WHO IS A BETTER DREAMER: SUBJECTIVE AND OBJECTIVE QUALITY ASSESSMENT FOR TEXT-TO-3D GENERATED CONTENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-3D content generation, often referred as "Dreamer," has emerged as a rapid solution for 3D modeling, yet current systems still suffer from pronounced quality deficiencies that undermine user experience. To systematically evaluate the quality of 3D generated contents (3DGCs) and to inform the future development of Dreamers, we present a comprehensive subjective and objective assessment. We benchmark 11 representative Dreamers using 484 prompts and collect 4,691 3DGCs, forming 3DGCQA-LS, the largest 3DGC quality assessment dataset to date. Through controlled multi-subject evaluation, we uncover substantial performance discrepancies across existing Dreamers and identify nine recurring distortion types. To further enable automated quality prediction, we introduce Hyper-VG, a novel 3DGC quality assessment framework that integrates the Huanyuan world model, persistent homotopy theory, and cube projection. Extensive experiments demonstrate that Hyper-VG achieves state-of-the-art (SOTA) performance across multiple benchmark datasets, providing a strong foundation for advancing reliable Dreamers. This work will be made open source to advance the field.

1 INTRODUCTION

With the rapid advancement of the metaverse and virtual/augmented reality (V/AR) Zhang et al. (2023c;d; 2025b), 3D content has become a fundamental medium for bridging the physical and digital worlds, enabling a wide range of applications in industrial manufacturing, healthcare, and film and entertainment Du et al. (2025); Circelli (2025). However, conventional 3D content creation pipelines remain labor-intensive and technically demanding, typically involving multiple stages. These processes require substantial domain expertise and are inherently limited in terms of design efficiency and scalability. As shown in Fig. 1, generative artificial intelligence (GAI) Zhou et al. (2023a) has emerged as a transformative paradigm for 3D content creation, substantially improving efficiency. In particular, Text-to-3D approaches Tripo (2025); Lai et al. (2025); Poole et al. (2022); Zhu et al. (2023), which directly convert descriptions into 3D assets, have attracted considerable attention and are often referred to as "Dreamers." Despite their promise, current Dreamers are constrained and the 3D generated contents (3DGCs) frequently exhibits severe quality defects, which significantly degrade visual fidelity and user experience. Consequently, the development of reliable methods for perceiving and evaluating quality



Figure 1: The critical importance of conducting quality assessments for 3DGCs. Effective and reliable evaluation metrics can guide algorithm optimisation and enhance user experience.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

issues in 3DGCs produced by different Dreamers is essential for guiding model optimization and improving downstream applications.

To address this problem, we conduct a comprehensive quality assessment of 3DGCs generated by representative Dreamers. First, to ensure broad coverage and diversity, we select 11 representative Text-to-3D models and design a set of carefully curated prompts. Specifically, 484 distinct prompts spanning 10 categories are employed to systematically evaluate the generative capabilities of different Dreamers. After filtering out failed generations, we collect a total of 4,691 valid 3DGCs, resulting in the construction of 3DGCQA-LS, the largest large-scale dataset for 3DGC quality assessment to date. To capture authentic user perceptual experience, we further conduct subjective experiments on the collected 3DGCs. The experimental results reveal a series of recurring quality issues across existing Dreamers and demonstrate substantial performance disparities. These findings provide strong empirical evidence of the necessity of rigorous quality assessment for 3DGCs. Building upon these insights, we propose Hyper-VG, a novel objective quality assessment framework for 3DGCs. The proposed method leverages world models Huang et al. (2025) and persistent homotopy Zomorodian & Carlsson (2004) to enable reliable and comprehensive evaluation. Extensive experiments across multiple 3DGCQA datasets show that Hyper-VG achieves state-of-the-art (SOTA) performance, offering robust objective metrics that can effectively support the optimization and development of Dreamers. In summary, the main contributions of this paper are threefold:

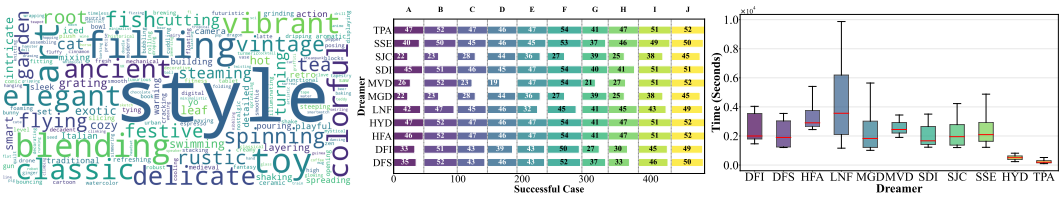
- We construct 3DGCQA-LS, a large-scale dataset for 3DGC quality assessment, comprising 4,691 3DGCs generated by 11 representative Dreamers using 484 diverse prompts. To the best of our knowledge, this is the largest dataset currently available for evaluating 3DGC quality.
- We conduct extensive subjective experiments on the 3DGCQA-LS dataset, identifying nine prevalent quality issues in Text-to-3D generation and performing a comprehensive analysis of the factors that influence 3DGC quality from multiple perspectives.
- We propose Hyper-VG, a novel 3DGC quality assessment framework that integrates structural and projection-based analyses to evaluate quality from geometric, topological, visual, spatial, and prompt-consistency perspectives. Experimental results demonstrate its SOTA performance across multiple datasets.

2 Related Works

2.1 Dreamer: Text-to-3D Generation Methods

Text-to-3D methods represent a highly challenging direction in GAI, aiming to synthesize 3DGCs directly from textual prompts. Owing to their cross-modal generative capability, these models are commonly referred to as "Dreamers." Based on their underlying 3D representations, Dreamers can be categorized into explicit surface, implicit surface, and tri-plane methods. Explicit representations ♠ Tsalicoglou et al. (2024); Wei et al. (2023); Liu et al. (2024a); Wu et al. (2024) directly model geometry and texture, enabling intuitive editing and efficient rendering. Recently, 3D Gaussian Splatting Kerbl et al. (2023) has emerged as a novel explicit representation paradigm, further advancing the expressive capacity and rendering efficiency of Dreamers Di et al. (2025); He et al. (2024); Xu et al. (2023). In contrast, implicit representations ♣ Wang et al. (2023c); Huang et al. (2023); Kim et al. (2024); Chen et al. (2023) learn continuous functions over 3D space to encode properties such as color and density, offering strong detail fidelity at the cost of high computational overhead. Tri-plane representations Hong et al. (2024); Li et al. (2023); Liu et al. (2024b); Yi et al. (2023) provide a hybrid solution by projecting 3D information onto three orthogonal feature planes, achieving a balance between expressive capacity and computational efficiency while supporting efficient multi-view rendering.

From an evaluation perspective, commonly used metrics such as Fréchet Inception Distance (FID) Heusel et al. (2017) and CLIP-based Radford et al. (2021) similarity are widely adopted to measure visual fidelity and semantic consistency of 3DGCs. However, existing evaluation metrics exhibit notable inconsistencies with human perceptual judgments, limiting their effectiveness in guiding model optimization. Consequently, the development of objective quality assessment methods that are well aligned with human visual perception is of critical importance for enabling Dreamers to generate 3DGC that better satisfies user expectations and application requirements.



(a) Prompt Word Cloud (b) Successful Generation Cases (c) Generation Time

Figure 2: The visualization of statistical results detailing the constructed 3DGCQA-LS dataset. The letters A–J in subfigure (b) denote *Artifacts*, *Beverages*, *Ceramics*, *Desserts*, *Electronics*, *Pets*, *Spices*, *Textiles*, *Tools* and *Toys*.

2.2 AI-Generated Content Quality Assessment

With the rapid development of GAI, AI-generated content (AIGC) has emerged as a transformative paradigm for media creation. According to media modality, mainstream AIGCs can be broadly classified into AI-generated images (AGIs) Podell et al. (2023); Rombach et al. (2022); Ko et al. (2023), AI-generated videos (AGVs) Blattmann et al. (2023); Singer et al. (2022); An et al. (2023), and 3DGCs Lin et al. (2023); Shi et al. (2023); Lukoianov et al. (2024). Despite remarkable progress, existing GAI still suffer from notable quality deficiencies and frequently fail to meet user expectations. Consequently, systematic quality assessment of AIGC is of substantial practical importance, as it provides critical guidance for improving generative models and enhancing users' visual experiences. As summarised in Table **??**, a range of representative quality assessment datasets and objective evaluation frameworks have been proposed across different AIGC modalities. The distribution and scale of existing datasets indicate that current research efforts predominantly focus on AGIs and AGVs, thereby driving rapid methodological advances in these domains. Moreover, evaluation paradigms for AGIs and AGVs are gradually shifting from general-purpose benchmarks toward application-specific assessment protocols Zhou et al. (2025a); Lekadir et al. (2021); Khan & Jion (2025), enabling more targeted evaluation.

