Towards Preference-Aligned 3D Quality Assessment

JiHyuk Byun
Yonsei University
quswlgur123@yonsei.ac.kr

Seon Joo Kim
Yonsei University
seonjookim@yonsei.ac.kr

Abstract

3D content creation is rapidly expanding across applications such as AR/VR, robotics, and simulation, yet evaluating the perceptual quality of 3D assets remains challenging. Existing 3D quality assessment (3D-QA) benchmarks mainly rely on synthetic distortions, which fail to capture naturally occurring artifacts in real-world and generative 3D data. In this work, we introduce 3D-PAQA, the first large-scale preference-aligned 3D-QA dataset constructed using multimodal large language models (MLLMs) guided by human-labeled exemplar anchors. By leveraging exemplar-anchored prompting and relative ranking, our approach injects human preference signals into the MLLM's context and yields stable, human-aligned quality annotations for over 260K 3D assets across six perceptual criteria. We further train a lightweight 3D-Evaluator based on Point Transformer-v3 that predicts perceptual quality directly from point-cloud features. Remarkably, the evaluator surpasses its teacher MLLM in correlation with human judgments, indicating that MLLM-derived supervision can be distilled into a compact, robust metric. Beyond benchmarking, 3D-PAQA and its evaluator pave the way for scalable, human-aligned 3D assessment applicable to both dataset curation and generative model evaluation.

1 Introduction

3D content creation is rapidly gaining attention across applications such as AR/VR, robotics, and simulation. With the emergence of large-scale repositories like Objaverse and ShapeNet [2, 1], 3D assets have become widely accessible, fueling the recent surge in 3D generative modeling. However, assessing the quality of these assets remains a significant challenge. Ensuring high-quality meshes is essential for the usability of downstream 3D applications, yet existing benchmarks often lack meaningful quality indicators.

Most existing 3D quality assessment (3D-QA) approaches are focusing on synthetic distortions[10, 12, 36], where artificial degradations (*e.g., gaussian noise, downsampling*) are introduced to generate distorted samples from pristine references. While useful for controlled experiments, such distortions poorly reflect the artifacts commonly found in real-world 3D assets (See Figure 1.). Consequently, metrics trained on distortion-based datasets often misalign with human perception, highlighting the need for more realistic and subjective evaluation.

The same challenge arises in 3D generative modeling. As shape generation or text-to-3D models advance, evaluating their outputs remains difficult due to the absence of clear ground truth. Although user studies are commonly employed, they are costly, inconsistent, and unscalable. Moreover, generative model training itself requires careful quality control: low-quality assets can degrade performance, while preference-aligned QA can enable better dataset curation. Thus, scalable human-aligned datasets are essential for both evaluating and curating 3D content.

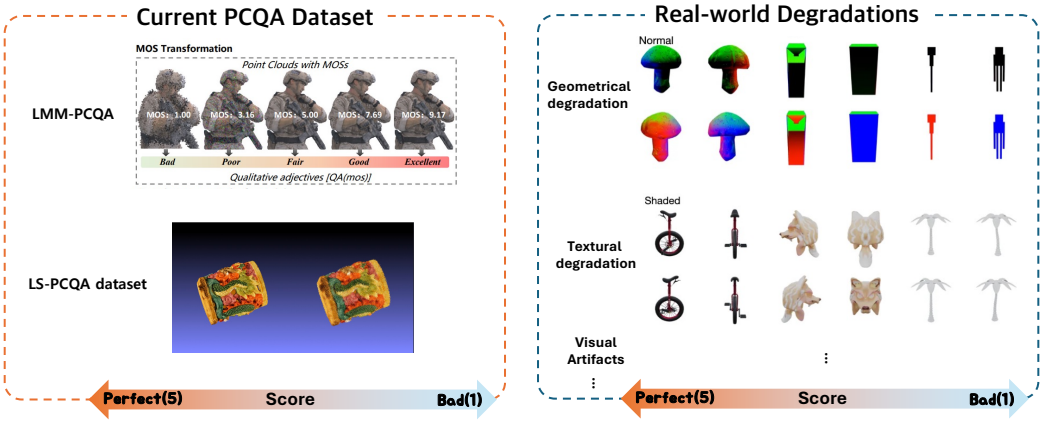


Figure 1: **Dataset degradation comparisons.** Existing 3D-QA datasets rely on synthetic distortions (e.g., geometrical or textural degradation), which fail to represent real-world artifacts (The figure is copied from [15, 10]). In contrast, 3D-PAQA captures naturally occurring degradations and aligns quality assessment with human perception through MLLM-based preference annotations.

To address these challenges, recent studies[5, 8, 22, 32] have explored automated 3D evaluation frameworks leveraging large models to approximate human perception. These can be broadly categorized into MLLM-based and non-MLLM-based approaches. MLLM-based methods exploit large vision-language models such as GPT-4V to assess 3D quality via pairwise A/B testing or instruction-based comparisons, demonstrating strong alignment with human judgment [32, 5, 22]. However, these methods require extensive comparisons and incur high computational costs, while still producing relative rather than absolute quality scores. In contrast, non-MLLM methods such as Objaverse++ [8] train lightweight curators on small human-labeled datasets ($\approx$10k samples), but their limited scale restricts generalizability. Our work combines the strengths of both paradigms—leveraging the human-alignment capability of MLLMs to construct a scalable preference-aligned dataset, which also enables the training of an efficient 3D evaluator.

