Textured Geometry Evaluation: Perceptual 3D Textured Shape Metric via 3D Latent-Geometry Network

Tianyu Luan[1], Xuelu Feng[1], Zixin Zhu[1*], Phani Nuney[1], Sheng Liu[1], Xuan Gong[1 2], David Doermann[1], Chunming Qiao[1], Junsong Yuan[1]

[1] State University of New York at Buffalo
[2] Harvard Medical School
{tianyulu, xuelufen, zixinzhu, phaneesw, sliu66, xuangong, doermann, qiao, jsyuan} @buffalo.edu

Abstract

Textured high-fidelity 3D models are crucial for games, AR/VR, and film, but human-aligned evaluation methods still fall behind despite recent advances in 3D reconstruction and generation. Existing metrics, such as Chamfer Distance, often fail to align with how humans evaluate the fidelity of 3D shapes. Recent learning-based metrics attempt to improve this by relying on rendered images and 2D image quality metrics. However, these approaches face limitations due to incomplete structural coverage and sensitivity to viewpoint choices. Moreover, most methods are trained on synthetic distortions, which differ significantly from real-world distortions, resulting in a domain gap. To address these challenges, we propose a new fidelity evaluation method that is based directly on 3D meshes with texture, without relying on rendering. Our method, named *Textured Geometry Evaluation* TGE, jointly uses the geometry and color information to calculate the fidelity of the input textured mesh with comparison to a reference colored shape. To train and evaluate our metric, we design a human-annotated dataset with real-world distortions. Experiments show that TGE outperforms rendering-based and geometry-only methods on real-world distortion dataset.

Introduction

3D reconstruction and generation is a field with broad applications, covering scenarios such as video games, AR/VR, and film production. High-fidelity 3D shapes with texture are crucial in representing the spatial content in these 3D applications. Current methods focus primarily on reconstructing and generating high-fidelity 3D shapes, and have already produced many widely appreciated results. However, most of the widely used shape fidelity evaluation metrics, such as Chamfer Distance (Borgefors 1984), are still not aligning with human perception well. Therefore, evaluating the fidelity of textured 3D shapes in a human-like manner can better support the generation of high-fidelity 3D models that align with user preferences.

There has been recent progress in fidelity evaluation metrics for textured 3D shapes that are aiming to align with human perception. Those works, such as (Nehmé et al. 2023) would largely rely on rendering. As shown in Fig. 1, these methods typically render 3D shapes into 2D images and evaluate shape fidelity based on the viewpoints of the rendered image. Although such methods can leverage 2D image quality assessment techniques to evaluate 3D shapes, they have limitations. Specifically, rendered images cannot fully reflect the overall structure of a 3D shape. To comprehensively assess the fidelity of a 3D shape, multi-view and full-coverage rendering is required, which significantly increases both the computational cost and the uncertainty of evaluation with viewpoint choice. In addition, most existing methods (e.g., (Nehmé et al. 2023)) rely on training datasets constructed from synthetic distortions. However, the distortions in synthetic data differ substantially from those encountered in real-world reconstruction and generation methods, leading to a significant domain gap. Models trained on synthetic datasets would experience substantial performance degradation when applied to real-world 3D fidelity evaluations.

To address the problems introduced by 2D-rendering-based evaluation and the domain gap caused by synthetic data, we propose a dataset constructed from real data and a 3D evaluation metric based directly on 3D meshes and textures without rendering. First, to eliminate the uncertainty introduced by viewpoint difference, our network is designed to take both 3D mesh and texture as input without reliance on viewpoints. Second, during the 3D fidelity annotation process, we rely on human subjects' perception of fidelity, with each object evaluated by multiple subjects. To reduce subjective inconsistency caused by varying viewpoints, we continuously rotate each 3D object across multiple angles and adjust lighting directions during the evaluation. This ensures that subjects can comprehensively observe the shape under various directions and make their fidelity evaluations more consistently. Additionally, to eliminate the domain gap caused by synthetic data, all distortions in our dataset are generated by different real-world reconstruction and generation algorithms. Through such training, our model can robustly evaluate 3D shapes without being largely affected by viewpoint changes or domain gap.

We name our method *Textured Geometry Evaluation* (TGE). Our network encodes both the input colored shape and the reference shape into the latent feature, and compares them in the latent space to determine the fidelity of the input shape. Specifically, we are inspired by PointNet++

*Corresponding author

	Ours	Graphic-LPIPS			
Input	3D distort mesh	Multiple view 2D images			
Output	0.5168	0.1593	0.2312	0.1981	0.2055
Score range	[0, 1]	[0, 1]			
View Consistency	✔	✘			

Figure 1: Comparison between previous rendering-based evaluation and our proposed 3D latent-geometry-based metric. Prior works, such as Graphic-LPIPS, rely on rendering 3D shapes into 2D images and assessing fidelity using image-based quality metrics, which makes this approach sensitive to viewpoint. Such methods produce inconsistent scores depending on rendering conditions (Right). In contrast, our method directly operates on textured 3D meshes, without relying on rendering (Left).

(Qi et al. 2017) and design a multi-stage Latent-Geometry network to extract features from colored meshes. We introduce a cross-attention mechanism to encode the color feature and fuse it with geometric features, resulting in shape representations that incorporate both texture and geometry. Using this encoding strategy on both the input and reference shapes, we are able to map the input and reference shapes into the same latent space. A fidelity comparison module is then used to compute the fidelity difference between the input and reference shapes, giving the final fidelity score for the input shape. We train and evaluate this network using our provided Colored Shape Fidelity dataset. With this design, the proposed fidelity shape evaluation metric does not rely on viewpoints and aligns well with human perceptual fidelity evaluation based on real-world data distribution.