In contrast, quality assessment for 3DGCs remains relatively underexplored. Evaluating 3DGCs poses dual challenges, namely the difficulty of large-scale data collection and the high computational complexity Zhang et al. (2025d; 2024a). Although several 3DGC quality assessment datasets Zhou et al. (2025d); Fu et al. (2025); Zhang et al. (2025a) have been introduced, they are limited in both scale and diversity. Furthermore, most existing objective 3D quality assessment methods Zhang et al. (2024b;c; 2023a) rely primarily on 2D projections, largely overlooking intrinsic geometric, structural, and spatial characteristics. To address these limitations, we construct the large-scale 3DGCQA-LS dataset and propose Hyper-VG, an objective quality assessment framework grounded in world models Huang et al. (2025) and persistent homotopy Zomorodian & Carlsson (2004). Together, these contributions enable a more comprehensive and perceptually aligned evaluation of 3DGCs.

3 Database Construction

3.1 Prompt Design

To enable a comprehensive evaluation of the generative quality of Dreamers, we design a total of 484 textual prompts for 3D content generation. These prompts are organized into ten semantic categories: *Pets*, *Beverages*, *Toys*, *Tools*, *Electronics*, *Artifacts*, *Ceramics*, *Textiles*, *Desserts*, and *Spices*. Collectively, these categories span a wide range of application domains, including everyday consumer products, professional design and manufacturing, as well as cultural and artistic content, thereby closely reflecting the practical demands of real-world 3D content applications. Each category comprises between 41 and 54 prompts, ensuring both content diversity and balanced representation across categories. This design not only enriches the variability of the dataset but also enhances the robustness and reliability of the resulting evaluation. To further characterize the properties of the designed prompts, we conduct statistical analyses of word frequency and semantic diversity, as visualized in Fig. 2(a).



Figure 3: Visualization of 3DGCs and representative distortions in the 3DGCQA-LS dataset. The upper panel presents the detailed information of the 3DGCQA-LS dataset, where the left and right sides (labeled A–J) represent ten different categories of prompts. Colors denote different prompt categories. The bottom panel illustrates typical distortion cases, demonstrating 9 distortion types.

3.2 3D Content Generation

To ensure the comprehensiveness of the proposed 3DGCQA-LS dataset, we select 11 representative Dreamers (listed in Table 1) to generate 3DGCs in response to 484 carefully designed prompts. For the closed-source Dreamers, 3DGCs are generated directly via their official web interfaces. For the remaining nine open-source Dreamers, we employ the publicly released source code provided by the respective authors, using default experimental settings to ensure fair and reproducible comparisons. Owing to differences in robustness and generalization capabilities, certain Dreamers exhibit sensitivity to specific prompts, leading to failed generation cases. To quantitatively analyze generative robustness, we categorize and statistically evaluate the successfully generated samples, with results reported in Fig. 2(b) and Table 1. The analysis reveals two key observations: 1) Substantial variability exists in generative robustness across different Dreamers: the closed-source Dreamers successfully generate 3DGCs for all prompts, demonstrating strong stability, whereas SJC Wang et al. (2023a) and MGD Lin et al. (2023) exhibit the lowest success rates, achieving only 67.6%; 2) Prompt-wise analysis indicates that existing open-source Dreamers encounter notable difficulties in generating content within the *Artifact* and *Textile* categories. As shown in Fig. 3, 4,691 valid 3DGCs generated by 11 Dreamers are collected to construct the dataset.

Table 1: Details of Dreamers employed. Symbols ♠ and ♣ are defined in Sec. 2.1, and $*$ denotes closed-source Dreamer. Sample size represents successfully generated cases.

Label	Methods	Type	Sample Size
TPA	Tripo AI* Tripo (2025)	♠	484/484
HYD	Hunyuan3D* Lai et al. (2025)	♠	484/484
SSE	SketchShape Sanghi et al. (2023)	♠	461/484
MGD	Magic3D Lin et al. (2023)	♠	327/484
SJC	Score Jacobian Chaining Wang et al. (2023a)	♠	327/484
DFI	DreamFusion-IF Poole et al. (2022)	♣	410/484
DFS	DreamFusion-SD Poole et al. (2022)	♣	437/484
HFA	HiFA Zhu et al. (2023)	♣	483/484
LNF	LatentNeRF Metzer et al. (2023)	♣	435/484
MVD	MVDream Shi et al. (2023)	♣	372/484
SDI	Score Distillation Lukoianov et al. (2024)	♣	471/484

Additionally, we investigate generation efficiency by recording the time required to produce each 3DGC. The corresponding statistical results are presented in Fig. 2(c). For closed-source Dreamers, we report the generation time observed through their web interfaces, whereas for open-source Dreamers, both generation and runtime measurements are conducted on servers equipped with 8 × A100 GPUs. The results demonstrate pronounced disparities in generation efficiency: open-source Dreamers generally incur substantially longer generation times, averaging approximately 2,000 seconds per sample, while closed-source Dreamers exhibit markedly higher efficiency.

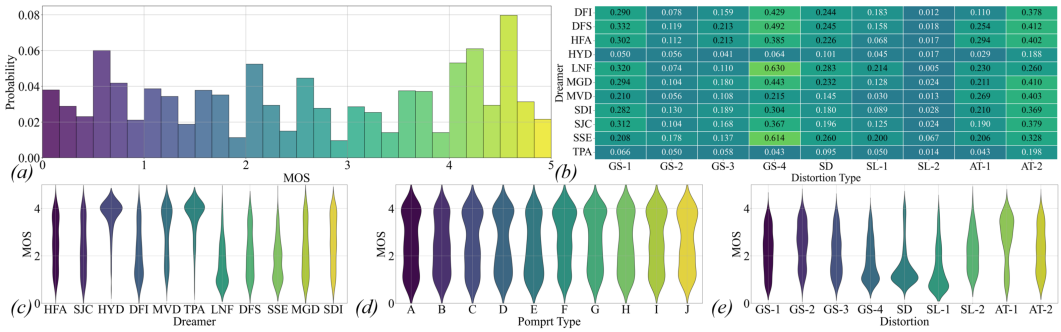


Figure 4: Statistical visualization of subjective evaluation results. (a) Distribution of Mean Opinion Scores (MOS); (b) Distortion occurrence rates across different Dreamers; (c–f) Effects of various influencing factors on the MOS distribution.

3.3 Subjective Experiment

To obtain reliable visual quality feedback, we recruit 40 participants (20 male and 20 female) to conduct subjective quality assessments on the 4,691 3DGCs in the 3DGCQA-LS dataset. Following the ITU-R BT.500-13 recommendations ITU (2002), all subjective experiments are performed in a well-controlled laboratory environment. To comprehensively present the 3DGCs and capture realistic viewing experiences, we adopt the lightweight subjective evaluation protocol for 3DGCs proposed by Zhou *et al.* Zhou et al. (2024). Specifically, all 3DGCs are preloaded into Blender Blender (2024), and Python scripts are used to sequentially and randomly display individual project files while recording participants' evaluation scores. All stimuli are presented through an iMac display with a resolution of 4096 × 2304. This interactive viewing setup allows participants to freely manipulate viewpoints, thereby enabling a more immersive assessment process and yielding subjective evaluations that better reflect real-world usage scenarios.

To reduce visual fatigue and dizziness associated with prolonged 3DGC observation, the subjective experiment is divided into 50 sessions, each consisting of no more than 100 assessment tasks. Participants are required to take a mandatory rest of at least 20 minutes between consecutive sessions and are limited to completing a maximum of 4 sessions per day. Prior to the formal evaluation, all participants undergo 30 minutes of training to familiarize themselves with the assessment criteria and the evaluation interface, ensuring consistency and reliability across subjective ratings.

