DIGITAL HERITAGE (2025)
S. Campana, D. Ferdani, H. Graf, G. Guidi, Z. Hegarty, S. Pescarin, and F. Remondino (Editors)

Enhancing Cultural Heritage with Generative AI: A Comparative Framework for the Evaluation of 3D Model Accuracy and Visual Fidelity

E. Balloni[1], M. Paolanti[2], J. Uggeri[1], P. Zingaretti [1], R. Pierdicca[3]

[1]Università Politecnica delle Marche, Department of Information Engineering (DII), Via Brecce Bianche, 12, 60131, Ancona, Italy
[2]University of Macerata, Department of Political Sciences, Communication and International Relations, Via Don Minzoni, 22A, 62100, Macerata, Italy
[3]Università Politecnica delle Marche, Department of Construction, Civil Engineering and Architecture (DICEA), Via Brecce Bianche, 12, 60131, Ancona, Italy

Abstract

The digitization of Cultural Heritage (CH) has become a vital tool for preservation and dissemination, with 3D reconstruction playing a key role in capturing intricate geometries and visual details of artifacts. While traditional methods like photogrammetry and laser scanning are effective, they often involve labor-intensive processes and struggle with complex material properties. Recent advancements in Generative AI (GenAI), particularly Large Reconstruction Models (LRMs) such as TRELLIS, offer promising alternatives for 3D generation. However, their application in CH remains underexplored. This paper introduces a novel comparative framework to evaluate the accuracy and visual fidelity of 3D GenAI models in the CH domain. Focusing on TRELLIS, the framework assesses single-view and multi-view 3D generation across five diverse CH scenes, employing both 2D (PSNR, SSIM, LPIPS) and 3D (Chamfer Distance, F-score, Accuracy) metrics. Results demonstrate superior performance for individual artifacts (e.g., Minareto, Greek Vase) compared to complex architectural scenes, with multi-view generation consistently outperforming single-view approaches. The study highlights the potential of GenAI for CH preservation while identifying challenges in large-scale reconstructions, paving the way for future hybrid methodologies and sparse-view optimizations.

CCS Concepts

• ***Computing methodologies*** → ***Artificial intelligence;*** *3D imaging;* ***Computer graphics; Image-based rendering;*** • ***Information systems*** → *Multimedia content creation;*

1. Introduction

The digitization of cultural heritage (CH) has emerged as a powerful paradigm for preserving, studying and disseminating our shared history and identity [HKP*22]. As many artifacts, monuments and archaeological sites are under threat from environmental degradation, conflict or the passage of time, 3D digitization provides a valuable way to document and protect their legacy. In this context, 3D reconstruction stands out for its ability to capture the intricate geometries and visual details of cultural artifacts. Traditional techniques, such as Photogrammetry and Terrestrial Laser Scanning, are key in this domain, enabling the creation of detailed 3D models from a series of images. However, these conventional methods often include labor-intensive processes, require significant expertise, and may struggle with accurately reproducing complex material properties or fine details.

In recent years, artificial intelligence (AI) has emerged as a transformative force in 3D reconstruction and generation. Techniques such as Neural Rendering, Neural Radiance Fields (NeRFs) and 3D Gaussian Splatting, have demonstrated remarkable capabilities in synthesising photorealistic 3D representations from 2D images. These methods show promise in the cultural heritage (CH) sector, offering new ways to digitise artifacts and sites [BGP*23; BCP*24; CFB*24]. Nonetheless, they also present some limitations. In particular, they often require extensive image datasets to achieve high-quality results and may produce outputs that lack the desired 3D fidelity when converted to a mesh representation for certain applications. Parallel to these developments, Generative AI (GenAI) has gained prominence in 3D object generation. Specifically, Large Reconstruction Models (LRMs) leverage vast 3D datasets to learn versatile representations of data [HZG*23], enabling the generation of diverse and high-quality 3D models from minimal and multimodal inputs [LGL*24; LLL*24]. Among the most recent state-of-the-art works, a notable example is TRELLIS [XLX*24], which introduces a unified Structured LATent (SLAT) representation to facilitate scalable and versatile 3D generation.