In this paper, we propose 3D Preference-Aligned Quality Assessment (3D-PAQA), a new framework for scalable and human-aligned 3D quality evaluation. Our goal is to construct a dataset and evaluation pipeline that capture subjective human preferences without relying on costly user studies or distortion-based assumptions. To this end, we leverage multimodal large language models (MLLMs) to annotate large-scale 3D assets through a structured prompting scheme. Specifically, we design a two-stage system consisting of (1) exemplar-anchored visual prompting, where human-labeled reference samples inject preference anchors into the MLLM's context, and (2) a relative ranking mechanism that stabilizes quality judgments across six perceptual dimensions (See Figure 2). This process yields a large-scale, preference-consistent 3D-QA dataset covering over 260K assets and enables the training of a lightweight 3D-Evaluator that distills MLLM annotations into an efficient point-based quality predictor. Compared to prior frameworks that depend on pairwise A/B testing or limited user labels, 3D-PAQA achieves improved scalability, stability, and human alignment, establishing a practical foundation for both dataset curation and generative model evaluation.

Overall, our work demonstrates that it is possible to construct a scalable, human-aligned 3D-QA dataset by leveraging MLLMs. At the same time, we formally introduce the 3D-PAQA task, establishing a benchmark and methodology for preference-aligned 3D quality assessment. Specifically, we make the following contributions:

- We present **the first large-scale human-aligned 3D quality dataset** with annotations for **260k** Objaverse assets across six criteria. This explicitly bridges the gap between subjective human aesthetic preferences and objective automatic QA in 3D.
- To generate large-scale annotations, we design a **simple in-context prompting scheme** that embeds exemplar-anchor sets into MLLM prompts, enhancing label stability and consistency without costly A/B testing.
- We show that a **light-weight 3D evaluator** trained on our dataset can efficiently predict preference-aligned quality scores, demonstrating superior performance compared to a 72B-parameter MLLM (Qwen2-VL 72B) on a human-labeled benchmark.

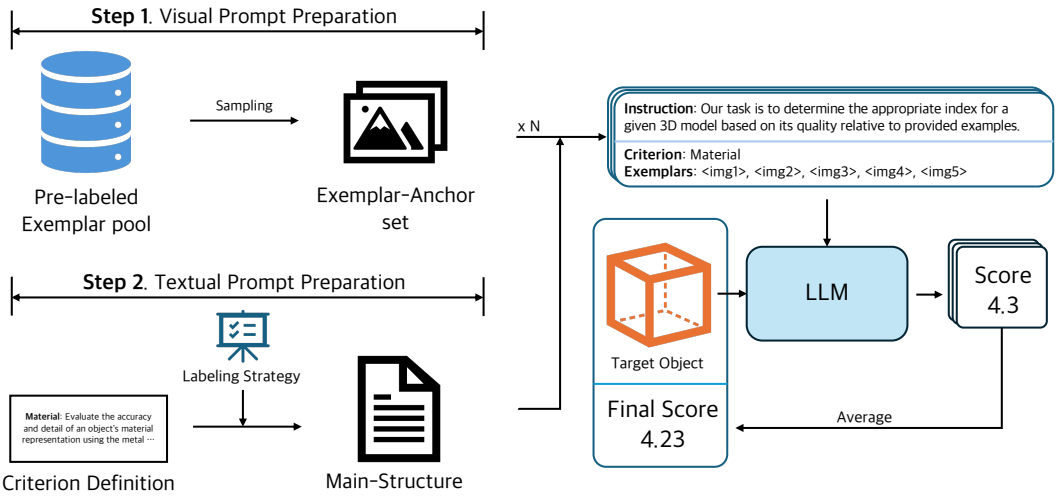


Figure 2: **Overview of the 3D-PAQA dataset construction pipeline.** Multi-view renderings of each object are evaluated through exemplar-anchored visual prompts and Labeling Strategy by an MLLM, yielding stable, preference-aligned annotations across six perceptual criteria. The exemplar anchor set is sampled N times, resulting in N quality scores predicted by the LLM for each criterion. These scores are then averaged to obtain the final annotation for the target object.

2 Related work

3D Quality Assessment and datasets. 3D quality assessment (3D-QA) aims to evaluate the perceptual quality of distorted 3D content. Most existing methods focus on synthetic distortions introduced for controlled degradations such as compression, transmission, or rendering. Benchmarks like WPC, LS-PCQA, and SJTU-PCQA [10, 12, 36] adopt this design, where pristine references are augmented with distortions (e.g., downsampling, Gaussian noise, compression) and scored by subjective ratings. To mitigate the scarcity of dataset scale, various methods have been proposed [13, 18, 19, 37, 39]: DiSPA disentangles perceptual factors via mutual information minimization, CoPA leverages contrastive pre-training on unlabeled point clouds, and IT-PCQA transfers perceptual knowledge from 2D IQA through domain adaptation. However, synthetic distortions poorly reflect real-world artifacts(See Figure 1.)—such as non-manifold geometry or unrealistic proportions—and the limited diversity of reference objects restricts generalization. This motivates moving beyond synthetic benchmarks. In this work, we release the first large-scale 3D-QA dataset capturing naturally occurring artifacts in generated 3D assets. Comprising over 260k unique samples, our dataset enables realistic and scalable quality assessment, bridging subjective human preferences with automatic evaluation across diverse and unseen 3D content.