Our contributions are summarized as follows:

- We propose a fully 3D-based fidelity evaluation metric for textured 3D shapes, without the need for rendering, which avoids the uncertainty introduced by viewpoint changes.
- We design a 3D-geometry-based architecture that jointly encodes geometric and color features of 3D shapes, enabling accurate fidelity representation and evaluation.
- To train and validate our metric, we design a human-annotated Colored Shape Fidelity dataset. We create distortions with real 3D reconstruction and generation techniques to minimize domain bias and display each model from diverse viewpoints to reduce viewpoint ambiguity in the annotation process.

Extensive experiments demonstrate that our proposed method achieves superior performance on multiple real-world reconstruction and generation datasets, significantly outperforming traditional 2D-rendering-based and geometry-only comparison methods.

Related Work

Metrics for 3D Generation and Reconstruction. Evaluating the fidelity of 3D shapes is crucial for assessing the performance of reconstruction and generation methods such as (Borgefors 1984; Hong et al. 2023; Ye et al. 2024; Tang et al. 2025; Luan et al. 2021; Zhai et al. 2023; Luan et al. 2023; Gong et al. 2022; Zhao et al. 2025; Luan et al. 2024a, 2025; Gong et al. 2023; Wu et al. 2024; Song et al. 2022; Zhang et al. 2021; Yang et al. 2025; Huang et al. 2025; Cui et al. 2024b; Reka, Pulli, and Vincze 2025; Xiang et al. 2024; Liu et al. 2024; Lee, Savva, and Chang 2024; Liu et al. 2021; Liu, Nie, and Hamid 2022). Traditional metrics such as Chamfer Distance (CD) (Borgefors 1984) and Earth Mover's Distance (EMD) (Rubner, Tomasi, and Guibas 1998) are widely used in the 3D reconstruction approaches. CD measures the average closest point distance between two point sets, and EMD provides a more accurate assessment by computing the minimal cost of transforming one distribution into another. Learning metrics that are based on CD are also proposed, including Learnable Chamfer Distance (LCD) (Huang et al. 2024), Density-aware Chamfer Distance (DCD) (Wu et al. 2021), and Hyperbolic Chamfer Distance (HyperCD) (Lin et al. 2023). Despite these advancements, many of these metrics focus solely on geometric discrepancies and often overlook the perceptual aspects of texture and color, which are essential for human-aligned evaluations. In 3D shape generation, evaluation metrics play a crucial role in quantifying the similarity between generated results and ground-truth shapes. Metrics such as Fréchet Inception Distance (FID) have been adapted from 2D image synthesis to 3D domains by projecting shapes into 2D renderings (Achlioptas et al. 2018). Such projection-based evaluations suffer from viewpoint dependency and may fail to reflect the structural and textural fidelity of the original 3D shapes. Our work addresses this gap by proposing a metric that directly operates on textured 3D shapes without relying on rendering or projection.

Recent Advances in 3D Shape Evaluation Metrics. To bridge the gap between computational metrics and human perception, recent research has introduced evaluation methods that incorporate both geometric and appearance features. Graphics-LPIPS (Nehmé et al. 2023) extends the Learned

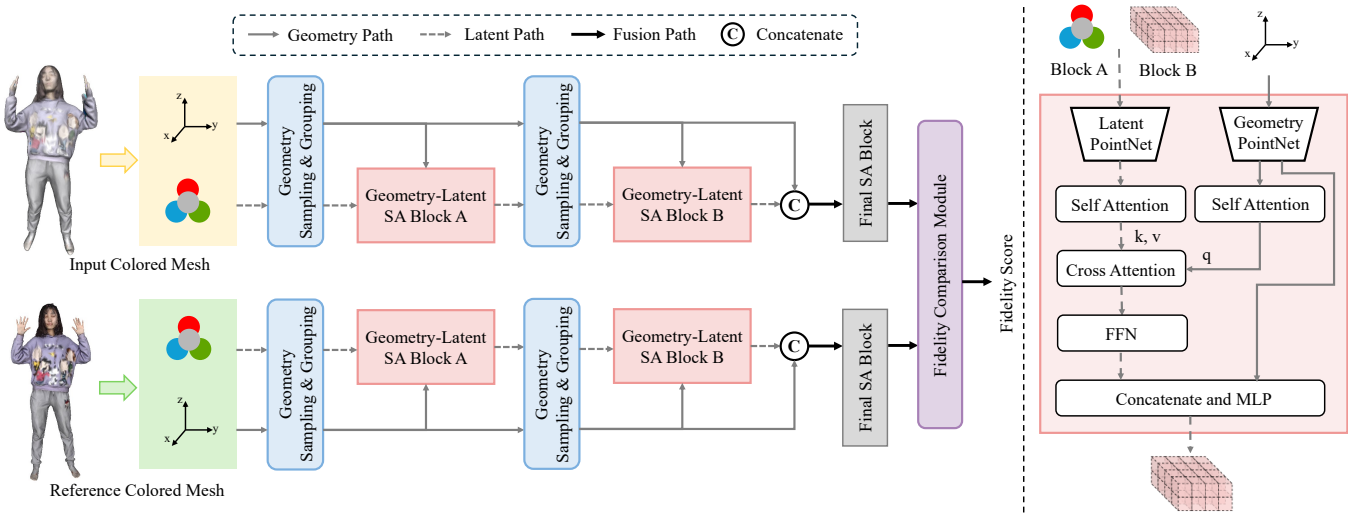


Figure 2: (a) Overview of the TGE pipeline. Given a pair of textured 3D meshes (input and reference), we extract hierarchical geometry and color features using a PointNet++-style pipeline. A novel Latent-Geometry Set Abstraction (LG-SA) block is introduced to jointly fuse geometry and color information at each level. The resulting global features from both meshes are compared by a shared MLP to predict a scalar fidelity score. This design allows perceptual fidelity evaluation without any rendering. (b) Illustration of the Latent-Geometry Set Abstraction (LG-SA) block. The module extracts geometry and appearance features via parallel self-attention modules and fuses them using a cross-attention mechanism. Geometry features serve as the query to attend to latent color features.

