SRAM: Shape-Realism Alignment Metric for No Reference 3D Shape Evaluation

Sheng Liu, Tianyu Luan*, Phani Nuney, Xuelu Feng, Junsong Yuan

State University of New York at Buffalo
{sliu66, tianyulu, phaneesw, xuelufen, jsyuan} @buffalo.edu

Abstract

3D generation and reconstruction techniques have been widely used in computer games, film, and other content creation areas. As the application grows, there is a growing demand for 3D shapes that look truly realistic. Traditional evaluation methods rely on a ground truth to measure mesh fidelity. However, in many practical cases, a shape's realism does not depend on having a ground truth reference. In this work, we propose a Shape-Realism Alignment Metric that leverages a large language model (LLM) as a bridge between mesh shape information and realism evaluation. To achieve this, we adopt a mesh encoding approach that converts 3D shapes into the language token space. A dedicated realism decoder is designed to align the language model's output with human perception of realism. Additionally, we introduce a new dataset, RealismGrading, which provides human-annotated realism scores without the need for ground truth shapes. Our dataset includes shapes generated by 16 different algorithms on over a dozen objects, making it more representative of practical 3D shape distributions. We validate our metric performance and generalizability through k-fold cross-validation across different objects. Experimental results show that our metric correlates well with human perceptions and outperforms existing methods, and has good generalizability.

Introduction

3D generation and reconstruction technologies are widely used in computer games, film and television production, and many emerging content creation applications. There is an increasing demand for 3D shapes that appear genuinely realistic to the human eye. Therefore, it is crucial to develop an evaluation metric aligned with human perception to assess the realism of 3D shapes, guiding the creation of more lifelike 3D shapes and contents.

Previous works, such as (Luan et al. 2024), have mainly focused on evaluating mesh fidelity. As shown in Fig. 1, these evaluations (shown on the left) require a ground truth as a reference. However, in many practical applications, a shape's realism does not necessarily depend on a ground truth. For instance, people typically consider a reconstructed animal as lacking realism. In many 3D generation applications, assessing the realism of a generated shape without relying on a ground truth reference is a common challenge, yet previous works have not adequately addressed this issue.

To overcome the dependence on ground truth for mesh realism evaluation, we design a metric that takes only the 3D shape as input. As shown in Fig. 1 right half, although realism can exist independently of a ground truth, capturing such information without it is challenging. Full-reference metrics, such as (Luan et al. 2024), measure fidelity by comparing the input shape with a ground truth. In contrast, a no-reference realism metric needs to understand the shape by itself at a detailed and semantic level. In full-reference approaches, the magnitude of the difference between two shapes is inversely related to fidelity, making the mapping more regular and easier to train. However, in no-reference metrics, the scale (or magnitude) of a mesh does not explicitly correlate with its realism, so the metric must instead capture the high-level semantic properties of the shape. Here, we refer "no reference fidelity" as "realism" since fidelity is typically based on comparisons while realism is not.

We propose to use a large language model (LLM) as a bridge to provide alignment between mesh shape and realism. LLMs have rich high-level knowledge priors and reasoning capabilities; a well-trained language model can provide the reasoning and knowledge priors needed for our metric to map from shape to realism. If we can align 3D shape information at the input side to the language model, and map the language model's output to realism at the output side, we can achieve better alignment from shape to realism. Through this alignment, we can evaluate shape realism without needing a ground truth shape reference.

We name our proposed metric Shape-Realism Alignment Metric (SRAM). To align mesh shape information with the language model input, we adopt the mesh encoding approach from (Yu et al. 2022) to encode mesh shapes into the language token space. We also design a realism decoder to align the language model's output with human perception of realism. Additionally, to provide the network with training data that includes human-annotated realism scores, we introduce a human-annotated dataset named RealismGrading. Unlike previous datasets such as (Nehmé et al. 2023), our dataset's scoring is based on the shape itself without requiring ground truth references. More importantly, the meshes in

*Corresponding author

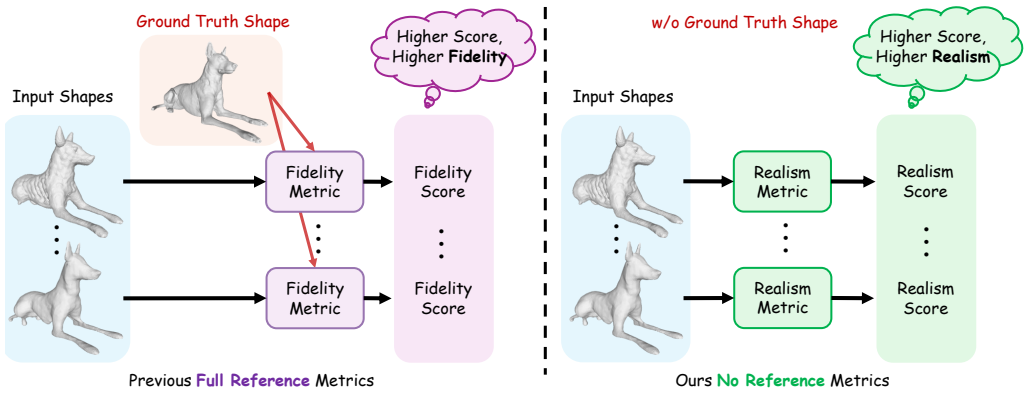

Figure 1: Full reference 3D shape evaluation vs. no reference 3D shape evaluation. Left: Traditional metrics require ground truth references to evaluate the fidelity of 3D shapes. Right: Our metric can evaluate 3D shape realism without a reference. In this paper, we refer "no reference fidelity" as "realism" since fidelity is typically based on comparisons while realism is not.

our dataset are based on real reconstruction and generation algorithms rather than synthetic ones, making this dataset better reflect the distribution of shapes in real-world 3D reconstruction and generation. RealismGrading comprises over a dozen objects, and for each object, we have 9-16 distortion shapes generated by a set of real-world reconstruction and generation algorithms. Furthermore, we obtained realism annotations for these distortion shapes from hundreds of human subjects. In our experiments, we validate our metric's performance and generalizability through k-fold cross-validation across different objects.