3D Generative models. 3D generative models have advanced rapidly, fueled by large-scale datasets such as Objaverse [2]. Existing approaches fall into three categories: optimization-based methods [7, 11, 16, 25], multi-view diffusion-based methods [4, 9, 20, 21, 31], and large reconstruction models [23, 26, 34, 35, 38]. Recent works such as TRELLIS [34] and CLAY [38] indicate that 3D models are beginning to approach human-level quality in generating realistic content. Despite this progress, evaluation and curation remain bottlenecks. Current assessments of generated assets often rely on statistical metrics such as 3D-IoU, FID or F-score, which fail to reflect human preferences and thus require costly user studies[31, 38]. At the same time, the quality of training data strongly influences generative performance. Since Objaverse is web-crawled, its assets vary widely in quality, leading researchers to adopt ad-hoc curation strategies ranging from manual filtering to CLIP-based heuristics. These approaches are either labor-intensive, computationally costly, or indirect, underscoring the need for scalable, preference-aligned 3D quality assessment.

MLLMs as human-aligned evaluators. The emergence of CLIP marked a milestone in multimodal learning by aligning vision and language within a shared space, and GPT-4V further demonstrated that MLLMs can interpret visual inputs in ways closely aligned with human perception. Building on this capability, recent works in IQA [27–30] showed that MLLMs can function as human-aligned

quality evaluators, bridging the gap between traditional distortion-based metrics and subjective human judgment. This trend has recently extended to 3D [5, 22, 14, 32], with approaches applying MLLMs to assess generative models. While these methods highlight the potential of preference-aligned 3D evaluation, they remain limited to comparisons between models rather than per-content scoring. Moreover, most rely on A/B testing aggregated into Elo ratings [3], which require a large number of pairwise comparisons, making them computationally costly and inherently heuristic. This reliance on relative scoring underscores a fundamental limitation of current MLLM-based evaluators in 3D QA.

3 3D-PAQA

The goal of this work is to achieve scalable and human-aligned 3D quality assessment by constructing a large-scale preference-aligned dataset and training an efficient 3D evaluator. To build 3D-PAQA, we leverage MLLMs guided by a stratified pool of human-labeled exemplar anchors that inject human preference into the prompting process. Conditioned on these anchors, the MLLM predicts perceptual scores for target samples, producing stable annotations used for evaluator training.

A key challenge in preference-based quality assessment lies in its inherent *subjectivity*, often making MLLM outputs unstable and inaccurate. To address this, we introduce two core components: **Exemplar-Anchors** and a **Relative Ranking strategy**, which jointly enforce consistency in MLLM judgments and yield more reliable preference-aligned annotations.

3.1 Task Definition: Preference-aligned 3DQA

3D quality assessment (3D-QA) traditionally aims to measure the perceptual fidelity of 3D assets, often with respect to synthetic distortions such as *downsampling, Gaussian noise, or compression artifacts*[10, 12, 13]. While such distortion-based benchmarks provide controlled testbeds, they fail to capture the naturally occurring artifacts—including non-manifold geometry, unrealistic proportions, or missing surfaces—that are prevalent in real-world 3D content and especially in assets generated by modern 3D generative models.

To bridge this gap, we define the task of **3D P**reference-**A**ligned **Q**uality **A**ssessment (3D-PAQA). Unlike *distortion-centric* QA, 3D-PAQA explicitly aligns quality evaluation with human perceptual preference. Formally, given a 3D asset X, the goal is to estimate a quality score vector

$$Q(X) = [q_1, q_2, \ldots, q_k] \tag{1}$$

across k perceptual criteria (e.g., geometry plausibility, texture fidelity, material consistency, artifact absence), where each q_i reflects human-aligned judgment.

The core distinction is that 3D-PAQA does not assume the availability of pristine references or artificial distortions. Instead, it targets absolute quality estimation of *arbitrary* 3D content—both curated datasets and generative outputs—anchored to subjective human preferences. This makes 3D-PAQA directly applicable to benchmarking generative models, curating large-scale repositories (e.g., Objaverse[2]), and improving training pipelines where data quality critically impacts downstream performance.

3.2 Dataset Construction with MLLM

To build a large-scale preference-aligned dataset, we leverage multimodal large language models (MLLMs) as scalable annotators. However, directly prompting MLLMs to score 3D assets often produces unstable and inconsistent results, due to the inherently subjective nature of preference-based assessment.

To address this challenge, we design a two-stage construction pipeline consisting of visual prompting with exemplar-anchors and a preference-assessing strategy (See Figure 3.).