Despite the rapid progress of 3D generative AI (GenAI) models, their application in the field of CH remains significantly underexplored [Spe24]. Most existing GenAI approaches have been de-

DOI: 10.2312/dh.20253146

veloped and benchmarked using generic datasets of everyday objects, either synthetic or real-world. These datasets often fail to capture the complexity and uniqueness of CH artifacts. CH assets differ fundamentally in terms of their morphology, surface textures and material compositions. They often have highly irregular shapes, intricate ornamental features, non-repetitive patterns and signs of historical degradation or restoration. These characteristics present a significant challenge to existing GenAI pipelines, which are usually optimised for clean, well-structured and frequently occurring object classes. Furthermore, CH artifacts are not merely aesthetic or functional objects; they carry historical, cultural and symbolic meaning. Consequently, their digital representations must meet stricter requirements in terms of visual and structural fidelity.

To address these limitations, this work introduces a comparative evaluation framework designed specifically to assess the suitability and performance of 3D GenAI models in the context of CH. The framework is designed to meet the specific requirements of CH digitization. These include preserving high-resolution detail, reproducing accurate geometry and handling incomplete or minimal input data. Our focus is particularly on the TRELLIS Generative AI Large Reconstruction Model (LRM) [XLX*24], a recent state-of-the-art model that uses a Structured Latent (SLAT) representation to enable scalable, multimodal 3D generation. Although TRELLIS has been shown to perform well in standard benchmarks, its capacity to generalise to CH scenarios has yet to be systematically studied. We conducted our evaluation on five different CH scenes, encompassing a range of object types, geometrical complexities, and cultural significances. Each scene is reconstructed from two input configurations: a single-view image and a set of multiple views. The generated 3D outputs are assessed using a dual-perspective methodology that incorporates 2D (image-based) and 3D (geometry-based) evaluation criteria. These include visual fidelity, structural coherence and spatial accuracy, which are crucial aspects for downstream applications such as virtual exhibitions, digital restoration or educational visualisation. This study aims to assess current capabilities and identify potential gaps and future directions in applying GenAI to the CH field. By grounding the evaluation in the specific constraints and demands of CH preservation, our framework helps lay the groundwork for more reliable, interpretable and culturally sensitive AI-driven 3D reconstruction methodologies.

The main contributions of this work lie in the development and application of a novel evaluation framework tailored to the specific requirements of CH digitization using GenAI. Unlike general-purpose benchmarks, our framework addresses the unique challenges posed by cultural artifacts, including their irregular geometries, intricate surface details and material complexity. This study provides empirical insights into the capabilities and limitations of GenAI models such as TRELLIS in the context of CH. It offers guidance on integrating these models into heritage documentation and preservation workflows, and highlights key areas for future research and improvement.

2. Related Works

The growing interest in applying AI to the creation of 3D content has led to remarkable advances in various fields, including entertainment, robotics and virtual reality [DPN*22; MGF*23; RSL*24]. However, when it comes to CH, the requirements for 3D digitization are particularly stringent, demanding visual fidelity, structural accuracy, and historical authenticity. This section reviews relevant research across three main areas: (i) generative AI methods for 3D generation, with a focus on recent developments in diffusion-based approaches; (ii) AI-driven 3D reconstruction applied specifically to cultural heritage artifacts; and (iii) evaluation methodologies designed to assess the quality and fidelity of reconstructed 3D models in CH contexts. This review contextualises our work within the broader landscape of GenAI and CH digitization, emphasising the necessity of specialised evaluation frameworks to address the distinctive challenges posed by heritage applications.