In summary, our contributions are as follows:

- We propose a metric for evaluating shape realism that does not require a ground truth mesh as a reference, enabling a no-reference assessment of a mesh's realism.
- We design a language-model-based Shape-Realism alignment pipeline. This pipeline can leverage the language model as a bridge to align the shape feature space with realism feature space, so that the metric can evaluate realism based only on 3D mesh shape without a ground truth mesh shape reference.
- We introduce the RealismGrading dataset, which provides human-annotated training data for our no-reference metric. In our dataset, annotations focus solely on the shape itself, without any comparison to a ground truth. Additionally, the meshes are produced by real reconstruction and generation algorithms rather than being synthetic.

Our experiments demonstrate that our SRAM metric correlates more closely with human perceptions of realism than existing state-of-the-art metrics.

Related Work

3D shape metrics in 3D generation & reconstruction. Evaluating 3D shape quality is challenging due to a mismatch between geometric metrics and human perception. Euclidean distance such as Chamfer Distance (CD) (Borgefors 1984; Luan et al. 2021, 2023; Zhao et al. 2025; Luan et al. 2025; Gong et al. 2023; Wu et al. 2024; Song et al. 2022; Zhang et al. 2021), IoU(Hu et al. 2021; Chen et al. 2021; Nie et al. 2020; Henderson and Ferrari 2018; Tang et al. 2022; Santhanam, Doiphode, and Shi 2023), F-score (Wang et al. 2018; Genova et al. 2020; Bechtold et al. 2021), and UHD (Wu et al. 2020) emphasize geometry but miss semantic and structural cues. For generative models, distribution-level metrics like MMD (Achlioptas et al. 2018), JSD (Kullback and Leibler 1951), TMD (Wu et al. 2020), SCEU (Wang et al. 2022b), and FPD (Shu, Park, and Kwon 2019) assess set-level quality, not individual fidelity. A parallel research has focused on evaluating shape quality in the context of 3D mesh compression and watermarking applications. Works such as (Wang et al. 2010; Bulbul et al. 2011; Lavoué 2009; Corsini et al. 2013). Our work addresses this gap by introducing a perceptually aligned metric for evaluating single 3D shapes.

Perceptual quality assessment for 3D shapes. To overcome the limitations of geometric metrics, recent works (Sarvestani, Tang, and Wang 2025; Yang et al. 2023, 2024; Cui et al. 2024; Nehmé et al. 2020, 2023) propose perceptual quality metrics using deep learning. HybridMQA (Sarvestani, Tang, and Wang 2025) fuses multiple modalities, while TSMD (Yang et al. 2023), TDMD (Yang et al. 2024), and SJTU-PQA (Cui et al. 2024) incorporate texture cues for textured mesh evaluation. However, these full-reference methods rely on ground truth and focus on surface appearance, limiting applicability to textureless or real-world data. Moreover, they are trained on synthetic distortions, unlike real-world artifacts. In contrast, our method requires no reference, targets geometry-only evaluation, and learns from authentic distortions in practical 3D tasks.

Method

Problem Formulation

We formulate the problem of defining a realism metric as follows. Given an input mesh x, we design metric function $F(\cdot)$, whose output is the realism of the mesh, i.e.,

$$s = F(x; \theta), \tag{1}$$

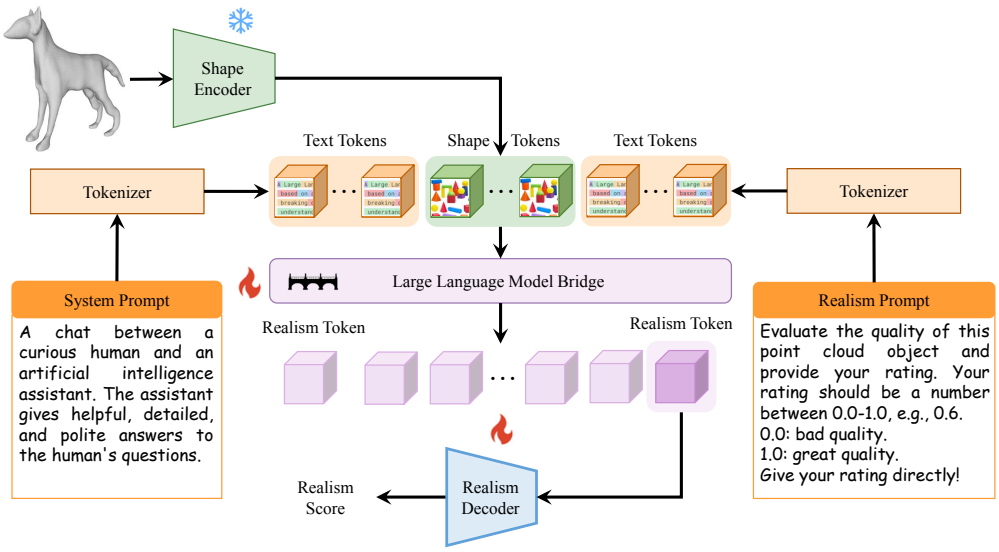

Figure 2: The pipeline of our Shape-Realism Alignment Metric (SRAM). Our metric can take a mesh shape as input and measure its realism without a ground truth mesh shape reference. It uses a language model as a bridge to achieve alignment from 3D shape to realism score. The language model bridge has 3 inputs: text tokens from the system prompt, 3D shape tokens from the 3D shape encoder, and another part of text tokens from the realism prompt. In the output part of our model, we design a token-based realism decoder to align language tokens with realism scores.

where $F(x;\theta)$ is a neural network, and θ is the learnable parameters of the network.

In the following sections, we will introduce how we design the metric $F(x;\theta)$ and how we train it. Overall, our training process can be expressed as:

$$\min_{\theta} \mathcal{L}(F(x;\theta) - y_x), \tag{2}$$

where y is the human annotation of mesh shape input x.

Shape-Realism Alignment Metric

We show our pipeline design in Fig. 2. Our metric can take a mesh shape as input and measure its realism without a ground truth mesh shape reference. It uses an LLM as a bridge to achieve alignment from 3D shape to realism score. Specifically, the input of the language bridge consists of three parts. The first part is the system prompt, which establishes a context for the language model. The second part is the shape tokens, where we use a PointNet-based 3D shape encoder to encode a 3D mesh into tokens. The third part is the realism prompt. This input helps elicit our language model's ability to reason about realism of meshes.