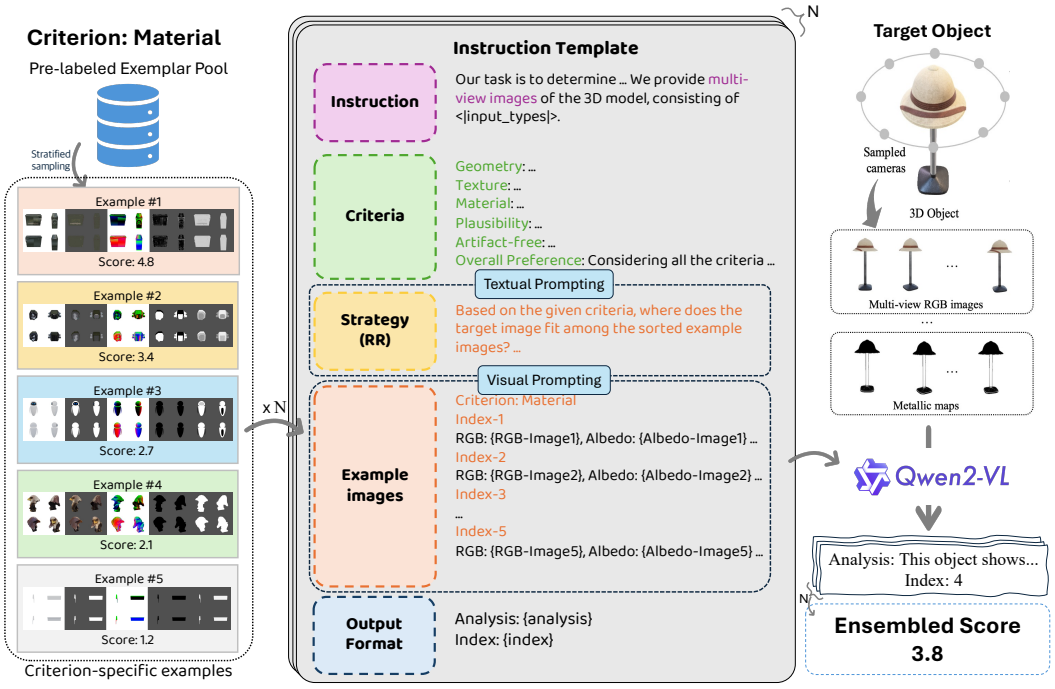


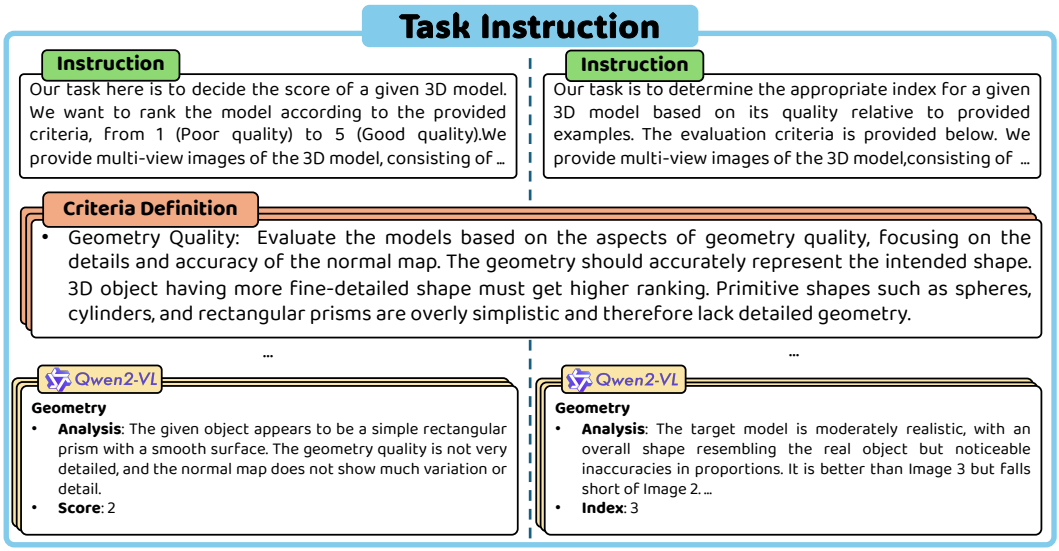
Figure 3: **Overview of the 3D-PAQA annotation process.** Our pipeline leverages multimodal large language models (MLLMs) to generate stable preference-aligned annotations. In the visual prompting stage, a set of stratified exemplar anchors—each representing a distinct quality level—are sampled from a human-labeled pool and presented alongside the target 3D asset. In the textual prompting stage, the MLLM receives criterion-specific instructions (e.g., geometry, texture, material) and predicts the relative rank of the target asset among exemplars. Repeating this process with multiple anchor sets and averaging the results yields consistent, human-aligned quality scores across six perceptual criteria.

Visual Prompting: Exemplar-Anchoring injection. Each target 3D asset is rendered into a set of multi-view images, which are fed to the MLLM along with carefully designed visual and textual instructions (See Figure 3-a). To stabilize outputs, we inject a set of *human-labeled exemplar anchors* into the prompt.

Concretely, we first prepare 100 pre-labeled reference objects spanning 10 categories (10 objects per category), and group them into five quality levels (high to low, 20 objects each). For each evaluation, we sample one object from each group, yielding a set of five anchors that cover the entire quality spectrum. These anchors are inserted alongside the query asset, providing explicit *reference points* for the MLLM to ground its prediction. To reduce anchor-specific bias, we repeat this evaluation five times with independently sampled anchors and average the resulting scores. This exemplar-anchoring scheme ensures that preference signals are explicitly embedded in the prompt and that final annotations remain consistent and robust across variations in anchor choice.

