Gen3DEval: Using vLLMs for Automatic Evaluation of Generated 3D Objects

Shalini Maiti[*†] Lourdes Agapito[†] Filippos Kokkinos[*]

[*]Meta AI [†]University College London

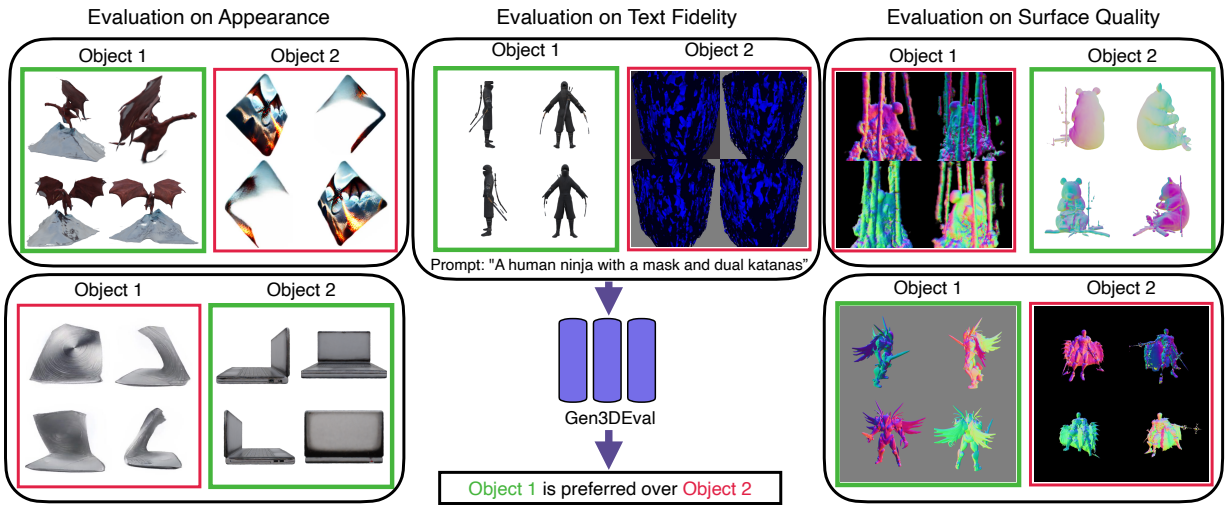

Figure 1. **Gen3DEval**: A holistic ranking metric to assess the quality of generated 3D objects on appearance, surface quality and text fidelity using a vision large language model (vLLM) which is trained to choose the better out of two objects on the three evaluation dimensions (appearance, text fidelity or surface quality).

Abstract

Rapid advancements in text-to-3D generation require robust and scalable evaluation metrics that align closely with human judgment, a need unmet by current metrics such as PSNR and CLIP, which require ground-truth data or focus only on prompt fidelity. To address this, we introduce Gen3DEval, a novel evaluation framework that leverages vision large language models (vLLMs) specifically fine-tuned for 3D object quality assessment. Gen3DEval evaluates text fidelity, appearance, and surface quality by analyzing 3D surface normals, without requiring ground-truth comparisons, bridging the gap between automated metrics and user preferences. Compared to state-of-the-art task-agnostic models, Gen3DEval demonstrates superior performance in user-aligned evaluations, placing it as a comprehensive and accessible benchmark for future research on text-to-3D generation. The project page can be found here: https://shalini-maiti.github.io/gen3deval.github.io/.

1. Introduction

The domain of text-to-3D generation has advanced significantly in recent years, driven by the rise of scalable architectures like diffusion models [43], neural radiance fields (NeRF) [36], and Gaussian splatting [25]. However, the field lacks standardized, human-aligned evaluation metrics that can reliably assess these assets and the methods that produce them. Existing metrics—such as CLIP [42] scores evaluate only limited aspects of the output like text fidelity and similarity-based measures like Peak-Signal-To-Noise Ratio (PSNR), SSIM [55], Chamfer Distance, and Fréchet Inception Distance (FID) [20] depend on ground-truth data making them inadequate and impractical for text-to-3D generation, where diverse outputs may correspond to a single prompt. In such cases, a unique, universally applicable reference does not exist, as multiple plausible 3D outputs can vary widely in style, appearance, and fidelity to the text. Meanwhile, FID computes a distributional similarity, which poses other challenges. Currently, there is no standardized large-scale validation set to serve as the ground-truth distri-

bution for 3D assets, making FID computation difficult and inconsistent. Moreover, generating sufficient 3D assets to estimate this distribution requires significant computational resources, casting FID as an expensive and time-intensive metric. As a result, these metrics fall short of capturing the nuanced requirements of evaluating text-to-3D generation, where a scalable, human-aligned approach is essential.

While prior work such as GPT4VEval [57] has leveraged GPT-4V [38] for assessing 3D asset quality, GPT-4V is a general-purpose model not specifically trained for 3D quality assessment, which limits its effectiveness in this domain. Furthermore, it can be costly to deploy at scale, and in our experiments, we found that GPT-4o (which is the successor to GPT-4V) performed significantly worse than our method in aligning with human judgments of 3D asset quality.

To bridge this gap, we introduce Gen3DEval, a vision-based large language model (vLLM) framework specifically fine-tuned to evaluate text-to-3D generation outputs in alignment with human preferences. Unlike existing metrics [42, 55, 63], Gen3DEval assesses not only text fidelity but also appearance and surface quality by analyzing rendered multi-view images. Supporting up to eight images as input, Gen3DEval enables comprehensive assessment by leveraging multi-view renderings, such as RGB and normal maps, of generated 3D objects. Using multi-view images as input allows for compatibility across diverse 3D representations [25, 36, 60].