In the output part of our model, we use a token-based realism decoder to regress realism scores. This way, we achieve alignment from mesh shape to realism score, thereby evaluating mesh shape realism without depending on ground truth mesh references. In the remainder of this subsection, we introduce the design of each module of our Shape-Realism Alignment Metric (SRAM).

System prompt. In the context of LLM, the system prompt primarily provides high-level context to the pretrained language model. It may also constrain and guide the language model's behavior. In our metric design, we set the system prompt as a user/assistant Q&A model. We follow (Xu et al. 2024b)'s system prompt to prepare the language model system with overall contextual information.

Shape encoder. The shape encoder extracts shape features. Considering the diversity of our input shapes and the limitations of our human-annotated 3D shape realism data in terms of shape variety, we need a 3D shape encoding model with good context to meet our generalization requirements for shape variety. Therefore, we consider using a large shape dataset pretrained model for this module. We follow the design in Point-BERT (Yu et al. 2022). Here, we treat the vertex of the input mesh shape as a point cloud. Point-BERT's structure can establish good alignment from point clouds to text tokens through the approach of point cloud masking. Point-BERT provides pretrained models on ScanObjectNN (Uy et al. 2019) and ShapeNet (Chang et al. 2015). We can leverage the shape priors inherent in these pretrained models to benefit our shape realism evaluation.

Realism prompt. The realism prompt is the core prompt for using LLM for realism evaluation. Here, we need to design a prompt that can better elicit the LLM's abilities and prior knowledge to benefit our shape evaluation. We experiment with multiple possible prompts to find the most suitable one for our task. Note the realism prompt only provides an initial LLM prior for our metric. Our final realism evaluation does not completely rely on this prompt. During the training phase, we also rely on finetuning to optimize our model, enabling it to better learn the human realism annotations provided in our dataset, thereby better evaluating shape realism.

Large language model bridge. The LLM bridge is a crit-

Object Type	CRM	DreamGaussian	InstantMesh	LGM	One2345	One2345++	PeRFlow	ShapE	ShapE-T	SplatterImage	TripoSR	ECON	ICON	PiFU	PiFUHD	HaMeR
Dog	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓					
Fish	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓					
Female Face	✓	✓	✓	✓	✓		✓		✓	✓	✓					
Male Face	✓	✓	✓	✓	✓		✓		✓	✓	✓					
Female Body A	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Female Body B	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Hand w/ Arm	✓	✓	✓	✓	✓		✓	✓	✓		✓					✓
Hand w/o Arm	✓	✓	✓	✓	✓		✓	✓	✓		✓					✓
Male Body A	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Male Body B	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Building	✓	✓	✓			✓	✓	✓	✓	✓	✓					
Bus	✓	✓	✓	✓		✓	✓	✓	✓		✓					
Keyboard	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓					
Mug	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓					
Plant	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓					
Shoe	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓					

Table 1: A list of the dataset and the 3D reconstruction/generation method we used. ✓ indicates that the corresponding method was employed to generate the mesh for that object category. (Part 1)

ical part in aligning shape information to realism information. Our model's high-level knowledge and reasoning abilities primarily depend on this part. We choose to initialize our LLM bridge with corresponding components from PointLLM (Xu et al. 2024b), a strong 3D LLM pre-trained using hundreds of thousands of 3D objects. The 3D LLM learns to understand semantics and details of 3D shapes during its training process. Such capability is crucial for realism evaluation. Our metric can utilize these capabilities along with appropriate finetuning to achieve better alignment between shape and realism.

Realism decoder. The realism decoder is another key element in our metric design. We need to align text tokens with realism scores using this decoder. Considering the LLM's strong expressive capability and excellent pretraining, and the fact that our output score is only one-dimensional, we can incorporate a major part of the decoder function through fine-tuning the LLM itself. This design allows the LLM's output to already have a good realism representation, rather than placing the main alignment functionality on the decoder. This design can save the number of parameters of the decoder and better utilize the LLM's pretraining. Therefore, we are able to design our decoder in a simple form of an MLP. We aim to only summarize the realism score through the decoder and backpropagate regression gradients through the decoder during training, working with the LLM finetuning to achieve better text-to-realism alignment. Our experiments show that our simplified design can achieve good alignment with human perception.

Training Process

Based on the pre-trained shape encoder and LLM bridge, we design an end-to-end training procedure. First, to leverage the prior knowledge on shapes and language in the shape encoder and LLM bridge, we initially adopt the pre-training model from (Yu et al. 2022) and (Touvron et al. 2023) for shape and language, respectively. Then, we randomly initialize the realism decoder. During training, the shape encoder is frozen while we simultaneously train the realism encoder and finetune the LLM bridge. This training strategy allows for better adaptation between the LLM and the realism decoder, enabling alignment from language tokens to realism scores through the light-weighted decoder.

Training Loss. Unlike the traditional LLM token classification loss, we employ a l_2 regression loss for the realism decoder. This design allows the network to learn continuous realism scoring information from human annotations. Our loss function is defined as follows:

$$\mathcal{L} = \|F(x) - y_x\|_2, \tag{3}$$

where x is the input mesh shape, y_x is the realism human annotation of input x, $F(\cdot)$ is our metric network, and $\|\cdot\|$ is the l_2 loss.

RealismGrading Dataset Design

We present a human-annotated dataset designed to capture the distortions introduced by real-world 3D reconstruction and generative methods. Our RealismGrading dataset includes human annotations that evaluate the mesh shapes without ground truth reference, and it consists of meshes from real-world 3D reconstruction and generation methods. Our benchmark dataset construction involves three main stages: data generation, annotation, and evaluation.