Textual Prompting: Absolute vs. Relative Labeling. To generate preference-aligned annotations at scale, we explored two strategies for scoring 3D assets (See Figure 4). The first is **Absolute Scoring (AS)**, where the MLLM directly predicts an absolute score (e.g., 1–5) for each criterion. While it is straightforward, AS often suffers from instability due to the subjective nature of quality assessment.

To obtain more reliable labels, our main dataset construction relies on a **Relative Ranking (RR)** strategy. Here, the MLLM compares the target asset against exemplar anchors and predicts its relative position within the set, from which the prediction score is derived by averaging the scores of the nearest anchors. This relative grounding induces MLLM to be more grounded on the given exemplar anchors, making the outputs more stable and consistent, which is crucial for constructing reliable annotations at scale.



(a) Absolute Scoring (b) Relative Rating

Figure 4: **Comparison of annotation strategies in 3D-PAQA.** We compare two methods for generating preference-aligned annotations using MLLMs. (a) **Absolute Scoring (AS)**: the MLLM directly assigns an absolute quality score (1–5) to the target asset for each criterion based on its overall analysis. (b) **Relative Ranking (RR)**: the MLLM ranks the target asset among exemplar anchors of varying quality, then derives the final annotation by averaging the scores of the nearest anchors. This relative scheme reduces subjectivity and produces more stable, consistent annotations across perceptual dimensions such as geometry, texture, and material.

3.3 3D Evaluator Training

Having established 3D-PAQA as a scalable and human-aligned dataset, we further validate its effectiveness by training a lightweight 3D evaluator on our annotations. Each 3D asset is sampled into a normalized point set and encoded using a Point Transformer-v3 backbone [33], followed by a small regression head that jointly predicts k preference-aligned quality scores across all perceptual criteria. The model is optimized using a Huber regression loss [6] against our 3D-PAQA annotations, without introducing any additional architectural complexity.

This experiment demonstrates that 3D-PAQA not only provides large-scale, human-aligned supervision but also enables the training of practical and efficient downstream evaluators. The resulting model achieves strong correlation with human-labeled benchmarks while being orders of magnitude smaller and faster than MLLM-based evaluators, confirming that our dataset serves as a reliable foundation for scalable 3D quality assessment (Section 4.3).

4 Experiments

4.1 Evaluation protocol

We report Spearman (SROCC), Pearson (PLCC), and Kendall's τ correlations (KROCC) with human-labeled subsets that were not included in evaluator training. To establish a benchmark, we conducted a user study on an Objaverse subset covering all six criteria, collecting 12k human responses. Baselines include zero-shot MLLMs (Qwen2-VL-72B, Qwen2-VL-7B) and curator-based methods (Objaverse++) for dataset analysis.

4.2 Dataset Analysis

Experimental Details. We implement an annotation pipeline that queries MLLM (Qwen2-VL-72B-int) with exemplar-anchored prompts [24]. For each object, we load its multi-view renderings and PBR maps (RGB, normal, roughness, metallic, albedo), and augment the prompt with five stratified exemplar anchors spanning the quality spectrum. To mitigate anchor-specific bias, we

Method		Preference		Plausibility		Artifacts		Geometry		Texture		Material		Average	
# of param	Strategy	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
7B	AS	0.416	0.488	0.608	0.566	0.366	0.249	0.583	0.552	0.797	0.711	0.540	0.568	0.552	0.522
	RR	0.372	0.245	0.401	0.275	0.276	0.045	0.528	0.500	0.591	0.595	0.183	0.163	0.392	0.354
72B	no-exp	0.582	0.605	0.586	0.578	0.413	<u>0.463</u>	<u>0.620</u>	<u>0.658</u>	0.794	0.772	<u>0.629</u>	<u>0.566</u>	0.604	0.607
	AS	<u>0.648</u>	**0.690**	<u>0.638</u>	<u>0.614</u>	<u>0.431</u>	0.431	0.599	0.611	**0.871**	**0.839**	**0.639**	**0.629**	**0.638**	**0.636**
	RR	**0.685**	<u>0.671</u>	**0.661**	**0.663**	**0.495**	**0.544**	0.567	0.534	<u>0.851</u>	<u>0.724</u>	0.530	0.514	<u>0.631</u>	<u>0.608</u>

Table 1: **Comparison of MLLM-annotated datasets.** We report PLCC and SROCC correlations across six perceptual criteria with human-labeled subsets, along with their averages.(no-exp: no exemplar-anchor, AS: Absolute Scoring strategy, RR: Relative Ranking strategy) Bold indicates best results, underline indicates second-best.

repeat the evaluation with five independent anchor sets and average the outputs to obtain robust annotations. This procedure yields preference-aligned quality labels for 260k Objaverse assets across six perceptual criteria (geometry, texture, material, 3D plausibility, artifacts, and overall preference). Additional implementation details are provided in the Appendix.