Generative AI for 3D Generation

Recent advances in diffusion-based generative models have enabled high-quality 3D content creation from 2D inputs [LHH*24; WLW*24]. Notable approaches focus on single-image 3D reconstruction and text-to-3D generation. For example, One-2-3-45 converts a single image into a textured 3D mesh in about 45 seconds without per-object optimization [LZW*23a], and its successor One-2-3-45++ further improves fidelity by fine-tuning diffusion models for consistent multi-view synthesis [LZW*23b]. Magic123 introduces a two-stage pipeline that first fits a coarse NeRF and then extracts a differentiable high-resolution mesh, guided by both 2D and 3D diffusion priors [QWZ*24]. Similarly, Make-It-3D lifts a single image to a NeRF, then to a textured point-cloud, using a two-stage diffusion-guided optimization to achieve high-fidelity geometry and texture [TWZ*23b]. DreamGaussian replaces the NeRF with a generative 3D Gaussian splatting model, yielding textured meshes in only a few minutes, about an order of magnitude faster than traditional score distillation methods [TWZ*23a].
Other works apply multi-view diffusion for 3D generation. RenderDiffusion is a diffusion model that explicitly predicts an intermediate 3D scene representation at each denoising step, enforcing 3D consistency while training on 2D image supervision [AXF*23]. MVDream learns a multi-view diffusion prior from both 2D and 3D data, enabling generation of consistent view sets from a text prompt, which can then be converted to 3D via existing pipelines [SWY*23]. Cycle3D tightly couples a 2D diffusion generator and a feed-forward 3D reconstruction network in a single diffusion loop: the 2D model proposes new views which are immediately "corrected" by a 3D reconstructor, yielding more consistent geometry and textures [TZC*25]. IM-3D iteratively alternates between multi-view generation (using a video diffusion model) and 3D reconstruction via Gaussian splatting, dramatically reducing computation while producing high-quality outputs with fewer artifacts [MLR*24]. [XLX*24] introduces Structured LATent (SLAT) representations, which allow decoding to different output formats, such as NeRF, 3DGS, and meshes, outperforming previous methods. Collectively, these models push the state of the art in image-to-3D, showing significant gains in speed, geometric accuracy, and visual quality.

AI-Driven 3D Reconstruction in Cultural Heritage

Digitization of CH objects often requires reconstructing damaged or incomplete artifacts from limited images. Recently, AI methods

have begun to tackle these challenges. [JS24] introduced a conditional diffusion model designed to reconstruct 3D point clouds of heritage objects. Their approach demonstrated the model's capability to accurately reproduce geometries specific to cultural artifacts, despite challenges related to data diversity and outlier sensitivity. In the area of digital restoration, [Dan*24] combined Stable Diffusion for image inpainting with NeRFs to repair and visualize degraded ceramic artifacts. This method enabled the creation of realistic 3D surrogates from incomplete 2D images, highlighting the potential of AI in enhancing museum exhibits and public engagement. [Shi25] applied AI-assisted 3D modeling to reconstruct temple arts from historical photographs. By leveraging platforms like 3DGS and NeRF, the study achieved detailed reconstructions, emphasizing AI's role in the continuous preservation of CH through evolved documentation and interpretation processes. Additionally, [DSG*24] proposed a pipeline leveraging 3DGS for efficient 3D digitization and segmentation of cultural heritage objects using only RGB images. This approach facilitates the creation of digital replicas without the need for manual annotation, making it accessible for widespread deployment.

Evaluation of 3D Models in Cultural Heritage

The CH domain demands rigorous evaluation of reconstructed 3D models in terms of geometric accuracy and visual fidelity. [LTGR23] propose a benchmark for heritage reconstruction, using laser-scanned artifacts as ground truth. They measure reconstruction quality by the Chamfer Distance between meshes, among other metrics, and highlight how lighting, surface properties, and occlusions affect different algorithms. [NSRR20] provides an overview of the different metrics to evaluate 3D suface reconstruction in CH, analyzing various 3D mesh quality metrics (e.g. F-score, accuracy) and discuss their relevance for CH digitization. Together, these works emphasize specialized evaluation protocols for heritage objects, using both quantitative benchmarks and qualitative criteria to assess the fidelity, completeness, and usability of generated 3D reconstructions. Additionally, the development of benchmarks like CUBE (CUltural BEnchmark for Text-to-Image models) allow the evaluation of cultural competence in AI-based generative models, focusing on cultural awareness and diversity [Kan*24]. These benchmark, while instrumental in assessing how well AI models represent diverse cultural artifacts, is limited on images, not including 3D representations.