Data Generation

We begin by selecting 16 image/object pairs sourced from various datasets (Yu et al. 2021; Wang et al. 2022a; Zhu et al. 2023; Yang et al. 2020; 3D 2025; Renderbot 2025; Downs et al. 2022; Sketchfab 2025; CGTrader 2025). For each pair, the input image is processed through 16 different algorithms to create distorted 3D meshes, while the associated object acts as the ground truth. The bject list are as follow: 01-Dog (Renderbot 2025), 02-Fish (3D 2025), 03-Female Face (Zhu et al. 2023; Yang et al. 2020), 04-Male Face (Wang et al. 2022a), 05-Female Human Body A (Yu et al. 2021), 06-Female Human Body B (Yu et al. 2021), 07-Hand w/ Arm (3D 2025), 08-Hand w/o Arm (3D 2025), 09-Male Human Body A (Yu et al. 2021), 10-Male Human Body B (Yu et al. 2021),11-Keyboard (Downs et al. 2022), 12-Building (Jensen et al. 2014), 13-Bus (3D 2025), 14-Mug (Downs et al. 2022), 15-Plant (Sketchfab 2025), and 16-Shoe (Downs et al. 2022), and the distortion algorithms are: CRM (Wang et al. 2024), DreamGaussian (Tang et al. 2023), InstantMesh (Xu et al. 2024a), LGM (Tang

	Object Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Overall
PLCC	NR-3DQA (Zhang et al. 2022)	0.47	0.70	0.59	0.70	0.18	0.51	0.60	0.76	0.41	0.51	0.82	0.55	0.61	0.41	0.34	0.74	0.555
	Ours	0.87	0.31	0.88	0.65	0.9	0.94	0.39	0.69	0.82	0.85	0.76	0.38	0.46	0.67	0.81	0.64	**0.689**
SROCC	Point-based	0.48	0.51	0.87	0.57	0.35	0.62	0.47	0.43	0.56	0.38	0.95	0.52	0.84	0.75	0.49	0.68	0.591
	Ours	0.92	0.44	0.88	0.61	0.74	0.88	0.39	0.7	0.85	0.9	0.89	0.62	0.25	0.51	0.9	0.66	**0.696**
KROCC	Point-based	0.31	0.40	0.72	0.39	0.37	0.44	0.35	0.31	0.42	0.18	0.85	0.39	0.71	0.56	0.33	0.53	0.454
	Ours	0.78	0.37	0.76	0.46	0.57	0.73	0.33	0.6	0.67	0.72	0.7	0.56	0.13	0.4	0.73	0.54	**0.565**

Table 2: We quantitatively compare our metric with an ad-hoc baseline using 3 correlation evaluations: PLCC (Pearson 1920), SROCC (Spearman 1910), and KROCC (Kendall 1938), which are defined in Sec. Evaluation Methods. Number 1-16 indicates the object index (aligns with object number in Tab. 1). Bold numbers indicate the best performance.

Datasets	ShapeGrading	RealismGrading
Number of objects	12	16
Mesh shape types	Synthetic	Real-world
Number of shape distortion types	7	16
Needs ground truth for annotation	Yes	No
Scoring range	$[0, 6]$	$[0, 1]$
Scoring 95% confidence interval	0.303	0.077

Table 3: We compare our dataset with ShapeGrading (Luan et al. 2024). Our datasets do not use ground truth reference for human annotation, and the mesh shape in our dataset comes from real-world 3D reconstruction and generation methods.

et al. 2024), One2345 (Liu et al. 2023), One2345pp (Liu et al. 2024), PeRFlowText (Yan et al. 2024), ShapE (Jun and Nichol 2023). ShapEText (Jun and Nichol 2023), Splatter-Image (Szymanowicz, Rupprecht, and Vedaldi 2024), TripoSR (Tochilkin et al. 2024), ECON (Xiu et al. 2023), ICON (Xiu et al. 2022), PiFU (Saito et al. 2019), PiFUHD (Saito et al. 2020), and HaMeR (Pavlakos et al. 2024). For text-driven models, we first generate descriptive prompts using an image captioning technique and then use these prompts to produce the meshes. In Tab. 1, we provide a complete list of the dataset and 3D reconstruction/generation method we used.

Annotation Process

To obtain reliable distortion scores, we adopt a pairwise comparison method inspired by (Ponomarenko et al. 2009) and (Luan et al. 2024). Each human evaluator compares 9-16 meshes of a specific object category with a given mesh material through 6 rounds of Swiss-system pairwise comparisons. Specifically, for each round, if the evaluator believes A mesh is more realistic than B mesh, then A mesh scores 1 in this round, and B mesh scores 0 in this round. Note that, in this process, the ground truth mesh is not shown to the human evaluators. After all 6 rounds, a mesh earns a score from 0 (losing every comparison) to 6 (winning every comparison). We gathered ratings from 319 subjects and got a total of 5,223 scores. These raw scores are then normalized to a $[0, 1]$ range. The overall reliability of the dataset is confirmed by an average 95% confidence interval of approximately 5%, computed with $\sigma_{\bar{x}} = z_{0.95}\sigma/\sqrt{N}$ (with $z_{0.95} \approx 1.96$).

In Tab. 3, we compare our dataset with the synthetic dataset presented in (Luan et al. 2024). Our datasets do not use ground truth reference for human annotation, and the mesh shape in our dataset comes from real-world 3D reconstruction and generation methods. We also visualize examples along with annotation scores in our dataset in Fig. 3. We observe that the annotated realism score rises as the realism of a mesh shape increases.

Evaluation Methods

To measure how closely our metric aligns with human perception, we employ three correlation coefficients. Pearson's linear correlation coefficient (PLCC) (Pearson 1920) quantifies the linear relationship between our metric and human ratings. PLCC can be represented as:

$$p = \frac{\sum_{i=1}^{n}(\hat{s}_i - \bar{\hat{s}})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n}(\hat{s}_i - \bar{\hat{s}})^2}\sqrt{\sum_{i=1}^{n}(s_i - \bar{s})^2}}, \tag{4}$$

where $\hat{s}_i$ and s_i are the input mesh and ground-truth realism scores of the sample i, respectively, and n is the data sample number. $\bar{\hat{s}} = \frac{1}{n}\sum_{i=1}^{n}\hat{s}_i$ and $\bar{s} = \frac{1}{n}\sum_{i=1}^{n}s_i$ are the mean realism prediction and ground truth realism scores.