Built upon the recent vLLM early fusion approaches [4, 5, 27, 29], our method processes these input renderings by first encoding each image through an image encoder, which translates them into visual tokens. These tokens are then integrated with text tokens and fed into a Llama3 model, allowing Gen3DEval to interpret both visual and textual features of the 3D objects holistically. To ensure robust performance, we curate data from human assessments and further enhance our training dataset with synthetically generated perturbations of artist-created 3D objects, incorporating artifacts like floaters, transparency errors, text fidelity inconsistencies, excessively smooth surfaces etc.

A key component of our framework is Gen3DEval-Bench, a benchmark dataset designed to standardize text-to-3D evaluations across various quality dimensions. Comprising 80 diverse prompts, Gen3DEval-Bench facilitates consistent, human-aligned assessments of visual fidelity and aesthetic preferences. Our evaluation pipeline involves two main stages: first, it performs pairwise comparisons of 3D objects using multi-view renderings. Then it applies ELO rating metrics [13] to generate scores that closely align with human judgment. This process ensures robust and reliable evaluations across a broad range of 3D generative methods. To sum up, our contributions are:

- A state-of-the-art holistic evaluation method for text-to-3D generation that ranks methods across appearance, surface quality and text fidelity.
- A vLLM fine-tuned on the Llama3 [2] model, using a synthetic dataset curated to reflect human preferences for evaluating generated 3D assets.
- A benchmark dataset, Gen3DEval-Bench, comprising 80 prompts for ranking existing and future text-to-3D generation methods in a standardized manner.

2. Related Work

Text-to-3D and Image-to-3D generation. In recent times, the landscape of text-to-3D generation has seen rapid growth with the advent of representations such as Neural Radiance Fields [36], occupancy fields [37], SDF [60] and Gaussian Splats [25], and the availability of large, publicly available datasets such as Objaverse [10, 11]. Some of the earlier methods in the space currently include [23, 28, 40, 41, 47, 48, 53, 56, 61, 64] that optimizes a randomly-initialized 3D model via gradient descent conditioned on sampled outputs of a text-to-image generation model. Another direction of work include methods such as [14, 21, 24, 31, 45, 46, 54] that use multi-view diffusion models to fine-tune text-to-image models to quickly generate highly consistent multiple views or videos simultaneously from a single input image. Notably, another family of methods [15, 22, 47] learns 3D priors from a large amount of data and a scalable architecture to directly output robust 3D outputs from text or image inputs. With such an impressive pace of growth in this research domain, it is imperative for the presence of robust evaluation metrics and benchmarks to ensure continued progress in this field, which is a gap that Gen3DEval attempts to bridge.

3D Evaluation Metrics and benchmarks Classical 3D metrics like PSNR, Chamfer Distance, LPIPS [63] and SSIM [55] were developed to measure the quality of a generated 3D asset against a ground-truth asset. However, these are similarity metrics and measure the distance between generated and the ground-truth data. This is infeasible since a single text prompt can be mapped to many generated 3D outputs, with their quality or fidelity being independent of their similarity to any single generated asset. We propose that instead of measuring similarity, we need to inject 3D quality priors into the method itself for the purpose of evaluation, which is the foundation of Gen3DEval.

Other metrics such as CLIP [42] scores have tried to measure alignment with text by using a standard benchmark of textual prompts and computing a corresponding score. However, they underwhelm with an increase in diversity of prompts. Another inadequacy is that only text-alignment is measured and not appearance or surface quality. Gen3DEval addresses both of these aspects. The work closest to ours in attempting to solve this problem is GPT-4V(ision) is a Human-Aligned Evaluator for Text-to-3D Generation [57]. However, it uses GPT4 [38] off the shelf,