3.4 Data Processing

Through the conducted subjective experiments, we collect a total of 187,640 = 40 × 4,691 individual evaluations, corresponding to 40 participants assessing 4,691 3D generative contents (3DGCs). Each evaluation consists of a subjective quality score q_{ij} assigned by the i-th participant to the j-th 3DGC, together with an annotated distortion label D_{ij}. Accordingly, each assessment can be represented as the tuple $\{q_{ij}, D_{ij}\}$. Notably, D_{ij} is a 9-dimensional binary distortion vector, where each element indicates the presence or absence of a specific quality defect in the corresponding 3DGC. To aggregate subjective ratings across participants and mitigate individual scoring biases, the raw quality scores q_{ij} are first normalized using z-score transformation:

$$z_{ij} = \frac{q_{ij} - \mu_i}{\sigma_i}, \tag{1}$$

where $\mu_i = \frac{1}{N_i}\sum_{j=1}^{N_i} q_{ij}$, $\sigma_i = \sqrt{\frac{1}{N_i - 1}\sum_{j=1}^{N_i}(q_{ij} - \mu_i)^2}$, with N_i denoting the number of 3DGCs evaluated by participant i. Following the rejection procedure specified in ITU-R BT.500-13 ITU (2002), ratings from unreliable participants are excluded. Ultimately, annotations from 33 participants are retained for analysis. The resulting Fleiss' Kappa coefficient of 0.84 demonstrates a high level of inter-rater agreement. The remaining normalized scores are then linearly rescaled to the range [0, 5], and the Mean Opinion Score (MOS) for each 3DGC is computed as the average of the rescaled z-scores. For distortion annotation, a majority-voting strategy is adopted: a distortion type is considered present in the j-th 3DGC only if it is identified by more than half of the participating subjects.

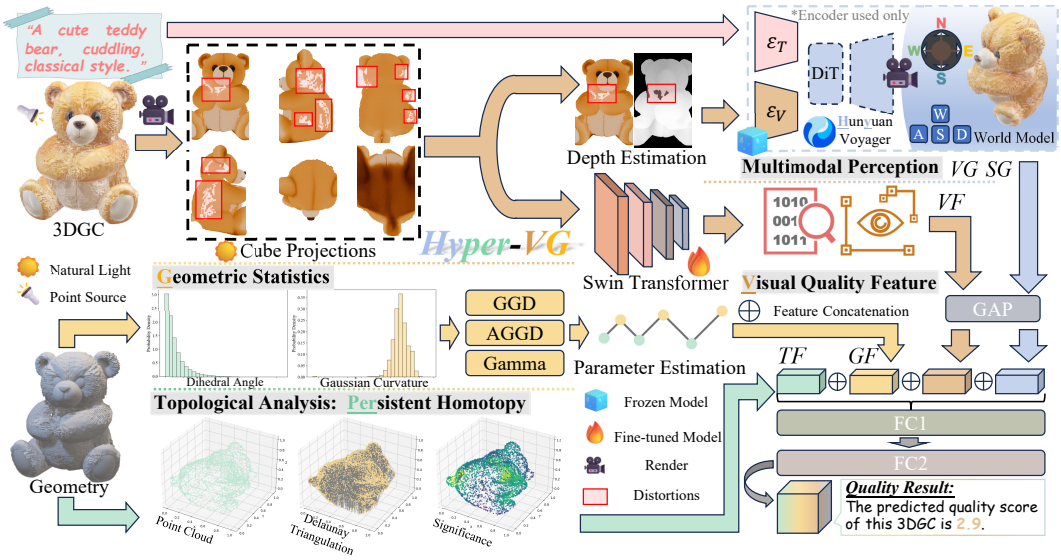


Figure 5: Framework of Hyper-VG. Detailed cubic projection process is discussed in Appendix C.

3.5 Mean Opinion Score Analysis

To facilitate a more intuitive interpretation of the subjective evaluation outcomes, we perform a comprehensive statistical analysis of the MOS distribution within the 3DGCQA-LS dataset, as illustrated in Fig. 4. Several key observations can be drawn from this analysis: 1) The overall MOS distribution approximates a uniform distribution, indicating that the constructed dataset contains a well-balanced range of 3DGCs spanning diverse quality levels. This balanced distribution supports the representativeness and generalizability of the dataset, while also highlighting substantial quality disparities among 3DGCs generated by existing Dreamers, thereby underscoring the necessity of systematic quality assessment; 2) Pronounced differences in generation quality are observed across Dreamers. Among the evaluated methods, HYD Lai et al. (2025) and TPA Tripo (2025) achieve the highest subjective quality scores, whereas LNF Metzer et al. (2023) consistently produces lower-quality 3DGCs; 3) Prompts from different semantic categories exhibit similar MOS distributions, suggesting stable generative performance across prompt types. This consistency further corroborates the robustness and general applicability of the proposed 3DGCQA-LS dataset.

3.6 Distortion Visualization & Analysis

Beyond assigning quality scores to 3DGC, a fine-grained statistical analysis of specific quality defects can provide more targeted insights for optimizing Dreamers. To this end, we identify and define 9 common distortion types (refer to Appendix A for more details) observed across Dreamers based on subjective experiments, as illustrated in Fig. 3. In addition, Fig. 4 presents the distortion occurrence rates for different Dreamers, enabling a quantitative comparison of their generative performance. Several key observations can be drawn from Fig. 4: 1) HYD Lai et al. (2025) and TPA Tripo (2025) exhibit the lowest distortion frequencies, whereas LNF Metzer et al. (2023) demonstrates the highest distortion rate, which is consistent with the trends observed in Sec. 3.5; 2) A distortion-wise analysis indicates that rough structures (GS-4) are the most prevalent defect across the majority of Dreamers, highlighting a critical area for further optimization; 3) Depth error (SD) has a pronounced impact on user experience, while semantic mismatch (SL-1) leads to the most severe degradation in perceived quality. This is because semantic mismatches cause 3DGCs to deviate from user expectations.

4 Proposed Method: Hyper-VG

4.1 Visual Quality Feature Extraction

The visual appearance of 3DGCs directly stimulates the human visual system and therefore constitutes the most intuitive dimension for quality assessment. As illustrated in Fig. 5, we first apply a cube projection to each 3DGC to obtain 6 projections, which are subsequently processed by Swin

Transformer Liu et al. (2021) for feature extraction:

$$VF = \frac{1}{6}\sum_{k=1}^{6} SwinT(I_k), \tag{2}$$

where, I_k denotes the k-th projection obtained via cube projection. The Swin Transformer, denoted by $SwinT(\cdot)$, is employed to extract high-level visual representations, and the resulting visual quality feature VF is computed by averaging the features across all projections.

4.2 World-Model-Based Multimodal Perception

To mitigate the loss of spatial information introduced by cube projection, we incorporate a world model with strong spatial reasoning capabilities for 3DGC multimodal perception. Specifically, we utilize the depth estimation and encoding modules of Hunyuan Voyager Huang et al. (2025) to extract multimodal features from both cube projections and their corresponding textual prompts. Each of the six projected views is first processed by a depth estimator to recover depth cues, and the resulting depth map is concatenated with the original RGB image to form an RGBD representation, which is subsequently encoded by a pre-trained visual encoder ε_V:

$$VG_k = \varepsilon_V(DE(I_k) \oplus I_k), \tag{3}$$

where $DE(\cdot)$ denotes depth estimation and $\oplus$ indicates channel-wise concatenation. The final visual–geometric feature VG is obtained by averaging the features extracted from all six views. In parallel, semantic–geometric consistency is captured by encoding the corresponding textual prompt using the pre-trained text encoder ε_T:

$$SG = \varepsilon_T(P), \tag{4}$$

where, P and SG denote the prompt used to generate the 3DGC and semantic-geometric features.

4.3 Geometric Statistics

Geometric structure plays a critical role in 3DGC quality assessment Zhang et al. (2023e). To characterize geometric features in 3DGCs, we analyze the statistical distributions of Gaussian curvature and dihedral angles. The Gaussian curvature at vertex v_m is defined as:

$$G_m = \frac{(2\pi - \sum_n \theta_{mn})}{A_m}, \tag{5}$$

where, θ_{mn} denotes the angles between adjacent edges at v_m and A_m is the area of the corresponding Voronoi cell. The dihedral angle θ_i at edge e_i is computed as:

$$\cos\theta_i = \frac{n_{i1} \cdot n_{i2}}{\|n_{i1}\| \, \|n_{i2}\|}, \tag{6}$$

where, n_{i1} and n_{i2} are the normal vectors of the two faces adjacent to e_i. Following the computation of global geometric statistics for each 3DGC, we estimate the distribution parameters using three widely adopted models: the Generalised Gaussian Distribution (GGD), the Asymmetric Generalised Gaussian Distribution (AGGD), and the Gamma Distribution. Resulting parameters are aggregated into a geometric feature vector GF, which characterizes the intrinsic geometric properties.