Additionally, Spearman's ranking order correlation (SROCC) (Spearman 1910) is also used to measure the ranking order correlation between the shape realism metric and human annotation. SROCC is denoted as:

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(R(\hat{s}_i) - R(s_i))^2}{n(n^2-1)}, \tag{5}$$

where $R(\hat{s}_i)$ and $R(s_i)$ are the rankings of $\hat{s}_i$ and s_i. n denotes. n is the data sample number.

Finally, we use Kendall's rank order correlation coefficient (KROCC) (Kendall 1938) as an alternative way to measure ranking order correlation. Unlike SROCC, which considers the magnitude of rank differences, this metric focuses solely on the relative ordering of ranks. KROCC can be represented as:

$$\tau = 1 - \frac{2}{n(n^2-1)}\sum_{i<j}\text{sign}(\hat{s}_i - \hat{s}_j))\,\text{sign}(s_i - s_j), \tag{6}$$

where $\text{sign}(\cdot)$ is the sign function:

$$\text{sign}(x) = \begin{cases} 1, & x > 0 \\ -1, & x < 0 \\ 0, & x = 0 \end{cases}$$

Note that each coefficient ranges from -1 to 1, with higher values indicating a stronger correlation.

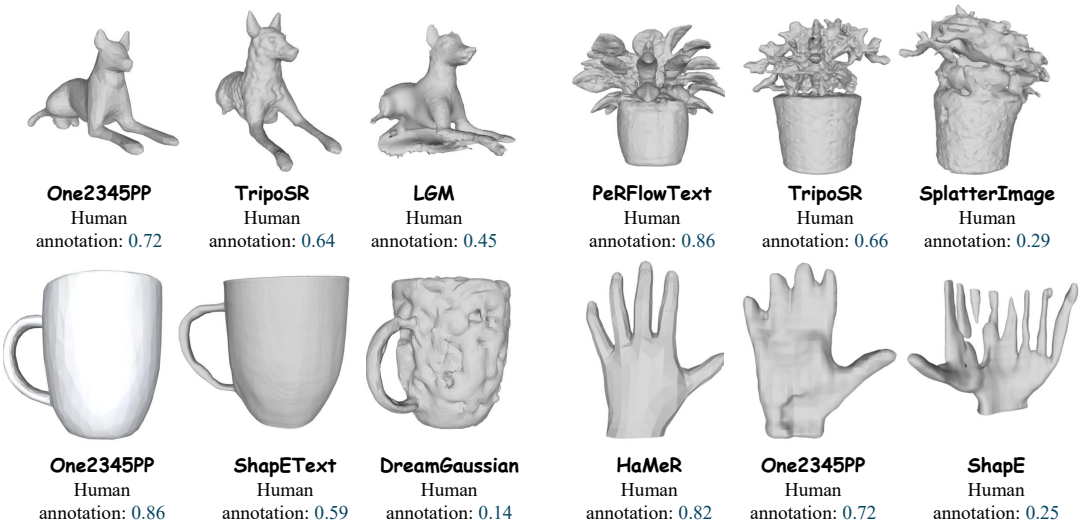

Figure 3: We show example meshes along with their human-annotated realism score from our RealismGrading dataset. Methods used to produce these meshes are shown as well, *e.g.*, "`One2345pp`". We observe that as the realism of a mesh increases, its annotated realism score also goes up.

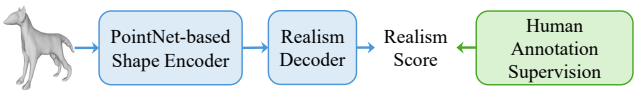

Figure 4: The point-cloud-based ad-hoc baseline design. This baseline employs PointNet (Qi et al. 2017) to extract shape features, which are then fed into the same realism decoder to produce a realism evaluation score.

Experiments

Implementation Details

Following (Xu et al. 2024b), we adopt PointBert (Yu et al. 2022) as our shape encoder. The maximum number of vertices of a mesh is set to be 8192. If there are more than 8192 vertices in a mesh, we use the farthest point sampling method to select 8192 vertices. The LLM bridge is initialized with corresponding components of PointLLM (Xu et al. 2024b). We adopt AdamW (Loshchilov and Hutter 2017) as our optimizer. The learning rate is set to be 2×10^{-4}. We train for 3 epochs and set the batch size to 12.

Evaluation on RealismGrading Dataset

We evaluate the alignment between our metric and human annotations using the three evaluation methods described in Sec. Evaluation Methods. The results are summarized in Tab. 2.

For comparison, we propose an ad-hoc baseline which is based on PointNet. As illustrated in Fig. 4, this baseline extracts shape features with PointNet (Qi et al. 2017). The shape features are then fed into the same realism decoder to generate a realism score.

As shown in Tab. 2, our metric demonstrates better alignment with human perception than the baseline. Our metric achieves PLCC of 0.69, SROCC of 0.70, and KROCC of 0.57. In addition, PLCC, SROCC, and KROCC of our metric are 0.13, 0.11, and 0.11 higher than those of the baseline, respectively. These numbers not only show that our metric is well aligned with human annotations, but also show that our method more faithfully reflects human perception of realism than the baseline. This illustrates the effectiveness of our LLM bridge in evaluating the realism of various 3D shapes and its generalization ability across different types of 3D shapes.

Method		PLCC ↑	SROCC ↑	KROCC ↑
LoRA	r=2	0.608	0.626	0.518
	r=4	0.656	0.639	0.514
	r=8	0.550	0.564	0.441
	r=16	0.524	0.512	0.401
Prefix Tuning		0.081	0.066	0.039
Full Finetune (Ours)		**0.689**	**0.696**	**0.566**

Table 4: Ablation Studies on finetuning methods. We compare three different methods, including two parameter-efficient finetuning methods, *i.e.*, LoRA and prefix tuning, and full finetuning. We also explore how the performance changes with different LoRA ranks (denoted as "`r`"). The results show that full fine-tuning achieves better k-fold validation results and good generalizability than the two-parameter efficient finetuning methods.