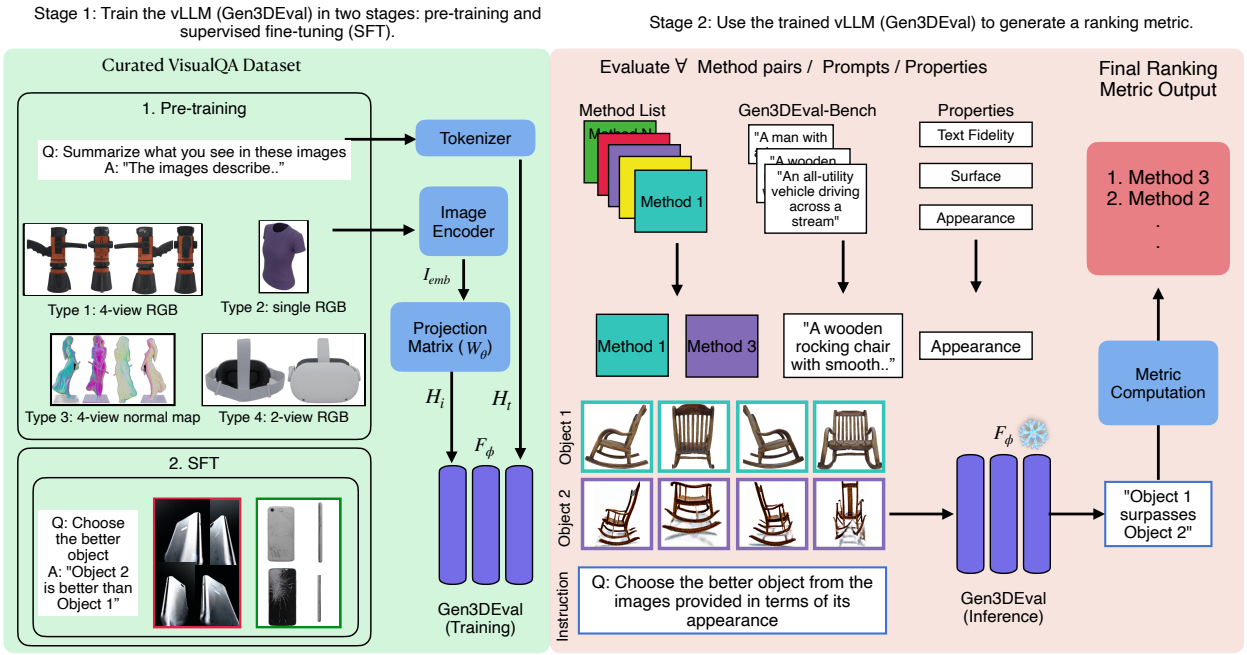

Figure 2. **Gen3DEval framework:** In stage 1, we train a vLLM to choose which object is better in terms of appearance, surface quality or text fidelity. This is further divided into 2 parts. In pre-training, we train the vision-to-language projector using image summary VQA. In the supervised fine-tuning (SFT) stage, we use comparison data to train for instruction following and preference evaluation. In stage 2, we compute a ranking metric for the set of methods by applying the trained vLLM from stage 1 pairwise on Gen3DEval-Bench prompts.

which is a task-agnostic vLLM trained on half a trillion parameters whose API and checkpoints are not publicly available, making it costly to scale, whereas Gen3DEval has been *specifically* trained to evaluate text-to-3D objects, on 8.35 billion parameters and will be made available for public usage. Moreover, in our experiments, we found that Gen3DEval performed significantly better than GPT4-o in aligning with human judgments of 3D asset quality.

In the absence of quantifiable metrics, user studies have been popularly employed as the gold standard to evaluate 3D generation methods. However, this is time-consuming, cost-ineffective and lacks a standard procedure. Certain benchmarks such as T^{3}Bench [17] and Dreamfusion prompts [40] have been created to reduce the lack of standardization in this process. Gen3DEval takes a step further in this direction by curating a benchmark Gen3DEval-Bench with diverse prompts in terms of types objects, length and compositionality.

Large Multi-modal Models The past couple of years has seen great strides made in the development of Large languages models (LLMs) like Llama [52], GPT-4 [38], Claude [3], Gemini [49], and consequently, led to development of vLLMs such as LLaVA [29], BLIP [27], FUYU [5] and more [2, 3, 38]. They are powerful multi-modal models that display strong image and language reasoning. However, since these are general purpose models, they do not perform well on evaluating generated 3D objects. [57] showcases capabilities of GPT4 [38] to be able to align with human preference for the assessment of 3D objects. Gen3DEval takes this effort further by fine-tuning a Llama3 [2] model using a curated synthetic dataset for the specific purpose of introducing 3D aesthetic preference into the vision-language space and transforming that into a ready-to-use evaluation ranking metric.

3. Method

The proposed method, Gen3DEval is a vision-based large language model (vLLM) that interprets and assesses the quality of 3D generated objects. We train Gen3DEval using a carefully curated Visual-Question-Answering (VQA) dataset, as detailed in Section 4.1. This training enables Gen3DEval to learn associations between visual cues in multi-view images and quality indicators such as text fidelity, surface detail, and appearance. We use up to eight multi-view RGB images—renderings that include RGB and normal maps—to capture comprehensive object details from multiple angles; see Figure 2 for an overview.

3.1. Model Details

The Gen3DEval training process builds upon the LLaVA architecture [29] and is organized into two sequential phases: pre-training and supervised fine-tuning. In the initial phase, each image is processed by an image encoder to produce visual embeddings I_{emb}. These embeddings are transformed

into language-compatible tokens H_i via a linear projection matrix W_θ, enabling them to integrate seamlessly with the language-based representations. At the same time, a language tokenizer converts the natural language Question-Answer (QA) pairs into text tokens H_t. The model F_ϕ then receives both the image and text tokens as input, learning to predict the next token H_y by maximizing the likelihood of the correct token.

For the image encoder, we initially consider CLIP [42], which is commonly used across various vision-language models [29, 40, 44]. However, since the input images are rendered views of 3D objects, we also evaluate two additional encoders: DinoV2 [39] and Fit3D [62]. DinoV2 generates visually consistent embeddings, while Fit3D is specifically designed to encode 2D images into features consistent with the underlying 3D scenes, making it particularly suited to our task. A comparison of these feature encoders is provided in Table 1 and discussed in Section 4.2.

During the pre-training phase, the weights of both the LLM and the image encoder are frozen, and only the weights of the linear projection matrix W_θ are updated. This selective tuning establishes alignment between the visual embeddings and language tokens, forming a foundation for integrated visual-language comprehension.

In the fine-tuning phase, we unfreeze both the projection matrix and the LLM, allowing them to be fine-tuned jointly, while keeping the image encoder's weights frozen. This stage further specializes the model for 3D quality assessment, enhancing its sensitivity to features such as surface texture, text fidelity, and overall visual coherence across various prompts and multi-view renderings.

3.2. Multi-view Input

To effectively evaluate the quality of generated 3D objects, Gen3DEval leverages multi-view input, using up to 8 rendered images uniformly panning each object. This approach is essential for capturing the complete appearance and surface consistency of 3D objects, as single-view images may overlook aspects like hidden surfaces and occlusions that become visible when observed from multiple viewpoints.