In this context, our work aims to address this issue by presenting a novel evaluation framework tailored specifically for assessing 3D generative AI methods within the CH domain. Although previous studies have introduced AI-based techniques for heritage reconstruction and discussed general evaluation metrics for 3D models, there is still a lack of comprehensive frameworks that systematically evaluate the visual fidelity and geometric structure of 3D objects generated by state-of-the-art GenAI models when applied to heritage artifacts. Our framework addresses this issue by offering a thorough, multi-perspective assessment process that combines 2D and 3D analysis. It evaluates the quality of generated objects in terms of visual realism, structural coherence and geometric accuracy, taking into account the specific challenges of CH digitization such as complex ornamentation, surface degradation and material diversity. By focusing on the TRELLIS model, our work benchmarks its performance across a representative set of CH scenes and offers critical insights into the capabilities and limitations of current GenAI techniques for heritage preservation. Our goal is to bridge the gap between recent advances in generative modelling and the stringent requirements of CH applications, guiding future research towards more robust, culturally sensitive and application-ready AI-driven 3D reconstruction methods.

3. Materials and Methods

3.1. Preliminaries

The baseline model for 3D generation used in our framework is TRELLIS [XLX*24], a state-of-the-art genAI method for 3D model generation. The core of this method is the Structured LATent (SLAT) representation, which enables decoding into various 3D output formats such as Radiance Fields, 3D Gaussians, and meshes. SLAT defines a set of local latents on a sparse 3D grid to represent both geometry and appearance information:

$$z = \{(z_i, p_i)\}_{i=1}^{L}, \quad z_i \in \mathbb{R}^C, \quad p_i \in \{0, 1, \ldots, N-1\}^3, \quad (1)$$

where $L \ll N^3$ due to the sparsity of 3D data, allowing construction at relatively high resolutions ($N = 64$ by default). This structured latent representation is encoded from 3D assets by fusing dense multiview visual features extracted by a vision foundation model and can be decoded into diverse 3D representations.

The generation process follows a two-stage pipeline:

1. Sparse Structure Generation: The first stage generates the sparse structure of SLAT using rectified flow models. These models employ a linear interpolation forward process between data samples and noise, producing a binary 3D grid:

$$\text{Active voxels: } \{p_i\}, \quad \text{where } p_i \in \{0, 1\}. \quad (2)$$

2. Local Latent Generation: In the second stage, local latent vectors are generated for non-empty cells in the sparse structure. These latents capture both geometry and appearance information.

To handle the sparsity in SLAT efficiently, the framework employs rectified flow transformers with 3D shifted window attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (3)$$

where queries Q and keys K are normalized using root mean square normalization (RMSNorm) to ensure training stability.

The overall training objective incorporates reconstruction losses tailored to each representation:

$$L_{\text{total}} = L_{\text{geo}} + \lambda_1 L_{\text{color}} + \lambda_2 L_{\text{reg}}, \quad (4)$$

where L_{geo} and L_{color} measure geometric and appearance fidelity, respectively, and L_{reg} includes regularization terms such as consistency and deviation penalties.

The framework is trained on 500K high-quality 3D assets from 4 publicly available datasets: Objaverse (XL) [DLW*23], ABO [CGD*22], 3DFUTURE [FJG*21], and HSSD [KMJ*24]. This allows the model to generalize to diverse object structures and visuals.