4.4 Persistent Homotopy Analysis

Beyond local geometric analysis, topological modeling provides a complementary macro-level perspective for assessing the structural integrity of 3DGCs, which remains largely underexplored in existing quality assessment methods. Within Hyper-VG, we introduce persistent homotopy Zomorodian & Carlsson (2004) to characterize the multiscale topological properties of 3DGCs. Specifically, each 3DGC is first converted into a point cloud, from which an Alpha complex is constructed by progressively increasing the filtration parameter, yielding a nested sequence of topological spaces. Persistent homotopy is then computed over dimensions (H_0, H_1, H_2) to obtain persistence pairs (b, d), where b and d denote the birth and death of topological features, respectively, and the feature lifetime is defined as $L = d - b$. To encode topological information into discriminative

Table 2: Performance of different methods on databases. ♦ and ♥ denote methods based on manual features and deep learning (DL). All DL-based methods undergo training on the databases while maintaining identical experimental settings. Best in **RED**, second in **BLUE**.

Type	Method	3DGCQA				3DGCQA-LS				3DGCQA + 3DGCQA-LS			
		SRCC↑	PLCC↑	KRCC↑	RMSE↓	SRCC↑	PLCC↑	KRCC↑	RMSE↓	SRCC↑	PLCC↑	KRCC↑	RMSE↓
IQA	♦ BRISQUE Mittal et al. (2012)	0.2091	0.3347	0.1444	0.7414	0.0506	0.1790	0.0339	1.5023	0.1011	0.1834	0.0530	1.4665
	♦ CPBD Narvekar & Karam (2011)	0.2099	0.4797	0.1335	0.7217	0.0194	0.1263	0.0140	1.5147	0.0383	0.1504	0.0715	1.4671
	♦ IL-NIQE Zhang et al. (2015)	0.1481	0.1573	0.1131	0.6600	0.0716	0.0329	0.0481	1.5261	0.0769	0.04463	0.0520	1.4473
	♦ NFERM Gu et al. (2014)	0.2797	0.4062	0.1999	0.7222	0.3621	0.3854	0.2471	1.3177	0.3717	0.3986	0.2682	1.2971
	♦ NFSDM Gu et al. (2013)	0.3189	0.4468	0.2235	0.6935	0.3391	0.3703	0.2301	1.3185	0.3788	0.3748	0.2972	1.2810
	♦ NIQE Mittal et al. (2013)	0.2079	0.2594	0.1413	0.8050	0.1639	0.1864	0.1122	1.5002	0.1654	0.1879	0.1395	1.4549
	♥ TOPIQ Chen et al. (2024)	0.3598	0.4293	0.2601	0.7162	0.7425	0.7477	0.5545	1.6520	0.7295	0.7269	0.5336	1.5860
	♥ CLIP-IQA Wang et al. (2023b)	0.1633	0.2428	0.1081	0.4160	0.4722	0.2712	0.3317	1.3872	0.4654	0.2869	0.3284	1.2893
	♥ HyperIQA Su et al. (2020)	0.2027	0.3038	0.1387	0.4046	0.7482	0.7470	0.5587	1.6257	0.7105	0.7284	0.5335	1.5424
	♥ Q-Align Wu et al.	0.0746	0.0764	0.0498	0.8311	0.5653	0.1423	0.4064	1.4692	0.5586	0.1638	0.3845	1.4391
	♥ DBCNN Zhang et al. (2020)	0.5381	0.5147	0.3946	0.4700	0.7124	0.7317	0.5241	1.6004	0.7042	0.7168	0.5316	1.5376
	♥ StairIQA Sun et al. (2021)	0.3813	0.4566	0.2653	0.5802	0.6854	0.7022	0.5095	1.2124	0.6754	0.6876	0.4949	1.1745
VQA	♦ VIIDEO Mittal et al. (2015)	0.0765	0.2617	0.0558	0.7731	0.0229	0.1317	0.0164	1.4579	0.0630	0.1444	0.0431	1.3460
	♦ VIDEVAL Tu et al. (2021a)	0.2176	0.3506	0.1565	0.7416	0.3614	0.3873	0.2448	1.3229	0.2963	0.3309	0.2031	1.2628
	♦ TLVQM Korhonen (2019)	0.2665	0.3775	0.1867	0.7338	0.3172	0.3631	0.2173	1.3340	0.2449	0.3045	0.1686	1.2795
	♦ V-BLIINDS Saad et al. (2014)	0.0555	0.1961	0.0417	0.8396	0.1303	0.1440	0.0885	1.5149	0.1042	0.1311	0.0700	1.4858
	♥ RAPIQUE Tu et al. (2021b)	0.2257	0.4074	0.1579	0.7275	0.1841	0.2407	0.1263	1.4193	0.1466	0.1829	0.0999	1.3363
	♥ SimpVQA Sun et al. (2022)	0.4689	0.4994	0.3249	0.7090	0.6122	0.6117	0.4356	1.2063	0.5957	0.5956	0.4222	1.1987
	♥ VSFA Li et al. (2019)	0.5677	0.5984	0.4415	0.5984	0.6605	0.6575	0.4770	1.1697	0.6676	0.6649	0.4857	1.1285
	♥ MDTVSFA Li et al. (2021)	0.5631	0.5912	0.4381	0.6044	0.6689	0.6675	0.4843	1.1602	0.6808	0.6758	0.4889	1.1212
	♥ DOVER Wu et al. (2023)	0.5944	0.6025	0.4427	0.5923	0.6901	0.7014	0.5027	1.1125	0.7139	0.7045	0.4728	1.1081
	♥ FAST-VQA Wu et al. (2022)	0.4787	0.5080	0.3505	0.8095	0.6239	0.6158	0.4442	1.3373	0.6126	0.6087	0.4368	1.3202
	♥ BVQA Li et al. (2022)	0.4251	0.4346	0.3085	2.3792	0.4806	0.3210	0.3345	2.5810	0.5449	0.4846	0.3823	2.6086
3DQA	♦ 3D-NSS Zhang et al. (2022)	0.2865	0.3138	0.1809	0.7822	0.2273	0.2068	0.1250	1.2861	0.2305	0.2321	0.1258	1.2559
	♥ MM-PCQA Zhang et al. (2025c)	0.6328	0.6507	0.4531	0.4963	0.6577	0.6493	0.4962	1.1826	0.6521	0.6485	0.4981	1.1414
	♥ ViT-MQA Zhou et al. (2023b)	0.3517	0.3724	0.2609	0.8780	0.4305	0.4264	0.3429	1.2432	0.4352	0.4262	0.3398	1.2219
	♥ Dual-PCQA Zhang et al. (2023f)	0.6583	0.6578	0.4718	0.4549	0.6685	0.6674	0.5134	1.1722	0.6769	0.6668	0.5190	1.1262
	♥ CT-PCQA Zhou et al. (2025c)	0.6577	0.6498	0.4762	0.4635	0.6726	0.6635	0.5078	1.1764	0.6771	0.6627	0.5085	1.1333
	♥ HyperScore Zhang et al. (2025a)	**0.7484**	**0.7605**	0.5530	**0.4029**	**0.7781**	**0.7794**	**0.5844**	**1.0896**	**0.7625**	**0.7680**	**0.5721**	**1.0955**
	♥ T23DAQA Fu et al. (2025)	0.7279	0.7523	**0.5645**	0.4264	0.7398	0.7435	0.5216	1.1443	0.7343	0.7489	0.5308	1.1242
	♥ CAP Zhou et al. (2025b)	0.7098	0.7566	0.5386	0.4245	0.7298	0.7331	0.5582	1.1494	0.7368	0.7341	0.5560	1.1069
	♥ **Hyper-VG (Ours)**	**0.7656**	**0.7881**	**0.5872**	**0.3839**	**0.8071**	**0.8089**	**0.6122**	**1.0472**	**0.7880**	**0.7794**	**0.6035**	**1.0877**

feature representations, we extract 3 descriptors. First, persistence images are constructed by projecting persistence pairs onto the birth–death plane and smoothing them with a Gaussian kernel. Given the set of persistence pairs $D = (b_i, d_i)$ and a weighting function $w(b, d) = d - b$, the persistence image P_I at location (x, y) is computed as:

$$P_I(x, y) = \sum_{(b,d)\in D} w(b, d) \cdot \phi_{(b,d)}(x, y), \tag{7}$$

where $\phi_{(b,d)}$ denotes a Gaussian kernel centered at (b, d) with bandwidth. Second, Betti curves are computed by sampling the Betti numbers $\beta_k(\alpha)$ as functions of the filtration parameter α, capturing the evolution of topological features across scales Bochner & Yano (1954). Finally, we extract statistical descriptors of persistence distributions, including the number of features, mean and standard deviation of lifetimes, maximum and total lifetimes, mean birth and death times, and persistence entropy. Collectively, these topological features TF characterize global structural properties of 3DGCs, such as connected components, loops, and voids, across multiple dimensions.