In pre-training, the input images range from a set of 1 to 4 RGB images or rendered surface normals panning the object in a 360° round-table manner alongside a short summary in a QA pair. In fine-tuning, we input two sets of images, for object 1 and object 2. Each set consists upto 4 multi-view images each, therefore training a VQA sample consists upto 8 images and QA capturing the preference for the preferred object. The number of tokens per image is 576 and approx. 250 for the text of the question.

4. Training Details

The following section provides an in-depth look at the Gen3DEval dataset, outlining its composition, structure, and the methodologies used to create a diverse and robust training set. We describe the dataset's sources and organization, the approach to rendering multi-view images, and the use of human judgment and synthetic perturbations to enhance model alignment with 3D quality standards. Each component of the dataset is designed to support the pre-training and fine-tuning stages, ensuring comprehensive coverage of key attributes like appearance, surface consistency, and text fidelity.

4.1. Dataset

Gen3DEval's training dataset is designed to train and evaluate the model's ability to assess 3D object quality across various dimensions, including appearance, surface consistency, and text fidelity. It comprises three subsets a) 3D artist-created meshes, b) human preference data on generative method outputs and c) synthetic 3D comparison data.

3D Meshes: Comprising 140,000 high-quality 3D meshes created by artists, this internal dataset spans diverse semantic categories and provides a robust foundation for generalizing to different types of 3D content. Each asset comes with an accompanying text caption generated with Llama3.2 [2]. We render each 3D asset from multiple viewpoints, creating three types of visual inputs: RGB images, alpha masks, and surface normals. These multi-view renderings allow Gen3DEval to capture comprehensive visual cues necessary for accurate 3D evaluation.

Human Annotations: To account for the nuances and irregularities inherent in 3D generative methods, we conducted a large-scale human preference study, collecting over 5K comparative data points across 13 different 3D generative methods [6, 8, 15, 16, 18, 30, 32–34, 46, 47, 51, 54]. Annotators viewed 360° videos of two 3D assets side-by-side and selected the preferred asset based on appearance and alignment with the corresponding generation prompt.

This preference data was incorporated into the Gen3DEval training dataset, enhancing the model's alignment with human aesthetics and text fidelity expectations. We conducted an in-depth analysis of these annotations to identify common artifacts in modern text-to-3D generation methods, including (1) disconnected components, (2) Janus artifacts, (3) opacity inconsistencies, (4) floating elements, (5) overly smooth or irregular surfaces, (6) texture seams. Based on these insights, we replicated these artifacts in the 3D mesh data to scale up our dataset and further improve Gen3DEval 's performance. Examples of these artifacts are illustrated in the appendix.

Synthetic Data: To expand the training dataset, we applied controlled perturbations using Blender, NeRF, and Gaussian splatting techniques, simulating common artifacts and misalignments found in text-to-3D generative outputs. For 3D meshes, we introduced perturbations such as Laplacian smoothing [19], beveling, random surface extrusions,

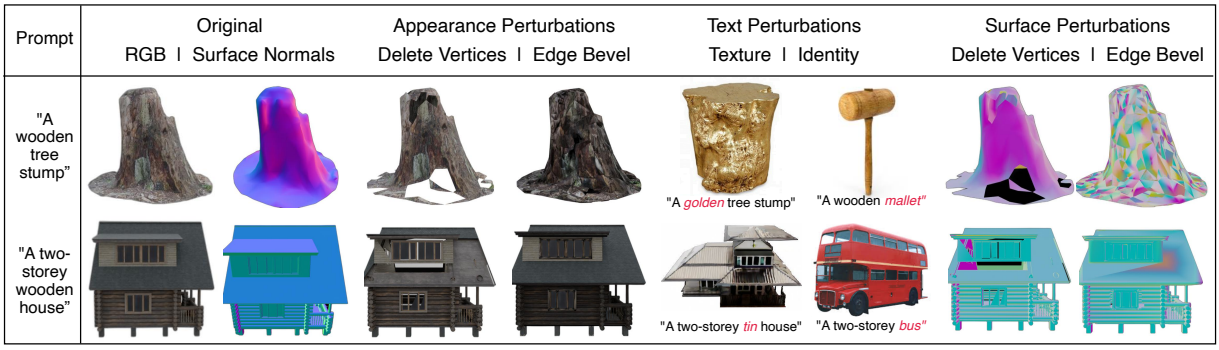

Figure 3. **Training Dataset** We use single and multi-view RGB and surface normals renderings of a 3D object generated from a prompt. We take these objects and perturb them to simulate common appearance, surface and text-related artefacts in generative 3D methods.

and texture map alterations like blurring and seam introduction. Additionally, we fitted NeRF and Gaussian splats to the renderings of the artist-created 3D assets, providing a broader foundation for synthetic training data.

To further enrich this dataset, we introduced additional perturbations to the NeRF and Gaussian splatting by manually adding transparency artifacts, floating elements, and disconnected components, mimicking frequent issues observed in text-to-3D generative methods.

To scale the text fidelity comparison dataset, we used the multi-view video diffusion model from IM-3D [33] and trellis [58] to generate single and multiple views of 3D objects with varied captions, focusing on changes to appearance attributes and composition of the objects. Textual perturbations were created with Llama3.2 [2] by prompting the model to modify the original captions, introducing subtle variations. This approach allowed us to generate a large, diverse synthetic dataset tailored for text fidelity comparison. To ensure high relevance and quality, we applied CLIP [42] to filter out examples with low image-text similarity, resulting in a refined synthetic dataset for evaluating text fidelity. Figure 3 showcases some samples of our SFT dataset.