Perceptual Image Patch Similarity (LPIPS) metric to 3D graphics by assessing the perceptual similarity of rendered images using an LPIPS (Zhang et al. 2018)-based approach. SJTU-TMQA (Cui et al. 2024a) and TSMD (Yang et al. 2023) are datasets designed for textured mesh quality assessment, providing benchmarks for evaluating the visual quality of 3D models. CMDM (Nehmé et al. 2021) introduces a full-reference metric that combines curvature-based geometric features with color information to assess the quality of colored meshes. While these methods represent significant strides toward perceptually aligned evaluations, they often rely on 2D renderings of 3D models, making them susceptible to variations in viewpoint and lighting conditions. Furthermore, the dependence on synthetic distortions in some datasets may not accurately reflect the complexities encountered in real-world scenarios, highlighting the need for evaluation metrics that operate directly on 3D data and are trained on real-world distortions.

Methods

Problem Formulation

We aim to design a human-aligned fidelity metric that quantifies the perceptual similarity between an input textured 3D shape and a reference shape. Given an input 3D mesh $\hat{m}$ and its corresponding ground-truth mesh m, our metric $F(\hat{m}, m; \theta)$ with learnable parameters θ outputs a scalar fidelity score $\hat{s} \in \mathbb{R}$:

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

where a higher score indicates higher perceptual similarity. Each mesh $m = \{(v_i, c_i)\}_{i=1}^{N}$ consists of N vertices, with $v_i \in \mathbb{R}^3$ as the 3D coordinate and $c_i \in \mathbb{R}^3$ as the RGB vertex color. During training, we minimize the discrepancy between predicted scores and human-annotated ground-truth scores s:

$$\min_{\theta} \mathcal{L}(\hat{s}, s), \tag{2}$$

where $\mathcal{L}$ is a hybrid loss function designed to align with both absolute and relative score accuracy (see Section).

Pipeline Overview

To align with human perceptual judgment and overcome the limitations of rendering-based evaluations, we propose the *Textured Geometry Evaluation* (TGE) framework. As shown in Fig. 2(a), our model directly processes 3D textured meshes without rendering, and extracts features through a two-stream design that jointly models geometry and color. Our framework is designed based on PointNet++ (Qi et al. 2017). First, given a pair of 3D meshes $(\hat{m}, m)$, where $\hat{m}$ is the input mesh to be evaluated and m is the reference, we first sample points from each mesh and extract local geometric features from vertex coordinates $\{v_i\}$ using sampling and grouping from (Qi et al. 2017). For each level of the Set Abstraction block, we design a novel *Latent-Geometry Set Abstraction* (LG-SA) block, in which the new LG-SA block uses both color and geometry information in the latent space, while the original SA block only considers the 3D geometry information. This design is motivated by our observation that human perception relies on both shape and appearance cues. By explicitly fusing these two streams, we aim to bridge the gap between geometric alignment and perceptual realism. In Sec. Latent-Geometry Set Abstraction (LG-SA)

Block, we will elaborate on the detailed design of our LG-SA block. After 2 layers of sampling & grouping along with LG-SA block, we use the original final SA block in (Qi et al. 2017) to get the encoded feature of the input mesh. For the referenced mesh, we use the same network to extract the 3D geometry-color feature.

After feature extraction, we obtain extracted feature $\mathbf{f}_{\text{input}}$ and $\mathbf{f}_{\text{ref}}$ for the distorted and reference meshes, respectively. These features are passed into a fidelity comparison module $\mathcal{C}$:

$$s = \mathcal{C}(\mathbf{f}_{\text{input}}, \mathbf{f}_{\text{ref}}), \tag{3}$$

where $\mathcal{C}$ is a multi-layer perceptron that predicts the final scalar fidelity score s. This score reflects the perceptual alignment between the input and reference shapes, taking into account both geometry and texture in a rendering-free manner. In summary, our pure 3D-based structure is carefully designed to address the 2 major challenges in 3D textured mesh fidelity evaluation: avoiding viewpoint sensitivity caused by rendering, and jointly utilizing texture and 3D geometry as an essential perceptual cue.

Latent-Geometry Set Abstraction (LG-SA) Block

To effectively combine geometry and appearance features, we design the *Latent-Geometry Set Abstraction (LG-SA)* block, which is a core component of our network. This block extends the original Set Abstraction (SA) module in PointNet++ (Qi et al. 2017) by introducing color-aware fusion in the latent space, thus enabling perceptually informed encoding of textured 3D shapes.

Each LG-SA block receives a set of 3D point coordinates $\{v_i\}_{i=1}^N$ and their corresponding RGB vertex colors $\{c_i\}_{i=1}^N$ as input. These inputs are processed in the geometry path and the latent feature path. For geometry path, we first use the standard farthest point sampling (FPS) and ball query (Qi et al. 2017) to select a subset of points and their local neighborhoods. Each neighborhood is encoded using a PointNet module to extract geometry features $\mathbf{g}_i \in \mathbb{R}^3$ for each sampled point. To further enhance local spatial relations, we apply a self-attention mechanism:

$$\mathbf{g}'_i = \text{Attn}(\mathbf{g}_i, \mathbf{g}_i, \mathbf{g}_i). \tag{4}$$

Similarly, we encode latent feature $\mathbf{l}_i \in \mathbb{R}^{d_l}$ using another PointNet followed by a self-attention module to yield latent features $\mathbf{l}'_i \in \mathbb{R}^{d_l}$. The color-based self-attention is defined as:

$$\mathbf{l}'_i = \text{Attn}(\mathbf{l}_i, \mathbf{l}_i, \mathbf{l}_i). \tag{5}$$

Here, d_l is the latent feature dimension. For the first LG-SA block, the input feature is just the RGB color and $d_l = 3$. For the second LG-SA block, the input feature is the output feature of the first LG-SA block, where $d_l = 256$.