4.5 Feature Fusion and Regression

To comprehensively consider the visual quality, multimodal information, geometric structure, and topology of 3DGC, we perform the following fusion of the extracted features:

$$\mathcal{F} = GAP(VF) \oplus GAP(VG) \oplus GAP(SG) \oplus GF \oplus TF, \tag{8}$$

where $GAP(\cdot)$ denotes global average pooling operation, $\oplus$ represents feature concatenation. The fused feature $\mathcal{F}$ is then subsequently fed into two Fully Connected (FC) layers to regress the predicted quality score $\hat{q}$. Mean Squared Error (MSE) is employed as the loss to enable continuous optimization.

5 Experiments

5.1 Experiment Details & Criteria

To validate the effectiveness of Hyper-VG, we conduct extensive performance comparisons against 31 representative objective quality assessment methods. Specifically, the competing approaches include 12 image quality assessment (IQA) methods, 11 video quality assessment (VQA) methods,

Table 3: Ablation study results, where '*w/o*' stands for 'without'. Best in **RED**, second in **BLUE**.

Dimension	3DGCQA				3DGCQA-LS				3DGCQA + 3DGCQA-LS			
	SRCC↑	PLCC↑	KRCC↑	RMSE↓	SRCC↑	PLCC↑	KRCC↑	RMSE↓	SRCC↑	PLCC↑	KRCC↑	RMSE↓
w/o VF	0.4874	0.4990	0.3052	0.7235	0.4625	0.4437	0.3152	1.3552	0.4994	0.4702	0.3164	1.3298
w/o (VG, SG)	0.6752	0.6899	0.4907	0.4431	0.7439	0.7285	0.5630	1.1774	0.7398	0.7268	0.5566	1.1342
w/o GF	**0.7362**	**0.7421**	**0.5588**	**0.4206**	**0.7843**	**0.7802**	**0.5997**	**1.0956**	**0.7714**	**0.7679**	**0.5725**	**1.0893**
w/o TF	0.7045	0.7314	0.5479	0.4298	0.7781	0.7696	0.5762	1.1240	0.7668	0.7578	0.5640	1.0942
Hyper-VG (Ours)	**0.7656**	**0.7881**	**0.5872**	**0.3839**	**0.8071**	**0.8089**	**0.6122**	**1.0472**	**0.7880**	**0.7794**	**0.6035**	**1.0877**

and 8 3D quality assessment (3DQA) methods. In terms of datasets, evaluations are conducted on both 3DGCQA dataset Zhou et al. (2025d), which contains 313 3DGCs, and the 3DGCQA-LS dataset. In addition, we perform joint evaluation by combining the two datasets to further examine the robustness and generalizability of each method. A five-fold cross-validation protocol is adopted, with no content overlap between folds, and the average performance across the five test splits is reported as the final evaluation score. Four quality assessment metrics are employed to quantitatively measure the performance of each method: Spearman's Rank-Order Correlation Coefficient (SRCC), Pearson's Linear Correlation Coefficient (PLCC), Kendall's Rank-Order Correlation Coefficient (KRCC), and Root Mean Square Error (RMSE). Among them, SRCC and KRCC primarily measure the monotonicity of prediction quality, while PLCC assesses the linearity and consistency of the algorithm's prediction quality. RMSE is employed to evaluate the accuracy of the prediction.

5.2 Performance Analysis

The evaluation results of all compared methods are summarized in Table 2. Several observations can be drawn: 1) The proposed Hyper-VG consistently outperforms the second-best method by more than 1.5% in terms of SRCC across multiple datasets, demonstrating its superior performance and validating its effectiveness; 2) Several IQA and VQA approaches achieve competitive results, highlighting the rationality of projection-based visual feature extraction. Notably, CAP Zhou et al. (2025b), which jointly models visual features and geometric structure, attains improved performance, further confirming the critical role of geometric characteristics in quality perception; 3) Hyper-VG maintains SRCC values exceeding 0.76 across all dataset configurations, indicating strong robustness and generalization capability across datasets of varying scales.

5.3 Ablation Experiments

To validate the design rationale, we conduct ablation studies on each feature component, with the results reported in Table 3. Several conclusions can be drawn: 1) All feature components contribute positively to the overall performance, confirming the effectiveness of Hyper-VG; 2) Visual features exhibit the most significant impact on 3DGC quality assessment, further validating the effectiveness of cube projection and highlighting the heightened sensitivity of the human visual system to visual plausibility; 3) As the dataset scale increases, the relative contributions of geometric and topological features gradually diminish, likely because Hyper-VG increasingly leverages richer and more semantically informative mid- to high-level representations learned from larger datasets.

6 Conclusion

Text-to-3D methods enable users to express design concepts through textual prompts and automatically generate corresponding 3D generated contents (3DGCs), and are therefore referred to as "Dreamers." Despite their promise, the absence of human supervision and the limitations of current generative technologies result in pronounced quality deficiencies in 3DGCs, substantially degrading user experience. To enable accurate and reliable assessment of 3DGC quality, we conduct a comprehensive study combining both subjective and objective evaluation. Specifically, we select 11 Dreamers to generate 3DGCs using 484 carefully designed prompts, resulting in a total of 4,691 3DGCs that form the large-scale 3DGCQA-LS dataset. Based on extensive subjective experiments involving participants, we identify and define 9 prevalent categories of distortions, thereby systematically revealing quality disparities across different Dreamers. We further propose Hyper-VG, a novel objective quality assessment framework that leverages world models and persistent homotopy to evaluate 3DGC quality from complementary perspectives. Extensive experimental results demonstrate that Hyper-VG produces quality predictions that are more closely aligned with human perceptual judgments, providing effective guidance for the optimization and development of Dreamers.

AI USE STATEMENT

Large Language Models (LLMs) are utilized to assist in polishing the language, improving readability, and enhancing the overall clarity, specifically employed for sentence rephrasing, grammar refinement, and ensuring coherence across sections; it is important to clarify that the LLM is not involved in any aspect of the research ideation, experimental design, or data analysis, as all scientific concepts, methodologies, and conclusions are developed and validated solely by the authors, who take full responsibility for the entire content of this manuscript, including any text that is generated or refined with the assistance of LLMs, and we confirm that the use of LLMs complies with ethical guidelines.

REFERENCES

Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*, 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Blender. `https://www.blender.org/`, 2024.

Salomon Bochner and Kentaro Yano. *Curvature and Betti numbers*. Number 32. Princeton University Press, 1954.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 2024.

Yang Chen, Yingwei Pan, Yehao Li, Ting Yao, and Tao Mei. Control3d: Towards controllable text-to-3d generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1148–1156, 2023.

Paolo Cignoni, Marco Callieri, Massimiliano Corsini, Matteo Dellepiane, Fabio Ganovelli, Guido Ranzuglia, et al. Meshlab: an open-source mesh processing tool. In *Eurographics Italian chapter conference*, volume 2008, pp. 129–136. Salerno, 2008.

Stephanie Circelli. Immersive spaces: The art of entertainment architecture in shaping perception, experience, and storytelling. Master's thesis, University of Cincinnati, 2025.

Donglin Di, Jiahui Yang, Chaofan Luo, Zhou Xue, Wei Chen, Xun Yang, and Yue Gao. Hyper-3dg: Text-to-3d gaussian generation via hypergraph. *International Journal of Computer Vision*, 133(5): 2886–2909, 2025.

Juan Du, Chengyu Tao, Xuanming Cao, and Fugee Tsung. 3d vision-based anomaly detection in manufacturing: A survey. *Frontiers of Engineering Management*, pp. 1–18, 2025.

Kang Fu, Huiyu Duan, Zicheng Zhang, Xiaohong Liu, Xiongkuo Min, Jia Wang, and Guangtao Zhai. Multi-dimensional quality assessment for text-to-3d assets: Dataset and model. *arXiv preprint arXiv:2502.16915*, 2025.

Ke Gu, Guangtao Zhai, Xiaokang Yang, et al. No-reference image quality assessment metric by combining free energy theory and structural degradation model. In *IEEE ICME*, pp. 1–6, 2013.

Ke Gu, Guangtao Zhai, Xiaokang Yang, et al. Using free energy principle for blind image quality assessment. *IEEE TMM*, 17(1):50–63, 2014.

Xianglong He, Junyi Chen, Sida Peng, Di Huang, Yangguang Li, Xiaoshui Huang, Chun Yuan, Wanli Ouyang, and Tong He. Gvgen: Text-to-3d generation with volumetric representation. In *European Conference on Computer Vision*, pp. 463–479. Springer, 2024.