4.2. Image Encoders

We evaluated 3 image encoders: CLIP [42], DinoV2 [39], Fit3D [62] and the combinations of CLIP with DinoV2 and CLIP with Fit3D; reshaping them to match the bigger of the two and adding the values. For CLIP embeddings, the effective resolution is 336x336 pixels and for both Dinov2 and Fit3D, which internally fine-tune the Dinov2 base model to introduce 3D awareness to image features, the model uses a ViT [12] backbone of patch size 14, effective resolution of 224x224 and embedding dimension 768. The results of these ablation experiments are reported in Table 1.

We observe that Gen3DEval performs equally well on synthetic surface assessment in all the configurations with lowest accuracy score of 0.89 in the case of Gen3DEval w/ Fit3D whereas on user-annotated out-of-domain (OOD) benchmark, CLIP clearly outperforms the rest. On the synthetic appearance benchmark, described in Section 5.2, Gen3DEval w/ Fit3D reports the lowest accuracy of 0.8, with the rest between .85-.89. On the other hand, on the human evaluation appearance dataset (in-domain methods) described in Section 5.1, Gen3DEval w/ CLIP as well as a combination of Fit3D and CLIP report the best accuracy score of 0.9, followed by CLIP and DinoV2 (0.86), then Fit3D (0.81) and finally DinoV2 (0.77). In terms of generalization with OOD benchmarks, Gen3DEval with CLIP outperforms the rest by a large margin.

On text fidelity benchmarks, in keeping with our earlier observation, the performance on OOD text fidelity benchmark is much better for Gen3DEval with CLIP (0.86) alone compared with the rest, with the nearest neighbour in CLIP and DinoV2 (0.74). Finally, on the synthetic text fidelity benchmark, Gen3DEval w/ DinoV2 alone underwhelms reporting 0.75. While standalone numbers for Fit3D is better, CLIP reports high scores by itself as well as in conjunction with Fit3D and DinoV2.

Overall, we noticed that Gen3DEval with CLIP embedding consistently performs well across all evaluation dimensions. As a result of this, Gen3DEval uses CLIP encoder to extract image embeddings.

4.3. Stage 1: Pre-training

The objective of this training stage is to train the projection matrix to learn correlations between the image encoding space and the language description space. To train the projection matrix, we use the renderings of the 141,000 3D artist-created meshes and their accompanying text prompts. We sample 40K single view image, 40K two-view images, 40K four-view images, and 10K four-view rendered surface normal images and their corresponding captions. We also combine 11.4K of the four-view images mentioned above and combine them to form an image grid. In the case of

surface normals, we process the captions to remove any aspect of appearance mentioned in them which is irrelevant to rendered normals. All multi-view images are sampled uniformly from a 360° azimuth with a fixed elevation angle.

The training process involved a batch size of 16, learning rate of $1e^{-3}$, a cosine learning rate scheduler with a warm-up ratio of 0.03, using the ADAM optimizer. We optimized the model using maximum likelihood for the next token prediction. Pre-training was conducted on 8 A100 GPUs over a period of 1 day, encompassing 8K iterations.

4.4. Stage 2: Supervised Fine-tuning

The objective of the supervised fine-tuning (SFT) stage is to jointly train the instruction-following large language model (LLM) and the pre-trained projection matrix for the task of selecting the best 3D object out of two based on text fidelity, 3D appearance and surface quality. For fine-tuning, we utilize the human-annotated data and the synthetically generated comparison data. The SFT dataset distribution is displayed in the appendix.

The fine-tuning process involved a batch size of 4, a learning rate of $2e^{-6}$ for the projector, and a learning rate of $1e^{-5}$ for the vLLM, with a cosine learning rate scheduler and a warm-up ratio of 0.03, using the ADAM optimizer. We optimized the model by using maximum likelihood of next token prediction. This stage was trained on 16 A100 GPUs for 18 hours, for 4K iterations.

5. Experiments

We evaluated the performance of Gen3DEval using three distinct datasets. Their details are explained in the following subsections. We also performed ablation studies to explore the impact of different image encoders on the performance of Gen3DEval. This involved varying the types of image encoders and delineating their respective contributions to the overall performance of the model in Section 4.2.

5.1. Human Evaluation Dataset

The purpose of the human evaluation dataset is to assess the alignment of Gen3DEval with human preference on text fidelity, appearance and surface quality. We curated this dataset in the same way as we curated the human preference data for the supervised fine-tuning stage of training in Section 4.1 under Human Annotations. Post curation and processing, we split the total data by holding out 10% of the total prompts (404) for the creation of this dataset and using the rest (90%) for the supervised fine-tuning stage of training. This dataset has 506 VQA comparison data samples for a total of 40 prompts, annotated on the basis of appearance. We also added 3 evaluation datasets, one each for appearance, surface and text fidelity generated from annotating pairwise evaluation and removing any ambiguities from methods as well as prompts that were not used as part of the training data. We use these to calculate out-of-domain (OOD) generalization performance of our Gen3DEval.

5.2. Synthetic Dataset

The second dataset is a synthetic evaluation dataset that includes objects generated by 3D artists along with their synthetically perturbed counterparts, enabling controlled experimentation in a low-noise environment. We created and processed this dataset in the same way as we processed the synthetic dataset for appearance, surface and textual perturbations detailed in Section 4.1 under Synthetic Data. Given that the VQA training is so diverse, we further ensured no overlap with training captions and objects by filtering the evaluation dataset using a sentence similarity threshold using sentence transformer embeddings [50]. We also curated a portion of the synthetic evaluation dataset by applying surface perturbations to artist-drawn meshes and rendering the surface normals of these pairs of objects. All the filtering mechanisms were similarly applied to this dataset.