Experiment Results

Full finetuning or parameter efficient finetuning. We explore whether parameter-efficient finetuning methods, *e.g.*, LoRA, prefix tuning, could lead to better performance. In Tab. 4, we compare three different finetuning methods: LoRA, prefix tuning, and full finetuning. We can see from the results that prefix tuning leads to poor performance. There is very little correlation between realism scores produced by the model trained via prefix tuning and human an-

notations. In addition, LoRA with ranks of 2 and 4 performs better than larger ranks of 8 and 16, with rank 4 achieving the best performance PLCC (0.656) and SROCC (0.639). However, full finetuning outperforms all LoRA settings across all three metrics. These results show that while LoRA and prefix tuning offer parameter efficiency, finetuning all parameters of the LLM bridge is more beneficial for our 3D realism evaluation task.

Method	PLCC ↑	SROCC ↑	KROCC ↑
Generation	0.245	0.250	0.217
Regression (Ours)	**0.689**	**0.696**	**0.566**

Table 5: Ablation study on realism score generation methods. We compare two methods: "`Generation`" and "`Regression`". For "`Generation`", we train the model to predict *text* tokens corresponding to quantized integer realism scores. For "`Regression`", we adopt the realism decoder to regress realism scores.

Prompt	PLCC ↑	SROCC ↑	KROCC ↑
W/ obj name	0.567	0.559	0.443
W/o obj name (Ours)	**0.689**	**0.696**	**0.566**

Table 6: Ablation study on prompt. We investigate whether including the object name in realism prompt helps our metric achieve better performance. If object name is included in the realism prompt "`Evaluate the quality of this point cloud object and provide your rating...`", we replace "`object`" with an object name, *e.g.*, "`dog`". Otherwise, we keep the original realism prompt. These two prompts are denoted as "`W/ obj name`" and "W/o obj name (Ours)", respectively.

Realism score: generation or regression. We investigate how the realism score should be produced. Specifically, we compare two methods: "`Generation`" and "`Regression`". For "`Generation`", we train the model to predict **text** tokens corresponding to quantized integer realism scores (more details can be found in our supplementary materials). For "`Regression`", we adopt the realism decoder introduced in Sec. Shape-Realism Alignment Metric to regress realism scores. We can see from 5 that regressing the realism scores leads to much better performance. Hence, we adopt the "`Regression`" method.

Prompt: Include object name or not. We explore whether or not we should include the object name in the realism prompt. If object name is included in the realism prompt "`Evaluate the quality of this point cloud object and provide your rating...`", we replace "`object`" with an object name, *e.g.*, "`dog`". Otherwise, we keep the original realism prompt. These two prompts are denoted as "`W/ obj name`" and "`W/o obj name (Ours)`", respectively. As shown in Tab. 6, the prompt without object name leads to better alignment of our metric with human annotations. The PLCC, SROCC, and KROCC of "`W/o obj name`" are at least 0.122 more than those of "`W/ obj name`".

Visualization. In Fig. 5, we show some visualization of 3D shapes, realism scores of our metric, and human-annotated realism scores. We can see that our method assigns high realism scores to realistic 3D shapes, while distorted shapes receive low scores. For example, as shown in the second row, the hand generated by HaMeR receives a score of 0.67. The hand generated by PeRFlowText received a lower score, as it has six fingers. The hand generated by One2345 receives the lowest score, as we can barely recognize that it is a hand. These visualizations illustrate that our metric correlates well with human annotations, which reflects how human annotators perceive mesh realism.

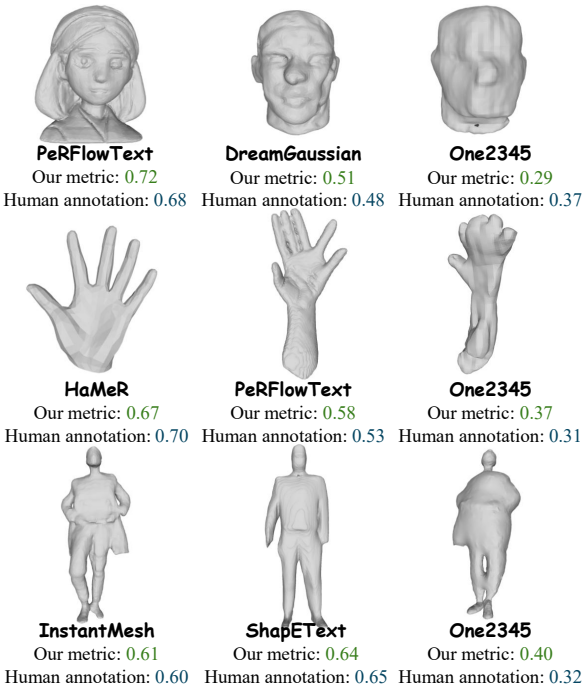

Figure 5: We present the realism scores from our metric alongside human-annotated realism scores for various meshes. The results show that our metric assigns high realism scores to realistic meshes, while severely distorted meshes receive low scores. Our metric correlates well with human annotations, which reflects how human annotators perceive mesh realism.

Conclusion

In conclusion, we have introduced a novel no-reference metric for evaluating 3D shape realism that operates solely on the 3D shape itself. Our method, which leverages a pretrained 3D language model and employs a LoRA-style finetuning approach, effectively integrates human realism annotations to capture high-level semantic features and align its scores with human perception. The introduction of a new dataset containing human-labeled scores for meshes generated by a diverse range of reconstruction and generation algorithms further validates the practical applicability of our approach. Experimental results demonstrate that our metric correlates strongly with human judgments and outperforms existing methods, offering a promising tool for advancing the evaluation of 3D shapes in content creation industries.

References

3D, A. 2025. Artec 3D - Professional 3D scanners.

Achlioptas, P.; Diamanti, O.; Mitliagkas, I.; and Guibas, L. 2018. Learning representations and generative models for 3d point clouds. In *ICML*, 40–49.

Bechtold, J.; Tatarchenko, M.; Fischer, V.; and Brox, T. 2021. Fostering generalization in single-view 3d reconstruction by learning a hierarchy of local and global shape priors. In *CVPR*, 15880–15889.

Borgefors, G. 1984. Distance transformations in arbitrary dimensions. *Computer vision, graphics, and image processing*, 321–345.