Scale and Coverage. We build 3D-PAQA upon the gobjaverse dataset, which provides 280K categorized and publicly available multi-view renderings and PBR assets[17]. From the total 265K usable samples, we leverage the provided category taxonomy spanning 10 semantic domains: *buildings and outdoor scenes, daily-used items, transportation, poor-quality artifacts, human-shaped assets, electronics, animals, furniture, plants, and food*. For each category, 10 representative samples are selected to form a 100-object exemplar-anchor pool, which serves as the human reference set during annotation. Another 100 samples (10 per category) are separately drawn to construct an evaluation set for validating annotation consistency. The remaining data are annotated through our MLLM pipeline under five configurations — 7B-AS, 7B-RR, 72B-no-exp, 72B-AS, and 72B-RR — corresponding to different model capacities and prompting strategies. This design provides semantically diverse and category-balanced annotations, supporting comprehensive and domain-agnostic evaluation of 3D quality.

Human-Alignment ability. We evaluate five annotation configurations (7B-AS, 7B-RR, 72B-no-exp, 72B-AS, 72B-RR) using the Qwen2-VL family. The results are summarized in Table 1. We observe that the 7B model exhibits a significant performance drop when exemplar-anchoring prompts are introduced, particularly under the relative-ranking (RR) strategy. This indicates that smaller models struggle to process multi-reference reasoning within complex prompts, which require higher model capacity. In contrast, exemplar anchoring substantially improves correlation scores for the 72B model, demonstrating its ability to interpret visual anchors effectively.

Moreover, the RR strategy consistently outperforms AS for *holistic* criteria such as preference, plausibility, and artifacts, while AS excels in *component-level* attributes like geometry, texture, and material. Interestingly, the no-exp configuration yields competitive scores in geometry and material, suggesting that these most structural attributes can be objectively assessed without exemplar guidance. For all subsequent experiments, we adopt the 72B-RR annotations as the primary training supervision, given their strongest human-aligned preference correlation.

Our 3D-PAQA has a correlation with object's complexity. We further analyze whether the annotated preference scores correlate with the intrinsic complexity of the 3D assets. Specifically, we group the 72B-RR annotated samples into five score intervals ($1 \leq \text{Score} < 2, \ldots, Score = 5$) and compute the average number of vertices and faces within each group, as shown in Table 2. The results exhibit a clear positive trend: higher-quality objects tend to have larger vertex and face counts, indicating that our 3D-PAQA annotations implicitly capture geometric richness and modeling fidelity. Interestingly, the lowest-quality group ($1 \leq \text{Score} < 2$) shows an atypical increase in vertex and face counts. As illustrated in Figure 5, this anomaly arises because low-preference objects often include geometrically dense yet perceptually degraded assets—such as monochromatic or structurally unnatural meshes. This observation suggests that the MLLM-inferred preference reflects not only geometric complexity but also the balance between geometry and texture realism, indicating that 3D-PAQA captures a holistic perception of 3D quality.

72B-RR	# of assets	Avg. of Vertex #	Avg. of Face #
Score = 5	2762	198737	207918
$4 \leq$ *Score* < 5	35227	166821	188674
$3 \leq$ *Score* < 4	88574	131384	152750
$2 \leq$ *Score* < 3	90028	113433	134664
$1 \leq$ *Score* < 2	28316	159991	196379

Table 2: **Distribution of assets and average vertex/face counts for 72B-RR.**

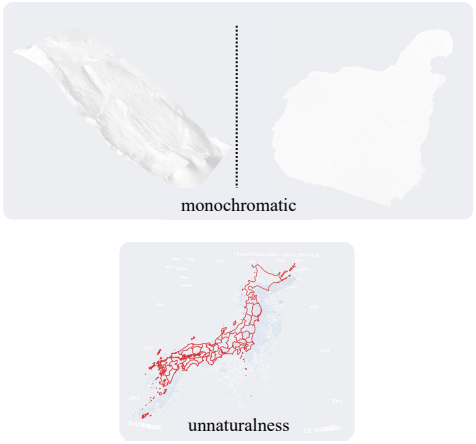


Figure 5: **Examples of textural degradations in Preference-1 examples.** Note that Preference-1 examples include geometrically complex but texturally simple or semantically unrealistic assets, showing that our 72B-RR dataset captures geometric, textural, and semantic quality cues comprehensively.

Comparison to Prior Datasets. Table 3 compares 3D-PAQA to previous 3D-QA datasets, including WPC [36], LS-PCQA [10], and SJTU-PCQA [12]. Prior datasets are primarily distortion-based and limited to synthetic or hybrid scenes with a few thousand samples, each requiring manual subjective ratings. In contrast, 3D-PAQA scales to over 260K real-world Objaverse assets with automatic MLLM-based annotation, covering six perceptual dimensions. Furthermore, while earlier datasets focus on artificially induced degradation, our dataset captures naturally occurring imperfections such as non-manifold geometry, unrealistic proportions, and material mismatches. This large-scale, human-aligned corpus provides a new foundation for training and evaluating open-domain 3D quality predictors.

4.3 3D-Evaluator

Experimental Details. We train a lightweight point cloud–based evaluator using Point Transformer-v3 PTv3) as the backbone architecture [33]. Each mesh asset is sampled into N points with coordinates and features, including RGB color, surface normals, and PBR attributes (roughness, metallic, albedo). When PBR attributes are unavailable, we assign default constant values to maintain feature consistency. These features are processed by the Point Transformer backbone and passed to a two-layer MLP regression head to predict each of 6 criteria. The evaluator is optimized using a Huber regression loss against 3D-PAQA annotations. Training employs AdamW with cosine learning-rate scheduling and standard point cloud augmentations (random rotation, jitter, scaling).