5.3. Benchmark Details

We create a diverse set of 80 prompts, Gen3DEval-Bench, which considers the diversity of objects, textures, and levels of composition. We determine the size of this benchmark with the consideration that text-to-3D generation is a time- and computation-intensive process, aiming to make the benchmark easily accessible. It is split between 40 animate (humanoids, animals) objects and 40 inanimate objects, as well as into 43 single object and 37 composite object prompts, i.e., combining multiple objects. The average number of words per prompt is 12.863. Refer to the supplementary for comparison with other prompt benchmarks.

5.4. Metric Computation

To compute the metric, Gen3DEval compares two methods at a time using the following procedure. First, it samples four images from a 360° RGB or surface normal video circling the object, at equal intervals covering a 360° view of the 3D asset per prompt per method for a pair of methods. For each prompt in Gen3DEval-Bench and each pair of methods, Gen3DEval is applied to a pair of assets at a time. It takes 8 input images (4 per object) alongwith the relevant QA prompt and parses the natural language output using [2] to determine which 3D asset is better. Subsequently, it applies the ELO rating system to the parsed outputs and extracts an overall ranking metric. The generation prompt is only provided in the case of evaluation on text fidelity and not for appearance and surface. We treat them as separate tasks which enables us to compare any two generated assets, irrespective of the generated prompts. It also allows us to evaluate image-to-3D methods more effectively.

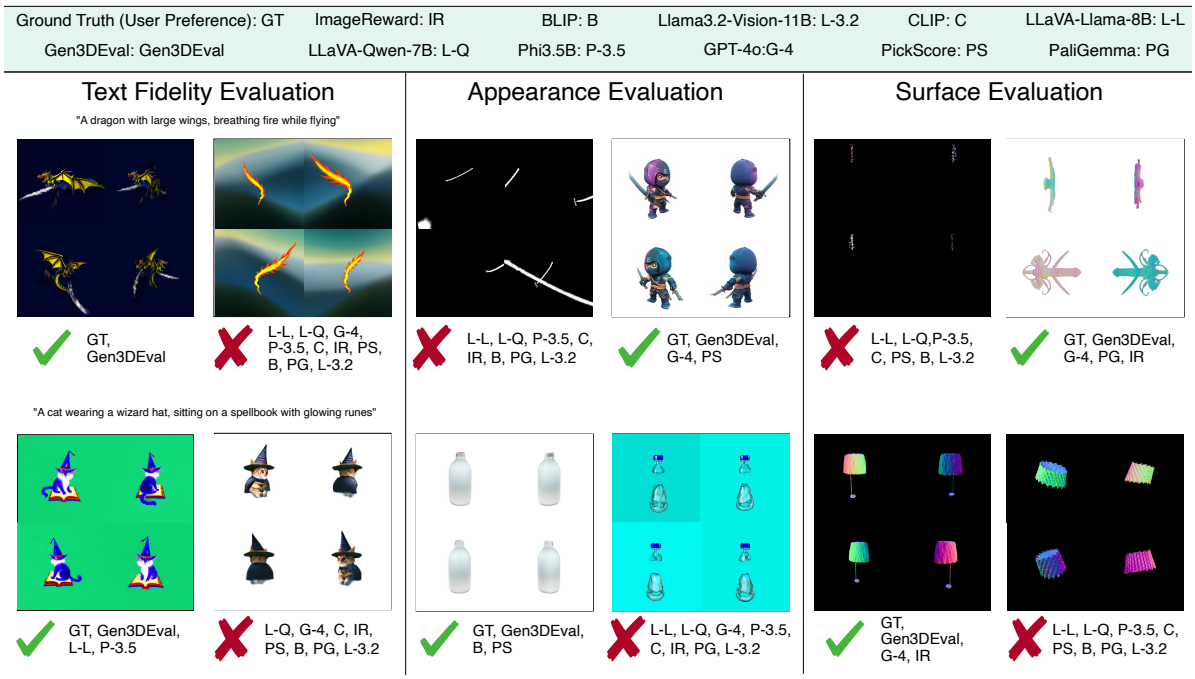

Figure 4. **Qualitative Comparison** of methods on samples of the evaluation dataset across text fidelity, appearance and surface evaluation.

6. Results

6.1. Gen3DEval and other evaluator methods

We compare Gen3DEval against classical baseline metrics such as CLIP as well state-of-the-art vLLMs on evaluation datasets described in Section 5.1 and Section 5.2. We show that the model outperforms all the current methods on assessing appearance preference by a large margin on synthetic, user preference and out-of-domain evaluation data, demonstrating a strong correlation with human preference in the context of text-to-3D asset generation. In terms of text fidelity, Gen3DEval outperforms CLIP [42], which is the most popular metric for text fidelity evaluation. Gen3DEval also narrowly outperforms ImageReward [59] and PickScore [26] on the out-of-domain benchmark which has been curated to remove any ambiguous samples. We also outperform our baselines of surface comparisons data using only synthetically perturbed surface normals.

Moreover, Gen3DEval is the first method that unifies text-to-3D generation metrics by incorporating appearance as well as text fidelity metrics in a holistic manner as evidenced in Table 1. We also provide a qualitative comparison of Gen3DEval with other methods on a few samples from the evaluation dataset in Figure 4.

We also report results for ablation studies for choice of image encoders used for the instruction tuning stage in Table 1. We note from the ablations that CLIP embeddings have a more consistent performance across all dimensions for the purpose of our metric.