This structured approach enables TRELLIS to generate 3D assets

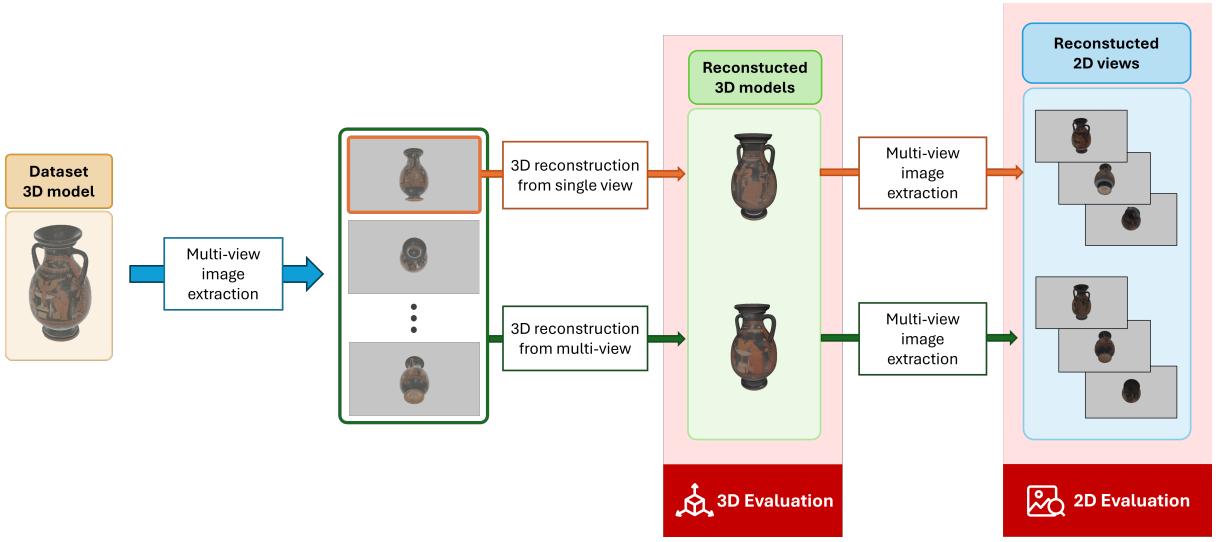


Figure 1: *Overview of the framework. Starting from a selected 3D object, multi-view images are extracted. Then, single-view and multi-view generation is performed. The generated 3D models are evaluated for 3D structural fidelity. Finally, 2D views are extracted from the 3D object and the images evaluated.*

in different formats. This is achieved through specialized decoders. In particular:

- 3D Gaussians: Each latent z_i is decoded into K Gaussians with position offsets o, colors c, scales s, opacities α, and rotations r:

$$D_{\text{GS}} : \{(z_i, p_i)\}_{i=1}^{L} \rightarrow \{\{(o_{ik}, c_{ik}, s_{ik}, \alpha_{ik}, r_{ik})\}_{k=1}^{K}\}_{i=1}^{L}, \quad (5)$$

  where the final positions x_{ik} of the Gaussians are constrained to the vicinity of their active voxel:

$$x_{ik} = p_i + \tanh(o_{ik}). \quad (6)$$

- Radiance Fields: For each active voxel, 4 orthogonal vectors v_x, v_y, v_z, v_c are predicted, representing the CP-decomposition of a local radiance volume:

$$V_{i,xyzc} = \sum_{r=1}^{R} v_{xi,r_x} v_{yi,r_y} v_{zi,r_z} v_{ci,r_c}, \quad (7)$$

  where $R = 16$ is the rank of decomposition.
- Meshes: The decoding process produces flexible parameters for mesh extraction:

$$D_{\text{M}} : \{(z_i, p_i)\}_{i=1}^{L} \rightarrow \{\{(w_{ji}, d_{ji})\}_{j=1}^{64}\}_{i=1}^{L}, \quad (8)$$

  where $w_{ji} = (\alpha_{ji}, \beta_{ji}, \gamma_{ij}, \delta_{ji})$ are interpolation and deformation parameters, and d_{ji} are signed distance values.

In our framework, we chose to focus on meshes as they are the most versatile and widely used representation in 3D modeling. Meshes provide a balance between geometric precision and computational efficiency, making them ideal for a wide range of applications.

3.2. Comparative Framework

Our novel evaluation framework, shown in Fig. 1, is structured to assess both visual fidelity and geometric accuracy of 3D models generated from CH artifacts using genAI [XLX*24]. The goal is to provide an easy to use and flexible framework to enable the evaluation of genAI approaches in a CH context.

Dataset creation The process begins with the creation of a synthetic dataset using Blender. The dataset was created by selecting five different CH scenarios and objects to ensure a wide variety of artifact types and scales. Specifically, the dataset includes two object (a Greek vase and a bust of Nefertiti), a minaret, a castle and a hermitage. Four of these models were sourced from publicly available repositories[†], while the minaret was created using photogrammetry and 3D modeling techniques [FWM*17]. Figure 1 details each dataset scene, with extracted images examples and the ground truth 3D models. A total of 125 high-resolution images (1920x1080 pixels) were rendered for each model, with camera positions uniformly distributed across a spherical sampling grid. From these, 100 images were selected as the training set for 3D generation, while the remaining 25 images served as a test set for 2D evaluation.