Results. Table 4 presents the quantitative comparison across 72B-RR, Objaverse++, and our trained 3D-Evaluator. The proposed 3D-Evaluator surpasses both baselines in overall correlation with human judgments, achieving the highest PLCC and SROCC across most perceptual dimensions. Compared to Objaverse++, which relies on limited manual curation, our model demonstrates significantly stronger human alignment with no-manual curation. Notably, the 3D-Evaluator even exceeds its teacher model (72B-RR) despite being trained on its annotations. We attribute this to a *distillation effect*, where the model consolidates noisy MLLM supervision into a more stable and human-consistent representation of perceptual quality. These results highlight the potential of the 3D-Evaluator as a lightweight, scalable curator for large-scale 3D datasets.

Dataset	Source	Domain	Annotation	Distortion	Agent	Scalability	Size
WPC [36]	80	Synthetic CAD	MOS (subjective)	Downsampling / Noise / Compression	Human	Low	∼4K
LS-PCQA [10]	104	Indoor / Outdoor	Subjective (single)	Voxel / Color / Rendering	Human	Low	∼8K
SJTU-PCQA [12]	108	Real + Synthetic	30-user study	Geometry / Texture / Compression	Human	Low–Med	∼10K
3D-PAQA (Ours)	**260K**	**Diverse real 3D (Objaverse)**	**MLLM-aligned (6 criteria)**	**Natural artifacts**	**Hybrid**	**High (auto-scalable)**	**260K**

Table 3: **Comparison of 3D-QA datasets.** Our 3D-PAQA uniquely provides large-scale, MLLM-aligned annotations reflecting natural 3D degradations.

Method	Preference		*Plausibility*		*Artifacts*		*Geometry*		*Texture*		*Material*	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
72B-RR	0.685	0.671	*0.661*	*0.663*	*0.495*	*0.544*	*0.567*	*0.534*	*0.851*	*0.724*	*0.530*	*0.514*
Objaverse^{++}[8]	0.505	0.514	–	–	–	–	–	–	–	–	–	–
3D Evaluator	**0.710**	**0.715**	*0.645*	*0.667*	*0.464*	*0.508*	*0.652*	*0.607*	*0.757*	*0.690*	*0.578*	*0.597*

Table 4: **Comparison between 3D-PAQA dataset, Objaverse++ and our 3D Evaluator.** Bold indicates best result, blue italics denote sub-criteria metrics.

5 Future Applications

Our 3D-PAQA dataset demonstrates strong human alignment, and the trained 3D-Evaluator not only inherits this property but also surpasses its teacher model (72B-RR) in correlation with human judgments. In this section, we discuss two promising directions that leverage the proposed dataset and evaluator for practical and scalable 3D quality assessment.

5.1 Generative model Evaluator

Our 3D-Evaluator can serve as a reliable metric for assessing the perceptual quality of 3D generative models. Prior approaches have explored this concept through MLLM-based pairwise A/B testing or relative comparison schemes, which require multiple competing models and numerous iterations to estimate consistent rankings. However, these methods cannot assign absolute scores to a single generative model and incur substantial computational overhead due to the use of large vision-language models. In contrast, our lightweight 3D-Evaluator enables absolute quality prediction with only 33M parameters, providing an efficient and model-agnostic way to evaluate generative outputs.

Extensions. Nevertheless, to generalize beyond human-created assets, future work should extend the 3D-PAQA dataset to include annotations on AI-generated artifacts. Incorporating such generative content will mitigate domain shift and enable the evaluator to consistently assess both human- and AI-generated 3D data.

5.2 Dataset Curator

Our 3D-PAQA (72B-RR) annotations themselves can serve as a high-quality curated dataset for training or evaluating 3D generative models. Since they provide stable, preference-aligned quality scores across diverse object categories, the 72B-RR subset can be directly used as a reliable benchmark for data filtering or generative training. Beyond this, our 3D-Evaluator offers a lightweight and transferable alternative for scalable assessment. With only 33M parameters, it enables absolute quality prediction without relying on costly MLLM inference, providing an efficient and model-agnostic way to evaluate or curate generative outputs across datasets.

6 Conclusion

We presented 3D-PAQA, a large-scale preference-aligned 3D quality assessment dataset, and a lightweight 3D-Evaluator trained upon it. By leveraging exemplar-anchored prompting and relative ranking with MLLMs, our dataset captures human perceptual preferences at scale, bridging the gap between subjective human evaluation and automatic 3D quality estimation. Comprehensive experiments demonstrate that the resulting 3D-Evaluator exhibits strong correlation with human judgments, even surpassing its teacher MLLM across multiple perceptual dimensions. This confirms

the effectiveness of distilling noisy multimodal annotations into a compact and stable quality predictor. Beyond benchmarking, our framework opens new directions for practical 3D quality control — serving as a generative model evaluator and a lightweight dataset curator. Future extensions will further enhance generalization by incorporating AI-generated 3D assets, enabling unified and scalable evaluation across both human- and machine-generated 3D content.