After obtaining the refined features $\mathbf{g}'_i$ and $\mathbf{l}'_i$, we perform cross-attention to allow geometric structure to attend to appearance cues:

$$\mathbf{f}_i = \text{CrossAttn}(\mathbf{g}'_i, \mathbf{l}'_i, \mathbf{l}'_i) + \text{FFN}(\mathbf{l}'_i), \tag{6}$$

where $\mathbf{g}'_i$ is used as query $\mathbf{q}_i$, and $\mathbf{l}'_i$ as key and value. This cross-attention allows each point's geometric understanding to be modulated by its corresponding appearance context, important for capturing visual distortions. The final fused feature $\mathbf{f}_i$ is then passed through an MLP and aggregated via max pooling to obtain a global shape-level descriptor. This design is inherently asymmetrical: geometry guides the attention, and appearance complements it. This design also reflects how human perception evaluates 3D shapes: structural integrity is the foundation, while texture is the most crucial content of the fidelity evaluation.

Importantly, this LG-SA block avoids dependency on mesh connectivity or 2D projection. By integrating local reasoning (via self-attention) and cross-modal modulation (via cross-attention), the LG-SA block captures shape and appearance interaction in a way that aligns with perceptual fidelity.

Training Strategy

Our training strategy is fully supervised and end-to-end, leveraging human-labeled fidelity scores from a dataset that includes real-world distortions. To ensure that the predicted fidelity score $\hat{s}$ aligns not only numerically but also highly correlated with human labels s, we optimize a hybrid loss function:

$$\mathcal{L} = \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{plcc}}\mathcal{L}_{\text{plcc}} + \lambda_{\text{srocc}}\mathcal{L}_{\text{srocc}}, \tag{7}$$

where The weights $\lambda_{\text{smooth}}, \lambda_{\text{plcc}}, \lambda_{\text{srocc}}$ are weight hyperparameters, and:

Smooth L1 loss

$$\mathcal{L}_{\text{smooth}} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}\frac{1}{2}(\hat{s}_i - s_i)^2, & \text{if } |\hat{s}_i - s_i| < 1\\ |\hat{s}_i - s_i| - \frac{1}{2}, & \text{otherwise}\end{cases} \tag{8}$$

This loss penalizes numerical discrepancies. Here, N is the number of data sample, $\hat{s}_i$ and s_i are the predicted and ground-truth scores for the i-th sample.

Pearson's correlation loss (PLCC Loss)

$$\mathcal{L}_{\text{plcc}} = 1 - \frac{\sum_{i=1}^{N}(\hat{s}_i - \bar{\hat{s}})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{N}(\hat{s}_i - \bar{\hat{s}})^2}\sqrt{\sum_{i=1}^{N}(s_i - \bar{s})^2}}, \tag{9}$$

This loss maximizes the linear correlation between predictions and human labels. $\bar{\hat{s}}$ and $\bar{s}$ are batch-wise means.

Spearman's rank-order correlation loss (SROCC Loss)

$$\mathcal{L}_{\text{srocc}} = 1 - \frac{6\sum_{i=1}^{N}(R(\hat{s}_i) - R(s_i))^2}{N(N^2-1)} \tag{10}$$

This term promotes correct ranking order, which is crucial when judging relative fidelity. $R(\cdot)$ is the soft-ranking operator as in (Blondel et al. 2020), making the function differentiable.

Colored Shape Fidelity Dataset

To evaluate the perceptual fidelity of textured 3D shapes in realistic settings, we construct a new human-annotated dataset, denoted as Colored Shape Fidelity. Unlike prior benchmarks that rely on synthetic distortions or rendering-based supervision, our dataset is grounded in real-world 3D reconstruction and generation pipelines, annotated through

(a) Reference Mesh Examples (b) Distorted Mesh Examples

Figure 3: Some examples in our Colored Shape Fidelity dataset: (a) Referenced meshes. (b) Distorted meshes reconstructed or generated from real-world methods.

Method	1	2	3	4	5	6	7	8	9	10	11	Average↑	Std↓
CD	0.6667	0.4959	0.7288	0.8090	0.7987	0.4820	0.9003	0.8191	0.5386	-0.0538	0.8046	0.6355	0.2570
IoU	0.2295	0.6756	0.5788	0.4970	0.4865	0.0943	0.8734	0.6475	0.5247	0.5705	**0.8427**	0.5473	0.2205
F-score	0.1738	0.0081	0.5147	0.7430	0.6970	0.4409	0.7770	0.7861	0.1562	0.1058	0.7429	0.4678	0.2907
P2S	0.7736	0.2731	0.6993	**0.8334**	**0.8263**	0.4206	0.8461	0.7987	0.4821	-0.0363	0.7252	0.6038	0.2729
ND	0.2091	0.4168	0.3633	0.7252	0.5080	0.0341	0.2543	0.2740	-0.3038	-0.0663	0.0335	0.2031	0.2803
UHD	0.7523	0.5299	0.4460	0.2933	0.3440	0.4479	0.7956	0.6987	0.5159	-0.2840	0.7359	0.4796	0.2908
G-LPIPS	0.7280	0.7399	0.7089	0.6173	0.8128	0.5663	**0.9151**	0.6291	**0.8923**	0.5205	0.3584	0.6808	0.1575
Ours	**0.8358**	**0.8387**	**0.7311**	0.7951	0.8024	**0.8132**	0.6585	**0.8737**	0.8812	**0.6424**	0.8042	**0.7887**	**0.0758**

Table 1: Per-object evaluation results across 11-fold object-level cross-validation using PLCC. Each object is used once as the held-out test set to assess generalization. Our method consistently achieves the highest overall performance and the lowest standard deviation across all metrics, demonstrating superior perceptual alignment and generalizability. The range of PLCC is [-1, 1], and higher values indicate stronger correlations. **Bold** number means the best.

rigorous perceptual protocols. The construction process involves three stages: distortion collection, perceptual scoring, and reliability verification.