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

Fangzhou Hong, Jiaxiang Tang, Ziang Cao, Min Shi, Tong Wu, Zhaoxi Chen, Shuai Yang, Tengfei Wang, Liang Pan, Dahua Lin, et al. 3dtopia: Large text-to-3d generation model with hybrid diffusion priors. *arXiv preprint arXiv:2403.02234*, 2024.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *ACM Transactions on Graphics (TOG)*, 44(6): 1–15, 2025.

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

ITU. Methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Union*, 2002.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Md Ashik Khan and Arafat Alam Jion. Fixed-threshold evaluation of a hybrid cnn-vit for ai-generated image detection across photos and art. *arXiv preprint arXiv:2512.21512*, 2025.

Seungwook Kim, Kejie Li, Xueqing Deng, Yichun Shi, Minsu Cho, and Peng Wang. Enhancing 3d fidelity of text-to-3d using cross-view correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10649–10658, 2024.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. Large-scale text-to-image generation models for visual artists' creative works. In *Proceedings of the 28th international conference on intelligent user interfaces*, pp. 919–933, 2023.

Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28(12):5923–5938, 2019.

Zeqiang Lai, Yunfei Zhao, Haolin Liu, Zibo Zhao, Qingxiang Lin, Huiwen Shi, Xianghui Yang, Mingxin Yang, Shuhui Yang, Yifei Feng, et al. Hunyuan3d 2.5: Towards high-fidelity 3d assets generation with ultimate details. *arXiv preprint arXiv:2506.16504*, 2025.

Karim Lekadir, Richard Osuala, Catherine Gallin, Noussair Lazrak, Kaisar Kushibar, Gianna Tsakou, Susanna Aussó, Leonor Cerdá Alberich, Kostas Marias, Manolis Tsiknakis, et al. Future-ai: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. *arXiv preprint arXiv:2109.09658*, 2021.

Bowen Li, Weixia Zhang, Meng Tian, Guangtao Zhai, and Xianpei Wang. Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):5944–5958, 2022.

Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *ACM International Conference on Multimedia*, pp. 2351–2359, 2019.

Dingquan Li, Tingting Jiang, and Ming Jiang. Unified quality assessment of in-the-wild videos with mixed datasets training. *International Journal of Computer Vision*, 129(4):1238–1257, 2021.

Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023.

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

Fangfu Liu, Diankun Wu, Yi Wei, Yongming Rao, and Yueqi Duan. Sherpa3d: Boosting high-fidelity text-to-3d generation via coarse 3d prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20763–20774, 2024a.

Ying-Tian Liu, Yuan-Chen Guo, Guan Luo, Heyi Sun, Wei Yin, and Song-Hai Zhang. Pi3d: Efficient text-to-3d generation with pseudo-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19915–19924, 2024b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Artem Lukoianov, Haitz Sáez de Ocáriz Borde, Kristjan Greenewald, Vitor Guizilini, Timur Bagautdinov, Vincent Sitzmann, and Justin M Solomon. Score distillation via reparametrized ddim. *Advances in Neural Information Processing Systems*, 37:26011–26044, 2024.

Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12663–12673, 2023.

A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "completely blind" image quality analyzer. *IEEE SPL*, 20(3):209–212, 2013.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE TIP*, 21(12):4695–4708, 2012.

Anish Mittal, Michele A Saad, and Alan C Bovik. A completely blind video integrity oracle. *IEEE Transactions on Image Processing*, 25(1):289–300, 2015.

Niranjan D Narvekar and Lina J Karam. A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). *IEEE TIP*, 20(9):2678–2683, 2011.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, June 2022.

Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind prediction of natural video quality. *IEEE Transactions on Image Processing*, 23(3):1352–1365, 2014.

Aditya Sanghi, Pradeep Kumar Jayaraman, Arianna Rampini, Joseph Lambourne, Hooman Shayani, Evan Atherton, and Saeid Asgari Taghanaki. Sketch-a-shape: Zero-shot sketch-to-3d shape generation. *arXiv preprint arXiv:2307.03869*, 2023.

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

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3667–3676, 2020.

Wei Sun, Xiongkuo Min, Guangtao Zhai, et al. Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training. *arXiv preprint arXiv:2105.14550*, 2021.

Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. A deep learning based no-reference quality assessment model for ugc videos. In *ACM International Conference on Multimedia*, 2022.

Tripo. `https://www.tripo3d.ai/`, 2025.

Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3d meshes from text prompts. In *2024 3DV*, pp. 1554–1563. IEEE, 2024.

Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. Ugc-vqa: Benchmarking blind video quality assessment for user generated content. *IEEE Transactions on Image Processing*, 30:4449–4464, 2021a.

Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. Rapique: Rapid and accurate video quality prediction of user generated content. *IEEE Open Journal of Signal Processing*, 2:425–440, 2021b.

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12619–12629, 2023a.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023b.

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. *Advances in neural information processing systems*, 36:8406–8441, 2023c.

Jiacheng Wei, Hao Wang, Jiashi Feng, Guosheng Lin, and Kim-Hui Yap. Taps3d: Text-guided 3d textured shape generation from pseudo supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16805–16815, 2023.

Haoning Wu, Zicheng Zhang, Weixia Zhang, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In *ICML*.

Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *European Conference on Computer Vision*, pp. 538–554. Springer, 2022.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 20144–20154, 2023.

Jinbo Wu, Xiaobo Gao, Xing Liu, Zhengyang Shen, Chen Zhao, Haocheng Feng, Jingtuo Liu, and Errui Ding. Hd-fusion: Detailed text-to-3d generation leveraging multiple noise estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3202–3211, 2024.

Xudong Xu, Zhaoyang Lyu, Xingang Pan, and Bo Dai. Matlaber: Material-aware text-to-3d via latent brdf auto-encoder. *arXiv preprint arXiv:2308.09278*, 2023.

Han Yi, Zhedong Zheng, Xiangyu Xu, and Tat-seng Chua. Progressive text-to-3d generation for automatic 3d prototyping. *arXiv preprint arXiv:2309.14600*, 2023.

L. Zhang, L. Zhang, and A. C. Bovik. A feature-enriched completely blind image quality evaluator. *IEEE TIP*, 24(8):2579–2591, 2015.

Weixia Zhang, Kede Ma, Jia Yan, et al. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE TCSVT*, 30(1):36–47, 2020.

Yujie Zhang, Bingyang Cui, Qi Yang, Zhu Li, and Yiling Xu. Benchmarking and learning multi-dimensional quality evaluator for text-to-3d generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18563–18574, 2025a.

Zicheng Zhang, Wei Sun, Xiongkuo Min, Tao Wang, Wei Lu, and Guangtao Zhai. No-reference quality assessment for 3d colored point cloud and mesh models. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7618–7631, 2022.

Zicheng Zhang, Wei Sun, Yingjie Zhou, Wei Lu, Yucheng Zhu, Xiongkuo Min, and Guangtao Zhai. Eep-3dqa: Efficient and effective projection-based 3d model quality assessment. In *2023 IEEE international conference on Multimedia and expo (ICME)*, pp. 2483–2488. IEEE, 2023a.

Zicheng Zhang, Wei Sun, Yucheng Zhu, Xiongkuo Min, Wei Wu, Ying Chen, and Guangtao Zhai. Evaluating point cloud from moving camera videos: A no-reference metric. *IEEE Transactions on Multimedia*, 2023b.

Zicheng Zhang, Yingjie Zhou, Wei Sun, Wei Lu, Xiongkuo Min, Yu Wang, and Guangtao Zhai. Ddh-qa: A dynamic digital humans quality assessment database. In *ICME*, pp. 2519–2524. IEEE, 2023c.

Zicheng Zhang, Yingjie Zhou, Wei Sun, Xiongkuo Min, Yuzhe Wu, and Guangtao Zhai. Perceptual quality assessment for digital human heads. In *ICASSP*, pp. 1–5. IEEE, 2023d.

Zicheng Zhang, Yingjie Zhou, Wei Sun, Xiongkuo Min, and Guangtao Zhai. Geometry-aware video quality assessment for dynamic digital human. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 1365–1369. IEEE, 2023e.

Zicheng Zhang, Yingjie Zhou, Wei Sun, et al. Simple baselines for projection-based full-reference and no-reference point cloud quality assessment. *arXiv preprint arXiv:2310.17147*, 2023f.