References

[1] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository, 2015.

[2] 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.

[3] Arpad E Elo. The proposed uscf rating system, its development, theory, and applications. *Chess life*, 22(8):242–247, 1967.

[4] 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.

[5] Yuze He, Yushi Bai, Matthieu Lin, Wang Zhao, Yubin Hu, Jenny Sheng, Ran Yi, Juanzi Li, and Yong-Jin Liu. T^3 bench: Benchmarking current progress in text-to-3d generation. *arXiv preprint arXiv:2310.02977*, 2023.

[6] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.

[7] 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, 2023.

[8] Chendi Lin, Heshan Liu, Qunshu Lin, Zachary Bright, Shitao Tang, Yihui He, Minghao Liu, Ling Zhu, and Cindy Le. Objaverse++: Curated 3d object dataset with quality annotations. *arXiv preprint arXiv:2504.07334*, 2025.

[9] 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.

[10] Qi Liu, Honglei Su, Zhengfang Duanmu, Wentao Liu, and Zhou Wang. Perceptual quality assessment of colored 3d point clouds. *IEEE Transactions on Visualization and Computer Graphics*, 29(8):3642–3655, 2022.

[11] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023.

[12] Yipeng Liu, Qi Yang, Yiling Xu, and Le Yang. Point cloud quality assessment: Dataset construction and learning-based no-reference metric, 2022.

[13] Yipeng Liu, Qi Yang, Yiling Xu, and Le Yang. Point cloud quality assessment: Dataset construction and learning-based no-reference metric. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2s):1–26, 2023.

[14] Shalini Maiti, Lourdes Agapito, and Filippos Kokkinos. Gen3deval: Using vllms for automatic evaluation of generated 3d objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18552–18562, 2025.

[15] Yana Nehmé, Johanna Delanoy, Florent Dupont, Jean-Philippe Farrugia, Patrick Le Callet, and Guillaume Lavoué. Textured mesh quality assessment: Large-scale dataset and deep learning-based quality metric, 2023.

[16] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion, 2022.

[17] Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9914–9925, 2024.

[18] Ziyu Shan, Yujie Zhang, Yipeng Liu, and Yiling Xu. Learning disentangled representations for perceptual point cloud quality assessment via mutual information minimization. *arXiv preprint arXiv:2411.07936*, 2024.

[19] Ziyu Shan, Yujie Zhang, Qi Yang, Haichen Yang, Yiling Xu, Jenq-Neng Hwang, Xiaozhong Xu, and Shan Liu. Contrastive pre-training with multi-view fusion for no-reference point cloud quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25942–25951, 2024.

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

[21] Yawar Siddiqui, Tom Monnier, Filippos Kokkinos, Mahendra Kariya, Yanir Kleiman, Emilien Garreau, Oran Gafni, Natalia Neverova, Andrea Vedaldi, Roman Shapovalov, et al. Meta 3d assetgen: Text-to-mesh generation with high-quality geometry, texture, and pbr materials. *Advances in Neural Information Processing Systems*, 37:9532–9564, 2024.

[22] Sitong Su, Xiao Cai, Lianli Gao, Pengpeng Zeng, Qinhong Du, Mengqi Li, Heng Tao Shen, and Jingkuan Song. Gt23d-bench: A comprehensive general text-to-3d generation benchmark. *arXiv preprint arXiv:2412.09997*, 2024.

[23] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024.

[24] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024.

[25] 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, 2023.

[26] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. In *European conference on computer vision*, pages 57–74. Springer, 2024.

[27] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-bench: A benchmark for general-purpose foundation models on low-level vision, 2023.

[28] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, Geng Xue, Wenxiu Sun, Qiong Yan, and Weisi Lin. Q-instruct: Improving low-level visual abilities for multi-modality foundation models, 2023.

[29] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtai Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. Equal Contribution by Wu, Haoning and Zhang, Zicheng. Project Lead by Wu, Haoning. Corresponding Authors: Zhai, Guangtai and Lin, Weisi.

[30] Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, Xiaohong Liu, Guangtao Zhai, Shiqi Wang, and Weisi Lin. Towards open-ended visual quality comparison, 2024.

[31] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *Advances in Neural Information Processing Systems*, 37:125116–125141, 2024.

[32] 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.

[33] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *CVPR*, 2024.

[34] 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.

[35] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. In *European Conference on Computer Vision*, pages 1–20. Springer, 2024.

[36] Qi Yang, Hao Chen, Zhan Ma, Yiling Xu, Rongjun Tang, and Jun Sun. Predicting the perceptual quality of point cloud: A 3d-to-2d projection-based exploration. *IEEE transactions on multimedia*, 23:3877–3891, 2020.

[37] Qi Yang, Yipeng Liu, Siheng Chen, Yiling Xu, and Jun Sun. No-reference point cloud quality assessment via domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21179–21188, 2022.

[38] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024.

[39] Wei Zhou, Guanghui Yue, Ruizeng Zhang, Yipeng Qin, and Hantao Liu. Reduced-reference quality assessment of point clouds via content-oriented saliency projection. *IEEE Signal Processing Letters*, 30:354–358, 2023.