Finally, we note that while user studies or user preference data is the current gold standard, it can be noisy and uncorrelated. For instance, when it comes to text fidelity comparison, sometimes, the preference is influenced by appearance. Moreover, with short or simple generation prompts, it is difficult to pick one over the other. In case of appearance comparison, sometimes, the background or scale can influence our choice.

6.2. Generative 3D Methods on Gen3DEval-Bench

Table 2 notes the results of Gen3DEval on Gen3DEval-Bench for a collection of 10 generative 3D methods, namely, Trellis [58], AssetGen [47], Fantasia3D [9], TripoSR [51], Magic123 [41], Magic3d [28], Vfusion-3d [15], Dreamfusion [40], LatentNerf [35] and Flex3D [16]. We rank them in the order of their performance for appearance, surface quality and text fidelity. The overall ranking is an average of the scores of appearance and text fidelity.

6.3. Limitations

Gen3DEval's assessment of objects with janus can be slightly erratic. There is room for improvement for out-of-domain performance for surface evaluation, because of limited availability of diverse, annotated surface comparison data. We also note that for image-to-3D generation methods, the performance of methods are inherently influenced by a text-to-image generation pipeline. Therefore, composing a strong and consistent image benchmark would be a logical next step. While Gen3DEval can be used for the pur-

	Appearance			Surface		T-Fidelity	
	Human	Synthetic	OOD	Synthetic	OOD	Synthetic	OOD
Classical							
Avg. CLIP Score [42]	0.3	0.4	0.17	0.3	0.45	0.78	0.8
Avg. Image Reward Score [59]	0.73	0.6	0.66	0.7	0.54	0.65	0.85
Avg. PickScore [26]	0.37	0.25	0.34	0.26	0.21	0.81	0.85
Vision Large Language models							
Phi-3.5-Vision [1]	0.53	0.47	0.54	0.49	0.5	0.64	0.65
LLaVA-Qwen-7B [29]	0.54	0.46	0.54	0.51	0.46	0.68	0.58
LLaVA-Llama3-8b [29]	0.5	0.5	0.47	0.47	0.48	0.49	0.5
Llama3.2-Vision-11B* [2]	0.06	0.04	0.05	0.1	0.07	0.04	0.5
BLIP* [27]	0.05	0.28	0.2	0.07	0.09	0.37	0.13
GPT-4o* [38]	0.59	0.48	0.69	0.54	0.54	0.61	0.55
PaliGemma* [7]	0.02	0.02	0.21	0.25	0.25	0.17	0.1
Gen3DEval (CLIP)	**0.9**	0.85	**0.89**	**0.99**	**0.67**	0.98	**0.86**
Gen3DEval (CLIP + Fit3D)	**0.9**	0.88	0.78	0.97	0.57	**1**	0.53
Gen3DEval (CLIP + DinoV2)	0.86	**0.89**	0.78	**0.99**	0.51	0.98	0.74
w/ Fit3D	0.81	0.8	0.55	0.89	0.44	0.93	0.44
w/ DinoV2	0.77	0.87	0.54	0.97	0.61	0.75	0.58

Table 1. We report accuracy for curated synthetic and out-of-domain human preference evaluation datasets for appearance, surface and fidelity to text. Additionally, for appearance, we compare methods on in-domain (unseen prompts from methods used for training data). We compare our method against classical metric methods as well as other vLLMs. For text fidelity, we do not provide prompts or pass empty strings for the classical methods. Methods with * next to their names do not currently support multi-image input and were passed either 4x2 grids composed of eight images or 8 input images in sequence in case of GPT-4o.

Methods	Appear.	Surf.	T-Fidelity	Overall
Trellis* [58]	1	1	1	1
AssetGen [47]	2	7	2	2
Flex3d* [16]	4	N/A	3	3
Latentnerf [35]	3	N/A	6	4
Magic123 [41]	5	4	4	5
Vfusion-3d* [15]	6	8	5	6
Magic3d [28]	7	5	7	7
Dreamfusion [40]	8	2	8	8
Fantasia3D [9]	N/A	3	N/A	N/A
TripoSR [51]	N/A	6	N/A	N/A

Table 2. Gen3DEval applied to 3D generation methods on Gen3DEval-Bench. Methods are ranked (Best , Worst) on text fidelity, appearance and surface quality score (if available). Only appearance and text fidelity are used for the overall score. Image-to-3D methods are denoted with *.

pose of evaluating any given pair of generated 3D objects, for the purpose of being used as a standard evaluation metric, comprehensive application across numerous methods is relevant to its performance; examples in the appendix.

7. Conclusion

In this paper, we have laid out the lack of an existing metric in the text-to-3D domain that caters to all its necessary parameters and established the relevance for developing such a metric. It is a difficult problem to solve because of the many ways in which 3D generation is supported. Diverging from the trend of using similarity metrics, which is impractical in the case of text-to-3D generation, we propose introducing 3D aesthetic preference into the vision-language space and transforming that into a ready-to-use evaluation ranking metric using vLLMs. We demonstrate that Gen3DEval, a vLLM containing 8.35 billion parameters and trained using a synthetically curated 3D-object dataset coupled with user preference data establishes itself as a comprehensive, accessible and competitive evaluation method displaying strong performance on relevant metric dimensions, i.e., appearance, texture and text fidelity, in alignment with user 3D object preference. We hope that Gen3DEval will provide a standard benchmark and metric for the comparison of existing and future methods.