Bulbul, A.; Capin, T.; Lavoué, G.; and Preda, M. 2011. Assessing visual quality of 3-D polygonal models. *IEEE Signal Processing Magazine*, 80–90.

CGTrader. 2025. 3D Models for VR / AR and CG projects.

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Chen, Z.; Kim, V. G.; Fisher, M.; Aigerman, N.; Zhang, H.; and Chaudhuri, S. 2021. Decor-gan: 3d shape detailization by conditional refinement. In *CVPR*, 15740–15749.

Corsini, M.; Larabi, M.-C.; Lavoué, G.; Petřík, O.; Váša, L.; and Wang, K. 2013. Perceptual metrics for static and dynamic triangle meshes. In *Comput. Graph. Forum*.

Cui, B.; Yang, Q.; Yang, K.; Xu, Y.; Xu, X.; and Liu, S. 2024. SJTU-TMQA: A quality assessment database for static mesh with texture map. In *ICASSP*, 7875–7879. IEEE.

Downs, L.; Francis, A.; Koenig, N.; Kinman, B.; Hickman, R.; Reymann, K.; McHugh, T. B.; and Vanhoucke, V. 2022. Google scanned objects: A high-quality dataset of 3d scanned household items. In *ICRA*, 2553–2560. IEEE.

Genova, K.; Cole, F.; Sud, A.; Sarna, A.; and Funkhouser, T. 2020. Local deep implicit functions for 3d shape. In *CVPR*.

Gong, X.; Song, L.; Zheng, M.; Planche, B.; Chen, T.; Yuan, J.; Doermann, D.; and Wu, Z. 2023. Progressive Multi-View Human Mesh Recovery with Self-Supervision. In *AAAI*.

Henderson, P.; and Ferrari, V. 2018. Learning to generate and reconstruct 3d meshes with only 2d supervision. *arXiv preprint arXiv:1807.09259*.

Hu, T.; Wang, L.; Xu, X.; Liu, S.; and Jia, J. 2021. Self-Supervised 3D Mesh Reconstruction from Single Images. In *CVPR*, 6002–6011.

Jensen, R.; Dahl, A.; Vogiatzis, G.; Tola, E.; and Aanæs, H. 2014. Large scale multi-view stereopsis evaluation. In *CVPR*, 406–413.

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

Kendall, M. G. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2): 81–93.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *The annals of mathematical statistics*, 79–86.

Lavoué, G. 2009. A local roughness measure for 3D meshes and its application to visual masking. *ACM Transactions on Applied Perception*, 1–23.

Liu, M.; Shi, R.; Chen, L.; Zhang, Z.; Xu, C.; Wei, X.; Chen, H.; Zeng, C.; Gu, J.; and Su, H. 2024. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *CVPR*, 10072–10083.

Liu, M.; Xu, C.; Jin, H.; Chen, L.; Varma T, M.; Xu, Z.; and Su, H. 2023. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In *NeurIPS*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Luan, T.; Li, Z.; Chen, L.; Gong, X.; Chen, L.; Xu, Y.; and Yuan, J. 2024. Spectrum auc difference (saucd): Human-aligned 3d shape evaluation. In *CVPR*, 20155–20164.

Luan, T.; Wang, Y.; Zhang, J.; Wang, Z.; Zhou, Z.; and Qiao, Y. 2021. Pc-hmr: Pose calibration for 3d human mesh recovery from 2d images/videos. In *AAAI*, 2269–2276.

Luan, T.; Zhai, Y.; Meng, J.; Li, Z.; Chen, Z.; Xu, Y.; and Yuan, J. 2023. High Fidelity 3D Hand Shape Reconstruction via Scalable Graph Frequency Decomposition. In *CVPR*.

Luan, T.; Zhai, Y.; Meng, J.; Li, Z.; Chen, Z.; Xu, Y.; and Yuan, J. 2025. Scalable High-Fidelity 3D Hand Shape Reconstruction Via Graph-Image Frequency Mapping and Graph Frequency Decomposition. *IEEE TPAMI*.

Nehmé, Y.; Delanoy, J.; Dupont, F.; Farrugia, J.-P.; Le Callet, P.; and Lavoué, G. 2023. Textured mesh quality assessment: Large-scale dataset and deep learning-based quality metric. *ACM Transactions on Graphics*, 42(3): 1–20.

Nehmé, Y.; Dupont, F.; Farrugia, J.-P.; Le Callet, P.; and Lavoué, G. 2020. Visual quality of 3d meshes with diffuse colors in virtual reality: Subjective and objective evaluation. *IEEE TVCG*, 27(3): 2202–2219.

Nie, Y.; Han, X.; Guo, S.; Zheng, Y.; Chang, J.; and Zhang, J. J. 2020. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *CVPR*, 55–64.

Pavlakos, G.; Shan, D.; Radosavovic, I.; Kanazawa, A.; Fouhey, D.; and Malik, J. 2024. Reconstructing Hands in 3D with Transformers. In *CVPR*.

Pearson, K. 1920. Notes on the history of correlation. *Biometrika*, 25–45.

Ponomarenko, N.; Lukin, V.; Zelensky, A.; Egiazarian, K.; Carli, M.; and Battisti, F. 2009. TID2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 30–45.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 652–660.

Renderbot. 2025. Animal 3D models - by Renderbot LLC.

Saito, S.; Huang, Z.; Natsume, R.; Morishima, S.; Kanazawa, A.; and Li, H. 2019. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2304–2314.

Saito, S.; Simon, T.; Saragih, J.; and Joo, H. 2020. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 84–93.

Santhanam, H.; Doiphode, N.; and Shi, J. 2023. Automated Line Labelling: Dataset for Contour Detection and 3D Reconstruction. In *WACV*, 3136–3145.

Sarvestani, A. S.; Tang, S.; and Wang, Z. 2025. Hybrid-MQA: Exploring Geometry-Texture Interactions for Colored Mesh Quality Assessment. In *CVPR*.

Shu, D. W.; Park, S. W.; and Kwon, J. 2019. 3d point cloud generative adversarial network based on tree structured graph convolutions. In *ICCV*, 3859–3868.