Zicheng Zhang, Wei Sun, Haoning Wu, Yingjie Zhou, Chunyi Li, Zijian Chen, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Gms-3dqa: Projection-based grid mini-patch sampling for 3d model quality assessment. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(6):1–19, 2024a.

Zicheng Zhang, Haoning Wu, Yingjie Zhou, Chunyi Li, Wei Sun, Chaofeng Chen, Xiongkuo Min, Xiaohong Liu, Weisi Lin, and Guangtao Zhai. Lmm-pcqa: Assisting point cloud quality assessment with lmm. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7783–7792, 2024b.

Zicheng Zhang, Yingjie Zhou, Chunyi Li, Kang Fu, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. A reduced-reference quality assessment metric for textured mesh digital humans. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2965–2969. IEEE, 2024c.

Zicheng Zhang, Wei Sun, Yingjie Zhou, Haoning Wu, Chunyi Li, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. Advancing zero-shot digital human quality assessment through text-prompted evaluation. *IEEE TIP*, 2025b.

Zicheng Zhang, Yingjie Zhou, Chunyi Li, Wei Sun, Xiongkuo Min, Xiaohong Liu, and Guangtao Zhai. Mm-pcqa+: Advancing multi-modal learning for point cloud quality assessment. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(4):1–22, 2025c.

Zicheng Zhang, Yingjie Zhou, Chunyi Li, Baixuan Zhao, Xiaohong Liu, and Guangtao Zhai. Quality assessment in the era of large models: A survey. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(7):1–31, 2025d.

Yingjie Zhou, Yaodong Chen, Kaiyue Bi, Lian Xiong, and Hui Liu. An implementation of multimodal fusion system for intelligent digital human generation. *arXiv preprint arXiv:2310.20251*, 2023a.

Yingjie Zhou, Zicheng Zhang, Wei Sun, et al. A no-reference quality assessment method for digital human head. In *ICIP*, pp. 36–40. IEEE, 2023b.

Yingjie Zhou, Zicheng Zhang, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Subjective and objective quality-of-experience assessment for 3d talking heads. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6033–6042, 2024.

Yingjie Zhou, Jiezhang Cao, Farong Wen, Li Xu, Yanwei Jiang, Jun Jia, Ronghui Li, Xiaohong Liu, Yu Zhou, Xiongkuo Min, et al. Who is a better player: Llm against llm. *arXiv preprint arXiv:2508.04720*, 2025a.

Yingjie Zhou, Farong Wen, Zicheng Zhang, Yanwei Jiang, Jun Jia, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Cap: An advanced no-reference quality assessment method for ai-generated 3d meshes. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2025b. doi: 10.1109/ICME59968.2025.11210150.

Yingjie Zhou, Zicheng Zhang, Wei Sun, Xiongkuo Min, and Guangtao Zhai. Ct-pcqa: A convolutional neural network and transformer combined method for point cloud quality assessment. *Signal Processing: Image Communication*, pp. 117371, 2025c.

Yingjie Zhou, Zicheng Zhang, Farong Wen, Jun Jia, Yanwei Jiang, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. 3dgcqa: A quality assessment database for 3d ai-generated contents. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025d.

Junzhe Zhu, Peiye Zhuang, and Sanmi Koyejo. Hifa: High-fidelity text-to-3d generation with advanced diffusion guidance. *arXiv preprint arXiv:2305.18766*, 2023.

Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. In *Proceedings of the twentieth annual symposium on Computational geometry*, pp. 347–356, 2004.

A Details of Subjective Quality Assessment

The primary objective of conducting subjective quality assessment experiments on 3D-generated contents (3DGCs) is to capture authentic user feedback, including perceptual ratings that reflect overall visual quality as well as the presence of specific distortion types. To enhance perceptual immersion and allow flexible viewpoint exploration, all 3DGC samples are rendered and presented using Blender Blender (2024), as illustrated in Fig. 6. Participants are permitted to freely manipulate the viewing perspective via standard mouse interactions, including rotation, panning, and zooming, before providing their assessments. Overall quality is rated using the Absolute Category Rating (ACR) protocol, while perceived distortion types are recorded through a multiple-choice selection interface.

Based on subjective annotations, we identify nine prevalent quality distortions in 3DGCs, as illustrated in Fig. 3. These distortions can be systematically categorized into four dimensions: geometric structure (GS), spatial depth (SD), semantic logic (SL), and appearance texture (AT). To further clarify the definitions and boundaries of each distortion type, the corresponding identification criteria are summarized in Table 4. An analysis of the distortion statistics in Fig. 3 reveals that the constructed 3DGCQA-LS dataset comprises 4,691 distorted instances, with a total of 7,900 distortion annotations. This observation indicates that individual 3DGC samples may exhibit multiple co-occurring quality issues, thereby highlighting the necessity and importance of comprehensive quality assessment, as addressed in this work.

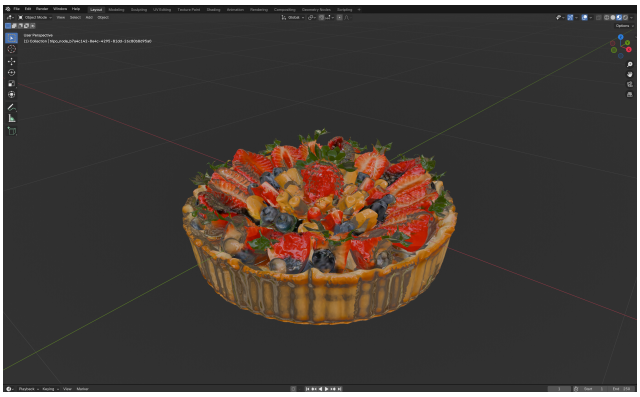


Figure 6: Blender-based subjective assessment interface with free-viewpoint control. Allowing unrestricted viewpoint exploration facilitates more comprehensive observation and improves the reliability of subjective evaluation results.

Table 4: Types of distortion observed in 3DGCs and their definitions.

Type	Label	Distortion	Description
Geometric Structure	GS-1	Geometric Loss	The 3DGC exhibits structural deficiencies and incompleteness.
	GS-2	Geometric Redundancy	The 3DGC contains redundant, extraneous structures.
	GS-3	Geometric Deformation	The 3DGC exhibits distorted, stretched, and irregular shapes.
	GS-4	Rough Structure	Noise, mottling, and jagged edges appear on the surface of 3DGC.
Spatial Depth	SD	Depth Error	Errors are occurred in spatial relationships, occlusion, and perspective of the 3DGC.
Semantic Logic	SL-1	Semantic Mismatch	The 3DGC fails to represent the content required by the prompt.
	SL-2	Multifaceted Repetition	Different components of the 3DGC exhibit unreasonable similarities.
Appearance Texture	AT-1	Anomalous Texture	The texture exhibits issues such as color spotting, bleeding, and adhesion.
	AT-2	Detail Loss	The 3DGC is overly simplistic and lacks intricate detail.

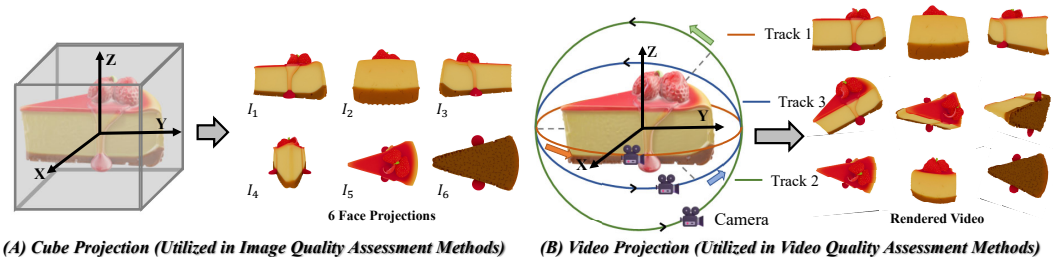


Figure 7: Visualization of different projection strategies.

B Implementation Details of Hyper-VG

Although Sec. 4 provides an overview of the architectural design of the proposed Hyper-VG framework, a more detailed discussion of its implementation is warranted. First, illumination conditions during the 3D content creation process have a substantial impact on the visual perception of 3DGCs. As illustrated in Fig. 5, point light sources are typically used to simulate localized artificial lighting, whereas natural light models the characteristics of sunlight, providing a global, uniform, and non-attenuating illumination. Compared with the variability introduced by multiple point light sources, the adoption of natural lighting in Hyper-VG standardizes the illumination environment across 3DGC samples. This design choice ensures consistent default lighting conditions and effectively mitigates issues of underexposure and overexposure caused by heterogeneous lighting setups.