Real-World distortion collection. We begin by selecting a diverse set of object categories and corresponding reference meshes from multiple 3D datasets, including Function4D (Yu et al. 2021), ARC3D (Artec3D 2025), RenderBot (Downs et al. 2022), Sketchfab (Sketchfab 2025), and CGTrader (CGTrader 2025). These objects serve as high-fidelity references for evaluating distortion quality. For each object, we generate distorted meshes using a set of state-of-the-art reconstruction and generation methods, spanning from neural implicit reconstruction (Xiu et al. 2022, 2023; Saito et al. 2020) to recent text-guided generative models (Tang et al. 2023; Wang et al. 2024; Xu et al. 2024; Tochilkin et al. 2024). For text-based models, image captioning tools are first used to generate prompts, which are then fed into text-to-3D models to generate shape variants. This procedure ensures that the distortions reflect actual artifacts in real systems, rather than artificial degradations.

Perceptual annotation protocol. To ensure perceptual consistency in scoring, we adopt a pairwise comparison protocol following the Swiss tournament system used in (Luan et al. 2024b). For each object-material combination, subjects compare the distorted meshes over 6 rounds of head-to-head matchups, where each mesh competes against dynamically selected peers. Scores are assigned based on win count, ranging from 0 (loses all rounds) to 6 (wins all rounds), and subsequently normalized to the $[0, 1]$ range. In total, we collect ratings from 180 unique human subjects across all object-material pairs. Each pair is evaluated by around 20 participants to reduce individual bias. In total, 1,696 annotations are obtained. This pairwise approach enforces score differentiation and prevents clustering artifacts common in absolute scoring settings.

Outlier handling and confidence estimation. To enhance annotation reliability, we detect and exclude outlier scores using the interquartile range (IQR) method (Dekking et al. 2005). Any score deviating more than $1.5\times$IQR from the 25^{th} or 75^{th} percentile is removed. This simple but robust strategy eliminates 7.0% of extreme annotations. We further quantify the consistency of annotations by computing the 95% confidence interval of each object's mean score using:$\sigma_{\bar{x}} = \frac{z_{0.95}\cdot\sigma}{\sqrt{N}}$, where σ is the sample standard deviation, N is the number of subjects per object, and $z_{0.95} \approx 1.96$ is the z-score for a 95% confidence level. Our statistic shows that the IOR process reduces the overall 95% confidence interval from 0.0823 to 0.067.

Evaluation metrics. To evaluate how well automated metrics align with human perception, we adopt three standard correlation measures: Pearson Linear Correlation Coefficient (PLCC) (Pearson 1920), Spearman Rank Correlation Coefficient (SROCC) (Spearman 1910), and Kendall Rank Correlation Coefficient (KROCC) (Kendall et al. 1946). These metrics respectively capture linear relationships, rank consistency, and ordinal agreement between predicted scores and human annotations. Each measure outputs values in the range $[-1, 1]$, with higher values indicating stronger perceptual alignment. In Fig. 3 we show a few examples of the groundtruth and distortions in the dataset.

Metric	1	2	3	4	5	6	7	8	9	10	11	Average↑	Std↓
CD	0.6167	0.5523	0.6071	0.5238	0.7381	0.7381	0.8286	0.7143	0.7381	0.0238	0.7866	0.6243	0.2113
IoU	0.3167	0.7782	0.7500	0.2857	0.2169	0.0476	**0.9429**	0.5952	0.5476	0.5714	**0.9624**	0.5468	0.2871
F-score	0.0833	0.0187	0.5988	0.2857	-0.1325	0.3810	0.2571	0.6667	0.0238	0.1429	0.7280	0.2776	0.2737
P2S	0.6667	0.5523	0.5714	0.5714	-0.0361	0.1667	0.9429	0.7381	0.7381	-0.2381	0.7364	0.4918	0.3502
ND	0.4333	0.5439	0.4643	0.7143	0.3253	0.0714	0.2571	0.0714	0.0952	-0.0952	0.3849	0.2969	0.2312
UHD	0.5833	0.5105	0.7143	0.1190	0.3494	0.3810	0.7714	0.5238	0.7857	-0.1429	0.7029	0.4817	0.2766
G-LPIPS	0.6333	0.7667	0.7748	0.2410	0.6506	0.4458	0.9276	0.7274	0.6752	0.5988	0.3530	0.6177	0.1911
Ours	**0.8833**	**0.8167**	**0.8469**	**0.8193**	**0.8470**	**0.8193**	0.7247	**0.9092**	**0.8593**	**0.6587**	0.9160	**0.8273**	**0.0731**

Table 2: Per-object evaluation results across 11-fold object-level cross-validation using SROCC. The range of SROCC is [-1, 1]. Higher values indicate stronger correlations. **Bold** number means the best.