Sketchfab. 2025. Sketchfab - the best 3D viewer on the web.

Song, L.; Gong, X.; Planche, B.; Zheng, M.; Doermann, D.; Yuan, J.; Chen, T.; and Wu, Z. 2022. Pref: Predictability regularized neural motion fields. In *ECCV*, 664–681. Springer.

Spearman, C. 1910. Correlation calculated from faulty data. *British journal of psychology*, 271.

Szymanowicz, S.; Rupprecht, C.; and Vedaldi, A. 2024. Splatter Image: Ultra-Fast Single-View 3D Reconstruction. In *CVPR*.

Tang, J.; Chen, X.; Wang, J.; and Zeng, G. 2022. Point scene understanding via disentangled instance mesh reconstruction. In *ECCV*, 684–701.

Tang, J.; Chen, Z.; Chen, X.; Wang, T.; Zeng, G.; and Liu, Z. 2024. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *ECCV*, 1–18. Springer.

Tang, J.; Ren, J.; Zhou, H.; Liu, Z.; and Zeng, G. 2023. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*.

Tochilkin, D.; Pankratz, D.; Liu, Z.; Huang, Z.; ; Letts, A.; Li, Y.; Liang, D.; Laforte, C.; Jampani, V.; and Cao, Y.-P. 2024. TripoSR: Fast 3D Object Reconstruction from a Single Image. *arXiv preprint arXiv:2403.02151*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Uy, M. A.; Pham, Q.-H.; Hua, B.-S.; Nguyen, D. T.; and Yeung, S.-K. 2019. Revisiting Point Cloud Classification: A New Benchmark Dataset and Classification Model on Real-World Data. In *ICCV*.

Wang, K.; Lavoué, G.; Denis, F.; Baskurt, A.; and He, X. 2010. A benchmark for 3D mesh watermarking. In *Shape Modeling International Conference*, 231–235. IEEE.

Wang, L.; Chen, Z.; Yu, T.; Ma, C.; Li, L.; and Liu, Y. 2022a. FaceVerse: a Fine-grained and Detail-controllable 3D Face Morphable Model from a Hybrid Dataset. In *CVPR*.

Wang, N.; Zhang, Y.; Li, Z.; Fu, Y.; Liu, W.; and Jiang, Y.-G. 2018. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, 52–67.

Wang, Z.; Jia, J.; Wu, H.; Xing, J.; Cai, J.; Meng, F.; Chen, G.; and Wang, Y. 2022b. Groupdancer: Music to multi-people dance synthesis with style collaboration. In *ACM MM*, 1138–1146.

Wang, Z.; Wang, Y.; Chen, Y.; Xiang, C.; Chen, S.; Yu, D.; Li, C.; Su, H.; and Zhu, J. 2024. Crm: Single image to 3d textured mesh with convolutional reconstruction model. In *ECCV*, 57–74. Springer.

Wu, R.; Chen, X.; Zhuang, Y.; and Chen, B. 2020. Multimodal shape completion via conditional generative adversarial networks. In *ECCV*, 281–296.

Wu, X.; Wu, X.; Luan, T.; Bai, Y.; Lai, Z.; and Yuan, J. 2024. Fsc: Few-point shape completion. In *CVPR*, 26077–26087.

Xiu, Y.; Yang, J.; Cao, X.; Tzionas, D.; and Black, M. J. 2023. ECON: Explicit Clothed humans Optimized via Normal integration. In *CVPR*.

Xiu, Y.; Yang, J.; Tzionas, D.; and Black, M. J. 2022. Icon: Implicit clothed humans obtained from normals. In *CVPR*.

Xu, J.; Cheng, W.; Gao, Y.; Wang, X.; Gao, S.; and Shan, Y. 2024a. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*.

Xu, R.; Wang, X.; Wang, T.; Chen, Y.; Pang, J.; and Lin, D. 2024b. PointLLM: Empowering Large Language Models to Understand Point Clouds. In *ECCV*.

Yan, H.; Liu, X.; Pan, J.; Liew, J. H.; Liu, Q.; and Feng, J. 2024. Perflow: Piecewise rectified flow as universal plug-and-play accelerator. *arXiv preprint arXiv:2405.07510*.

Yang, H.; Zhu, H.; Wang, Y.; Huang, M.; Shen, Q.; Yang, R.; and Cao, X. 2020. FaceScape: A Large-Scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction. In *CVPR*.

Yang, Q.; Jung, J.; Deschamps, T.; Xu, X.; and Liu, S. 2024. TDMD: A Database for Dynamic Color Mesh Quality Assessment Study. *IEEE TVCG*.

Yang, Q.; Jung, J.; Wang, H.; Xu, X.; and Liu, S. 2023. Tsmd: A database for static color mesh quality assessment study. In *VCIP*, 1–5. IEEE.

Yu, T.; Zheng, Z.; Guo, K.; Liu, P.; Dai, Q.; and Liu, Y. 2021. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *CVPR*, 5746–5756.

Yu, X.; Tang, L.; Rao, Y.; Huang, T.; Zhou, J.; and Lu, J. 2022. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *CVPR*, 19313–19322.

Zhang, J.; Wang, Y.; Zhou, Z.; Luan, T.; Wang, Z.; and Qiao, Y. 2021. Learning dynamical human-joint affinity for 3d pose estimation in videos. *IEEE TIP*.

Zhang, Z.; Sun, W.; Min, X.; Wang, T.; Lu, W.; and Zhai, G. 2022. No-reference quality assessment for 3D colored point cloud and mesh models. *IEEE TCSVT*, 32(11): 7618–7631.

Zhao, S.; Wang, Z.; Luan, T.; Jia, J.; Zhu, W.; Luo, J.; Yuan, J.; and Xi, N. 2025. PP-Motion: Physical-Perceptual Fidelity Evaluation for Human Motion Generation. In *ACM MM*.

Zhu, H.; Yang, H.; Guo, L.; Zhang, Y.; Wang, Y.; Huang, M.; Wu, M.; Shen, Q.; Yang, R.; and Cao, X. 2023. FaceScape: 3D Facial Dataset and Benchmark for Single-View 3D Face Reconstruction. *TPAMI*.