For network training, we employ the Swin Transformer Tiny Liu et al. (2021) as the backbone architecture, with a batch size of 32. The learning rate is set to 1×10^{-5}, and optimization is performed using the Adam optimizer Kingma (2014). All experiments are conducted on a server equipped with eight NVIDIA A100 GPUs with 80 GB memory each.

C Projection Schemes & Performance Experiment Details

As all 3DGCs within the 3DGCQA-LS dataset we established are in mesh data format, we can first define each 3DGC as follows:

$$Mesh = \{(v, n, e)|v \in V, n \in N, e \in E\}, \tag{9}$$

where V, N, and E represent the sets of vertices, normal vectors, and edges, respectively. Subsequently, we can calculate the center point position of the 3DGC using the following formula:

$$C = \frac{1}{N_v} \sum_{i=1}^{N_v} v_i, \tag{10}$$

where v_i represents the position of the i-th vertex, and N_v denotes the total number of vertices in the mesh. Using C, a 3D coordinate system is established to achieve different projection schemes, with the default direction aligned to the x-axis.

As illustrated in Fig. 7, this work considers two primary categories of projection strategies for 3DGCs: cube projection and video projection. The cube projection strategy deploys virtual cameras along the positive and negative directions of the x, y, and z axes in a Cartesian coordinate system, thereby capturing six canonical views of the 3DGC, namely the front, back, left, right, top, and bottom projections. In contrast, video projection samples a denser and more diverse set of viewpoints. Specifically, we adopt the three rendering trajectories proposed by Zhang $et\ al.$ Zhang et al. (2023b), performing equidistant circular rendering around the 3DGC at 45° intervals to generate a continuous video sequence. Compared with video projection, cube projection requires only six rendered views and thus offers substantially higher computational efficiency. Accordingly, cube projection is employed in the proposed objective evaluation framework.

In Sec. 5, we compare Hyper-VG with representative image quality assessment (IQA), video quality assessment (VQA), and 3D quality assessment (3DQA) methods. Three implementation details merit further clarification: 1) For IQA methods, cube projection is first applied to the 3DGC, and the final visual quality score is obtained by averaging the quality scores of the six rendered views; 2) For VQA methods, the 3DGC is rendered into a video sequence using the video projection strategy, and the resulting video quality score is used as the visual quality metric; 3) For 3DQA methods that are applicable only to point cloud data, the 3DGC is converted into a point cloud representation using MeshLab Cignoni et al. (2008) prior to quality evaluation.

D Persistent Homotopy Analysis

In conventional 3DQA, manually modeled or sensor-acquired 3D content generally preserves a high degree of macroscopic structural consistency with real-world objects. As a result, quality assessment methods that rely on local geometric features have been shown to be effective. However, as summarized in Table 4, contemporary 3DGCs still exhibit notable macroscopic geometric degradations, including geometric loss, structural redundancy, and global deformation. These observations indicate that geometric features alone are insufficient for reliably detecting such high-level structural anomalies. To address this limitation, we introduce persistent homotopy analysis Zomorodian & Carlsson (2004) to enable a more holistic topological characterization of 3DGC. Specifically, the 3DGC is first converted into a point cloud representation, from which an Alpha complex is constructed. By progressively increasing the parameter α, a sequence of nested topological spaces is generated, capturing the evolution of topological features across scales. For a point set P, the Alpha complex is defined as:

$$\mathcal{A}_\alpha(P) = \left\{ \sigma \subset P \mid \min_{x \in \mathbb{R}^3} \max_{p \in \sigma} \|x - p\| \leq \alpha \leq \min_{q \notin \sigma} \|x - q\| \right\}, \tag{11}$$

where $\min_{x \in \mathbb{R}^3} \max_{p \in \sigma} \|x - p\|$ is the radius of the smallest enclosing ball of simplex σ, $\min_{q \notin \sigma} \|x - q\|$ is the distance from the ball center to the nearest point not in σ. Building upon the Alpha complex, we compute persistent homology to obtain persistent pairs (b, d) across different dimensions (H_0, H_1, H_2), where b denotes the filtering parameter value at trait birth and d denotes the filtering parameter value at trait death. The trait's lifespan is $L = d - b$. To transform topological features into quantifiable feature vectors, we extract three types of feature representations. The first is the Persistence Image, obtained by mapping persistence onto the birth-death plane and smoothing with a

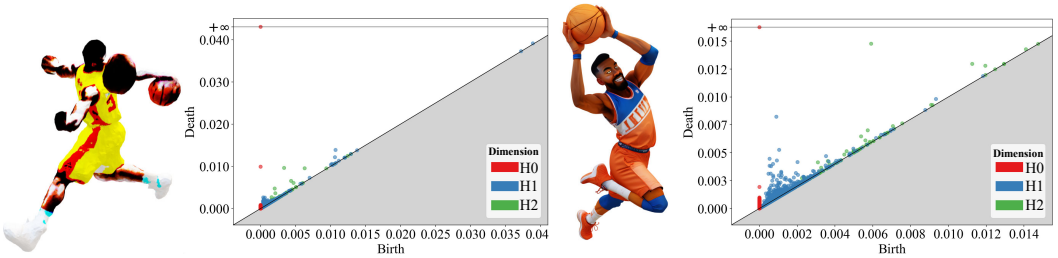


Figure 8: Visualization of persistent homotopy analysis results. Persistent pairs across different dimensions are marked with distinct colors. Both 3DGCs on the left and right originate from the same prompt, yet their topological analysis results exhibit significant differences.

Gaussian kernel. Let the set of persistence pairs be denoted as $D = \{(b_i, d_i)\}$, with the weighting function $w(b, d) = d - b$. For a Gaussian kernel ϕ_u centred at $u = (b, d)$ with bandwidth σ, the value of the persistence image at point (x, y) is given by:

$$I(x, y) = \sum_{(b,d)\in D} w(b, d) \cdot \phi_{(b,d)}(x, y). \tag{12}$$

Secondly, we compute the Betti curve characteristics and determine the function of the Betti number $\beta_k(\varepsilon)$ as a function of the filtering parameter ε at discrete sampling points, thereby capturing the evolution of topological features across different scales. Finally, we extract statistical features including the number of features, mean lifetime, lifetime standard deviation, maximum lifetime, total lifetime, mean birth time, mean death time, and sustained entropy E:

$$E = -\sum_{i=1}^{n} p_i \log p_i, \quad p_i = \frac{L_i}{\sum_j L_j}. \tag{13}$$

These topological features TF characterise the macroscopic structural properties of 3DGC from different dimensions, including connected components, loops, and voids, thereby offering a novel perspective for quality assessment.

To more intuitively illustrate the influence of persistent homotopy theory on the perceived quality of 3DGCs, we perform a qualitative visual analysis of 3DGCs produced by two different Dreamers using the same textual prompt, as shown in Fig. 8. As observed in Fig. 8, the 3DGC on the left suffers from pronounced repetitive distortions across multiple facets. In contrast to the high-quality 3DGC on the right, this results in a substantially different distribution of persistent pairs. The comparative topological analysis presented in Fig. 8 for both low- and high-quality samples demonstrates that persistent homotopy analysis is capable of effectively capturing macroscopic structural characteristics of 3DGCs, thereby contributing positively to comprehensive quality assessment.

E Comparison of Inference Efficiency

We conduct inference efficiency experiments to calculate the average inference efficiency of Hyper-VG. Specifically, we used the large-scale 3DGCQA+3DGCQA-LS dataset to perform a total of 5,004 3DGC inferences. We use the average inference time to measure the inference efficiency of each algorithm; the experimental results are shown in Table 5.

Table 5: Average inference time across various evaluation methods.

Methods	CT-PCQA	Dual-PCQA	MM-PCQA	CAP	Hyper-VG (Ours)
Runtim (Seconds/Item)	8.04	8.90	10.43	7.95	7.72

As can be seen from Table 5 , our proposed Hyper-VG achieves the highest efficiency among the 3D quality assessment algorithms. This superiority stems from the fact that our feature extraction module can run in parallel, ultimately enabling efficient inference.

F LIMITATIONS AND BROADER IMPACTS

This work still has limitations in terms of dataset construction and the baseline evaluation algorithms used. First, regarding dataset construction, although the 3DGCQA-LS dataset has been compiled to be as comprehensive as possible in covering existing Dreamers, some Dreamers have not been included. While Hyper-VG provides a comprehensive and effective evaluation framework for 3DGCs, its performance in more practical applications, such as reconstructed large-scale scenes and autonomous driving, still requires further validation. However, this work keeps pace with the latest technological advancements and provides a valuable evaluation benchmark for the future development and optimization of the Dreamers and even world models.