Acknowledgements

We thank Antoine Toisoul for help in curating the synthetic data for our experiments. Shalini Maiti has been supported by a sponsored research award from Meta.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 8

[2] Meta AI. The llama 3 herd of models, 2024. 2, 3, 4, 5, 6, 8

[3] Anthropic. 3

[4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2

[5] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. 2, 3

[6] Raphael Bensadoun, Yanir Kleiman, Idan Azuri, Omri Harosh, Andrea Vedaldi, Natalia Neverova, and Oran Gafni. Meta 3d texturegen: Fast and consistent texture generation for 3d objects. *arXiv preprint arXiv:2407.02430*, 2024. 4

[7] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 8

[8] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. 4

[9] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22246–22256, 2023. 7, 8

[10] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 2

[11] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words. *arXiv preprint arXiv:2010.11929*, 7, 2020. 5

[13] Arpad E Elo and Sam Sloan. The rating of chessplayers: Past and present. *(No Title)*, 1978. 2

[14] Ruiqi Gao*, Aleksander Holynski*, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole*. Cat3d: Create anything in 3d with multi-view diffusion models. *Advances in Neural Information Processing Systems*, 2024. 2

[15] Junlin Han, Filippos Kokkinos, and Philip Torr. Vfusion3d: Learning scalable 3d generative models from video diffusion models, 2024. 2, 4, 7, 8

[16] Junlin Han, Jianyuan Wang, Andrea Vedaldi, Philip Torr, and Filippos Kokkinos. Flex3d: Feed-forward 3d generation with flexible reconstruction model and input view curation. *arXiv preprint arXiv:2410.00890*, 2024. 4, 7, 8

[17] Yuze He, Yushi Bai, Matthieu Lin, Wang Zhao, Yubin Hu, Jenny Sheng, Ran Yi, Juanzi Li, and Yong-Jin Liu. T^3bench: Benchmarking current progress in text-to-3d generation, 2023. 3

[18] Zexin He and Tengfei Wang. Openlrm: Open-source large reconstruction models. https://github.com/3DTopia/OpenLRM, 2023. 4

[19] Leonard R Herrmann. Laplacian-isoparametric grid generation scheme. *Journal of the Engineering Mechanics Division*, 102(5):749–756, 1976. 4

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1

[21] Lukas Höllein, Aljaž Božič, Norman Müller, David Novotny, Hung-Yu Tseng, Christian Richardt, Michael Zollhöfer, and Matthias Nießner. Viewdiff: 3d-consistent image generation with text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[22] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2

[23] Yukun Huang, Jianan Wang, Yukai Shi, Boshi Tang, Xianbiao Qi, and Lei Zhang. Dreamtime: An improved optimization strategy for diffusion-guided 3d generation. *arXiv preprint arXiv:2306.12422*, 2023. 2

[24] Wonbong Jang and Lourdes Agapito. Nvist: In the wild new view synthesis from a single image with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10181–10193, 2024. 2

[25] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1, 2

[26] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: an open dataset

of user preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2023. Curran Associates Inc. 7, 8

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 2, 3, 8

[28] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 300–309, 2023. 2, 7, 8

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2, 3, 4, 8

[30] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. *arXiv preprint arXiv:2311.07885*, 2023. 4

[31] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2

[32] Yuxin Liu, Minshan Xie, Hanyuan Liu, and Tien-Tsin Wong. Text-guided texturing by synchronized multi-view diffusion. *arXiv preprint arXiv:2311.12891*, 2023. 4

[33] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, Natalia Neverova, Andrea Vedaldi, Oran Gafni, and Filippos Kokkinos. Im-3d: Iterative multiview diffusion and reconstruction for high-quality 3d generation. *arXiv preprint arXiv:2402.08682*, 2024. 5

[34] Meshy. Meshy text-to-3D v3.0, 2024. 4

[35] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022. 7, 8

[36] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, page 405–421, Berlin, Heidelberg, 2020. Springer-Verlag. 1, 2

[37] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*, pages 3501 – 3512, Piscataway, NJ, 2020. IEEE. 2

[38] OpenAI. Gpt-4 technical report. 2023. 2, 3, 8

[39] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023. 4, 5

[40] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 2, 3, 4, 7, 8

[41] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. 2, 7, 8

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 2, 4, 5, 7, 8

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1

[44] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Lit, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 4

[45] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model, 2023. 2

[46] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023. 2, 4

[47] Yawar Siddiqui, Tom Monnier, Filippos Kokkinos, Mahendra Kariya, Yanir Kleiman, Emilien Garreau, Oran Gafni, Natalia Neverova, Andrea Vedaldi, Roman Shapovalov, and David Novotny. Meta 3d assetgen: Text-to-mesh generation with high-quality geometry, texture, and pbr materials. *arXiv*, 2024. 2, 4, 7, 8

[48] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 2

[49] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3

[50] Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. Augmented SBERT: Data augmentation

method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online, 2021. Association for Computational Linguistics. 6

[51] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 4, 7, 8

[52] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 3

[53] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. *arXiv preprint arXiv:2212.00774*, 2022. 2

[54] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 2, 4

[55] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 1, 2

[56] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: high-fidelity and diverse text-to-3d generation with variational score distillation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 2

[57] Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22227–22238, 2024. 2, 3

[58] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 5, 7, 8

[59] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2023. Curran Associates Inc. 7, 8

[60] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[61] Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *CVPR*, 2024. 2

[62] Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2D Feature Representations by 3D-Aware Fine-Tuning. In *European Conference on Computer Vision (ECCV)*, 2024. 4, 5

[63] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2

[64] Junzhe Zhu, Peiye Zhuang, and Oluwasanmi Koyejo. Hifa: High-fidelity text-to-3d generation with advanced diffusion guidance. In *International Conference on Learning Representations*, 2023. 2