRR, abs/2403.12008*, ,
- Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G. (2023). Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 12619–12629).
- Wang, T., Zhang, B., Zhang, T., Gu, S., Bao, J., Baltrusaitis, T., ... Guo, B. (2023). Rodin: A generative model for sculpting 3d digital avatars using diffusion. *CVPR*, ,
- Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N.A., Khashabi, D., Hajishirzi, H. (2022). Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, ,
- Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J. (2024). Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems, 36*, ,
- Wang, Z., Wang, T., Hancke, G., Liu, Z., Lau, R.W. (2024). Themestation: Generating theme-aware 3d assets from few exemplars. *Acm siggraph 2024 conference papers* (pp. 1–12).
- Wu, T., Wang, J., Pan, X., Xu, X., Theobalt, C., Liu, Z., Lin, D. (2023). Voxurf: Voxel-based efficient and accurate neural surface reconstruction. *International conference on learning representations (iclr)*.
- Wu, T., Yang, G., Li, Z., Zhang, K., Liu, Z., Guibas, L., ... Wetzstein, G. (2024). Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 22227–22238).
- Wu, T., Zhang, J., Fu, X., Wang, Y., Jiawei Ren, L.P., Wu, W., ... Liu, Z. (2023). Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. *Ieee/cvf conference on computer vision and pattern recognition (cvpr)*.
- Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H. (2023). Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, ,

- Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S., Shan, Y. (2024). Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, , diffusion guidance.
- Zou, Z.-X., Yu, Z., Guo, Y.-C., Li, Y., Liang, D., Cao, Y.-P., Zhang, S.-H. (2023). Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. *arXiv preprint arXiv:2312.09147*, ,
- Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., ... Dong, Y. (2024). Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, ,
- Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., Lin, D. (2023). Pointllm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*, ,
- Xu, Y., Shi, Z., Yifan, W., Peng, S., Yang, C., Shen, Y., Gordon, W. (2024). Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arxiv:2403.14621*, ,
- Ye, J., Liu, F., Li, Q., Wang, Z., Wang, Y., Wang, X., ... Zhu, J. (2024). Dreamreward: Text-to-3d generation with human preference.
- Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., ... Shen, C. (2023). Metric3d: Towards zero-shot metric 3d prediction from a single image. *Proceedings of the ieee/cvpr international conference on computer vision* (pp. 9043–9053).
- Zheng, C., & Vedaldi, A. (2023). Free3d: Consistent novel view synthesis without 3d representation. *arXiv*, ,
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... others (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 46595–46623,
- Zhu, J., & Zhuang, P. (2023). Hifa: High-fidelity text-to-3d generation with advanced