Metric	1	2	3	4	5	6	7	8	9	10	11	Average↑	Std↓
CD	0.3889	0.3662	0.5238	0.2857	0.6429	0.5000	0.7333	0.5714	0.5000	-0.0714	0.7043	0.4677	0.2161
IoU	0.1667	**0.6480**	0.6190	0.2857	0.1482	0.0000	**0.8667**	0.4286	0.3571	0.4286	**0.8733**	0.4384	0.2746
F-score	0.1111	0.0000	0.5822	0.2143	-0.1482	0.2857	0.2000	0.5000	0.0000	0.0714	0.5916	0.2189	0.2372
P2S	0.4444	0.3662	0.4286	0.3571	0.0000	0.1429	0.8667	0.6429	0.5000	-0.2143	0.6480	0.3802	0.2944
ND	0.3333	0.4789	0.2381	0.5714	0.2224	0.0714	0.2000	0.0714	0.1429	-0.0714	0.2535	0.2284	0.1757
UHD	0.5000	0.3662	0.5238	0.0714	0.3706	0.2857	0.6000	0.4286	0.6429	-0.1429	0.5353	0.3801	0.2250
G-LPIPS	0.5000	0.6111	0.5855	0.2224	0.5401	0.3706	0.8281	0.5669	0.5401	0.4001	0.2858	0.4955	0.1608
Ours	**0.7778**	0.6111	**0.6831**	**0.6671**	**0.7715**	**0.6671**	0.5521	**0.7937**	**0.7715**	**0.4728**	0.8003	**0.6880**	**0.1033**

Table 3: Per-object evaluation results across 11-fold object-level cross-validation using KROCC. The range of KROCC is [-1, 1]. Higher values indicate stronger correlations. **Bold** numbers are the best.

Experiments

Implementation Details

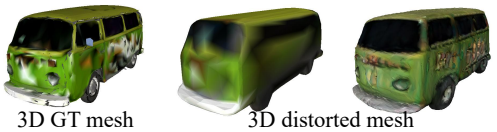
We follow the architectural design of the multi-scale grouping (MSG) variant of PointNet++ (Qi et al. 2017). Our fidelity module comprises a three-layer multilayer perceptron (MLP) with hidden dimensions of 1024, 512, and 256, respectively. For the attention module, each head is followed by a feed-forward network (FFN) consisting of two linear layers: the first expands the embedding dimension d_{emb} to $4d_{\text{emb}}$, followed by a ReLU activation, and the second projects it back to d_{emb}. Here, d_{emb} denotes the output channel of the last MLP layer in the PointNet-based encoder. During training, we minimize a weighted sum of three loss terms: smooth loss, PLCC loss, and SROCC loss. The corresponding weights λ_{smooth}, λ_{plcc}, and λ_{srocc} are set to 1, 0.2, and 0.2, respectively. The model is trained on a single NVIDIA RTX A6000 GPU using the AdamW optimizer (Loshchilov and Hutter 2017), with a learning rate of 1×10^{-3} and a weight decay of 1×10^{-4}. We use a batch size of 3. The implementation is based on PyTorch (Paszke et al. 2019).

Experiment Results

Human alignment evaluation and generalization analysis. To rigorously assess the generalization ability of our proposed metric, we adopt an 11-fold object-level cross-validation strategy. Specifically, we treat each of the 11 objects in our dataset as the held-out test set in turn, while training the model on the remaining 10 categories. This object-wise split ensures that the model is always tested on *unseen objects*, thereby validating the metric's performance along with generalizability on shapes. For each fold, we compute three commonly used perceptual correlation metrics, PLCC, SROCC, and KROCC, between the predicted scores and human annotations on the testing object. We report both the per-object results and the average performance across all folds, along with the standard deviation to reflect score consistency. The comparison methods include: Chamfer Distance (CD) (Borgefors 1984), Intersection of Union (IoU) (Henderson and Ferrari 2018), F-score (Wang et al. 2018), Plane-to-Surface (P2S), Normal Difference, Unidirectional Haudorff Distance (UHD) (Wu et al. 2020), and Graphic-LPIPS (G-LPIPS) (Nehmé et al. 2023).

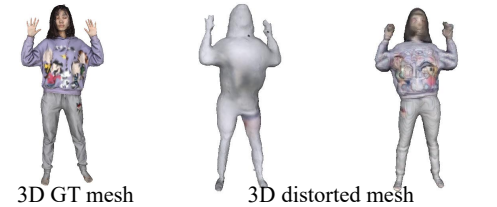
As shown in Tabs. 1 to 3, our method achieves the highest mean performance across all three metrics, while also maintaining the lowest standard deviations. This demonstrates not only superior prediction quality but also strong reliability across different shape types. By contrast, classical geometry-based metrics such as Chamfer Distance (CD) and IoU yield significantly higher variance and lower overall correlation with human ratings. Learning-based metrics like G-LPIPS show higher accuracy but also suffer from performance fluctuations across folds. These results highlight the importance of integrating texture and geometry in a 3D-native manner, as implemented in our method. Our evaluation setup and consistent cross-object results collectively demonstrate that the proposed fidelity metric is generalizable, stable, and better aligned with perceptual evaluation than existing alternatives.

Module necessity. In Tab. 5, we examine different strategies for module design. The 1st row "w/o Attention & Latent" is removing both the attention mechanism and the latent branch, leaving only the geometry branch, and we simply altered the original geometry Set Abstraction module from PointNet++ from 3 to 6 channels to take color input. The result justifies our design of the LG-SA module. The 2nd row "w/o Attention" would remain the Latent branch design, but only remove the attention parts. This result justifies the necessity of our Attention module. The 3rd row "w/o Self-Attention" has cross attention design, but removes the self-attention part. The result shows that the self-attention

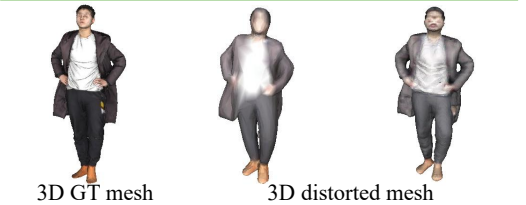


3D GT mesh	3D distorted mesh	
GLPIPS	0.4312	0.4327
Ours	0.5222	0.7469
Human Annotation	0.6250	0.6667

3D GT mesh	3D distorted mesh	
GLPIPS	0.3062	0.3502
Ours	0.6183	0.7055
Human Annotation	0.6667	0.7500



3D GT mesh	3D distorted mesh	
GLPIPS	0.3033	0.3051
Ours	0.3221	0.5973
Human Annotation	0.3333	0.6667



3D GT mesh	3D distorted mesh	
GLPIPS	0.2750	0.2903
Ours	0.5161	0.6710
Human Annotation	0.5833	0.7500

Figure 4: Visualized comparison of our metric vs. the previous metric G-LPIPS. Our metric aligns better with human annotation compared to the previous metric.

Metrics	CD	IoU	F-score	P2S	ND	UHD	G-LPIPS (Learnable)	Ours (Learnable)
GFLOPs	1.6	0.084	2.0	2.0	3.0	0.8	93.11	14.7

Table 4: Computational complexity comparison of different metrics.

Metric	PLCC ↑	SROCC ↑	KROCC ↑
w/o Attention & Latent	0.6180	0.6568	0.5366
w/o Attention	0.7153	0.7035	0.5924
w/o Self-attention	0.5589	0.6320	0.5099
w/o Geometry feature	0.6519	0.6464	0.5649
Ours	**0.7887**	**0.8273**	**0.6880**

Table 5: Ablation study: impact of modules.

Loss function	PLCC ↑	SROCC ↑	KROCC ↑
$\mathcal{L}_{smooth}$	0.7742	0.7456	0.6155
$+\mathcal{L}_{plcc} + \mathcal{L}_{srocc}$	0.7296	0.7383	0.5949
$+0.2\mathcal{L}_{plcc} + 0.2\mathcal{L}_{srocc}$ (Ours)	**0.7887**	**0.8273**	**0.6880**

Table 6: Ablation study: impact of loss functions/weights.

modules before the cross-attention module are necessary. The 4th row "w/o Geometry feature" means not concatenating the geometry feature back with the latent feature after the LG-SA module. The result indicates that even when used as the query for the cross-attention module in the geometry branch, the geometry feature would still be crucial for later encoding. Overall, the ablation proves the necessity of all proposed components in achieving optimal perceptual evaluation.

Loss function impacts. In Tab. 6, we investigate the effect of weighting in the composite loss function. Using only the Smooth L1 loss ($\mathcal{L}_{\text{smooth}}$) yields strong results. If adding a small correlation objectives ($\mathcal{L}_{\text{plcc}}, \mathcal{L}_{\text{srocc}}$) will boost the human alignment performance.

Computational complexity. We compare the computational complexity of various 3D fidelity metrics in terms of GFLOPs (billion floating point operations). The number of vertices is uniformly set to 10,000. For IoU, the resolution is set to $256 \times 256 \times 256$, and for G-LPIPS, the image size is set to 1600×1600, which are both the same as Tabs. 1 to 3. Traditional geometry-based metrics such as Chamfer Distance (CD), IoU, and F-score are lightweight. In contrast, recent learning-based methods like G-LPIPS incur significantly higher computational costs, reaching 93.11 GFLOPs. Our proposed method achieves a good trade-off, requiring only 14.7 GFLOPs while achieving better human alignment in fidelity. This demonstrates that our design is not only perceptually effective but also computationally efficient, making it suitable for real-world applications.

Visualization. We visualize some results comparison of our metric vs. the previous metric, G-LPIPS. For G-LPIPS, the lower the better. For our metric, the higher the better. As observed, our metric aligns better with human annotation compared to the previous metric.

Conclusion

We present *Textured Geometry Evaluation* (TGE), a human-aligned fidelity metric that evaluates textured 3D meshes directly without rendering. Unlike prior methods that rely on 2D projections or synthetic distortions, TGE jointly encodes geometry and color to assess perceptual fidelity against a reference mesh. We construct a new human-annotated dataset featuring real-world distortions to train and validate our method. Extensive experiments demonstrate that TGE achieves better alignment with human evaluation than previous rendering-based and geometry-only approaches.

References

Achlioptas, P.; Diamanti, O.; Mitliagkas, I.; and Guibas, L. 2018. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, 40–49. PMLR.

Artec3D. 2025. Artec 3D - Professional 3D scanners. https://www.artec3d.com/3d-models.

Blondel, M.; Teboul, O.; Berthet, Q.; and Djolonga, J. 2020. Fast differentiable sorting and ranking. In *ICML*, 950–959.

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

CGTrader. 2025. CGTrader - 3D Models for VR / AR and CG projects. https://www.cgtrader.com/free-3d-models/scanned/various/25-storey-building.

Cui, B.; Yang, Q.; Yang, K.; Xu, Y.; Xu, X.; and Liu, S. 2024a. SJTU-TMQA: A quality assessment database for static mesh with texture map. In *ICASSP 2024-2024 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 7875–7879. IEEE.

Cui, R.; Song, X.; Sun, W.; Wang, S.; Liu, W.; Chen, S.; Shang, T.; Li, Y.; Barnes, N.; Li, H.; et al. 2024b. LAM3D: Large Image-Point Clouds Alignment Model for 3D Reconstruction from Single Image. *Advances in Neural Information Processing Systems*, 37: 4454–4480.

Dekking, F. M.; Kraaikamp, C.; Lopuhaä, H. P.; and Meester, L. E. 2005. *A Modern Introduction to Probability and Statistics: Understanding why and how*, volume 488. Springer.

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.

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*.

Gong, X.; Zheng, M.; Planche, B.; Karanam, S.; Chen, T.; Doermann, D.; and Wu, Z. 2022. Self-supervised Human Mesh Recovery with Cross-Representation Alignment. In *ECCV*.

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

Hong, Y.; Zhang, K.; Gu, J.; Bi, S.; Zhou, Y.; Liu, D.; Liu, F.; Sunkavalli, K.; Bui, T.; and Tan, H. 2023. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*.

Huang, T.; Liu, Q.; Zhao, X.; Chen, J.; and Liu, Y. 2024. Learnable Chamfer Distance for point cloud reconstruction. *Pattern Recognition Letters*, 178: 43–48.

Huang, Z.; Boss, M.; Vasishta, A.; Rehg, J. M.; and Jampani, V. 2025. SPAR3D: Stable Point-Aware Reconstruction of 3D Objects from Single Images. *arXiv preprint arXiv:2501.04689*.

Kendall, M. G.; et al. 1946. The advanced theory of statistics. *The advanced theory of statistics*.

Lee, H.; Savva, M.; and Chang, A. X. 2024. Text-to-3D Shape Generation. In *Computer Graphics Forum*, volume 43, e15061. Wiley Online Library.

Lin, F.; Yue, Y.; Hou, S.; Yu, X.; Xu, Y.; Yamada, K. D.; and Zhang, Z. 2023. Hyperbolic chamfer distance for point cloud completion. In *Proceedings of the IEEE/CVF international conference on computer vision*, 14595–14606.

Liu, P.; Wang, Y.; Sun, F.; Li, J.; Xiao, H.; Xue, H.; and Wang, X. 2024. Isotropic3D: Image-to-3D generation based on a single clip embedding. *arXiv preprint arXiv:2403.10395*.

Liu, S.; Nie, X.; and Hamid, R. 2022. Depth-guided sparse structure-from-motion for movies and tv shows. In *CVPR*, 15980–15989.

Liu, S.; Song, L.; Xu, Y.; and Yuan, J. 2021. Nech: neural clothed human model. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*, 1–5. IEEE.

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

Luan, T.; Gao, Z.; Xie, L.; Sharma, A.; Ding, H.; Planche, B.; Zheng, M.; Lou, A.; Chen, T.; Yuan, J.; et al. 2024a. Divide and Fuse: Body Part Mesh Recovery from Partially Visible Human Images. In *European Conference on Computer Vision*, 350–367. Springer.

Luan, T.; Li, Z.; Chen, L.; Gong, X.; Chen, L.; Xu, Y.; and Yuan, J. 2024b. Spectrum auc difference (saucd): Human-aligned 3d shape evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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*, 16795–16804.

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.; Abid, M.; Lavoué, G.; Da Silva, M. P.; and Le Callet, P. 2021. Cmdm-vac: Improving a perceptual quality metric for 3d graphics by integrating a visual attention complexity measure. In *2021 IEEE International Conference on Image Processing (ICIP)*, 3368–3372. IEEE.

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.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32.

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

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Reka, M.; Pulli, T.; and Vincze, M. 2025. Multi-Modal 3D Mesh Reconstruction from Images and Text. *arXiv preprint arXiv:2503.07190*.

Rubner, Y.; Tomasi, C.; and Guibas, L. J. 1998. A metric for distributions with applications to image databases. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, 59–66. IEEE.

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.

Sketchfab. 2025. Sketchfab - the best 3D viewer on the web. https://sketchfab.com/.

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.

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*.

Tang, Z.; Zhang, J.; Cheng, X.; Yu, W.; Feng, C.; Pang, Y.; Lin, B.; and Yuan, L. 2025. Cycle3d: High-quality and consistent image-to-3d generation via generation-reconstruction cycle. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7320–7328.

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*.

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.; 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, 281–296. Springer.

Wu, T.; Pan, L.; Zhang, J.; Wang, T.; Liu, Z.; and Lin, D. 2021. Density-aware chamfer distance as a comprehensive metric for point cloud completion. *arXiv preprint arXiv:2111.12702*.

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

Xiang, J.; Lv, Z.; Xu, S.; Deng, Y.; Wang, R.; Zhang, B.; Chen, D.; Tong, X.; and Yang, J. 2024. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*.

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*, 13286–13296.

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

Yang, J.; Sax, A.; Liang, K. J.; Henaff, M.; Tang, H.; Cao, A.; Chai, J.; Meier, F.; and Feiszli, M. 2025. Fast3R: Towards 3D Reconstruction of 1000+ Images in One Forward Pass. *arXiv preprint arXiv:2501.13928*.

Yang, Q.; Jung, J.; Wang, H.; Xu, X.; and Liu, S. 2023. Tsmd: A database for static color mesh quality assessment study. In *2023 IEEE international conference on visual communications and image processing (VCIP)*, 1–5. IEEE.

Ye, J.; Liu, F.; Li, Q.; Wang, Z.; Wang, Y.; Wang, X.; Duan, Y.; and Zhu, J. 2024. Dreamreward: Text-to-3d generation with human preference. In *European Conference on Computer Vision*, 259–276. Springer.

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.

Zhai, Y.; Huang, M.; Luan, T.; Dong, L.; Nwogu, I.; Lyu, S.; Doermann, D.; and Yuan, J. 2023. Language-guided human motion synthesis with atomic actions. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5262–5271.

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, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

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*, 6840–6849.