3D Generation and Evaluation Pipeline At the core of our framework lies a robust 3D generation and evaluation pipeline that systematically assesses the capabilities of genAI models (TRELLIS [XLX*24] in our case) in generating high-fidelity 3D reconstructions of CH artifacts. The pipeline is designed to support both single-view and multi-view generation modalities, enabling a comprehensive analysis of performance under varying levels of input information. In the single-view generation scenario, the model is provided with a single randomly selected RGB image from the

[†] https://sketchfab.com

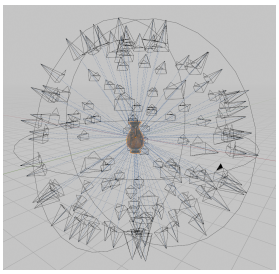
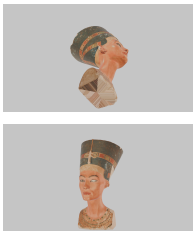
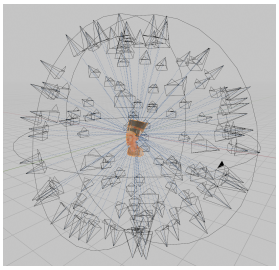
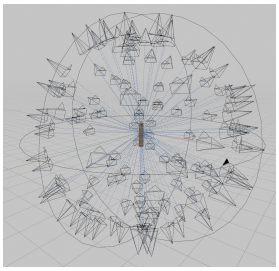
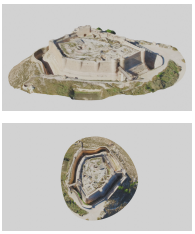
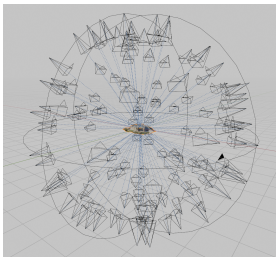
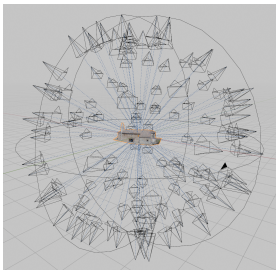
Dataset	Scene Type	Description	Extracted Images	Ground Truth
Greek Vase	Artifact	Pelike vase showing the love god Eros accompanied by two women near a basin.		
Nefertiti Bust	Artifact	Bust of Queen Nefertiti from the Egyptian Museum of Berlin		
Minaret	Architectural Structure	Minaret of the Omayyad Mosque, Damascus, Syria		
Castle	Architectural Structure	Castle of Chinchilla, Chinchilla de Monte-Aragón, Spain		
Hermitage	Architectural Structure	Hermitage of Santa Coloma, Guadalajara, Spain		

Table 1: *Dataset overview with related characteristics.*

training set of 100 views. This represents a challenging minimal-input setting, emulating real-world constraints where only limited visual documentation may be available. Conversely, in the multi-view generation scenario, the model receives the full set of 100 training images, offering a richly informed reconstruction process that leverages diverse viewpoints and lighting conditions. Once the

3D object is generated and it is exported in mesh format, a 2D re-rendering step is conducted. Here, 25 novel views are rendered from the generated 3D mesh using virtual cameras placed at the exact same positions as those used in the 25-image test set. This alignment allows for a pixel-level comparison between the ground truth and generated outputs, facilitating consistent and fair evaluation of visual fidelity.

The generated 3D models were evaluated in both 2D and 3D, assessing visual quality and structural fidelity. To evaluate the photorealism and perceptual fidelity of the generated objects, we compared the 25 ground-truth test images with the 25 images rendered from the generated 3D model under matching viewpoints. Based on the current literature [HBM*24; YPW23; WYZ*24], the following metrics were adopted:

- Peak Signal-to-Noise Ratio (PSNR): PSNR quantifies the difference between corresponding pixels of two images. It is computed as:

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{MAX^2}{\mathrm{MSE}}\right), \tag{9}$$

  where MAX is the maximum possible pixel value (typically 255 for 8-bit images), and MSE is the mean squared error between the predicted image I and the ground-truth image I^*:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(I_i - I_i^*)^2, \tag{10}$$

  Higher PSNR values indicate closer correspondence to the reference image.
- Structural Similarity Index Measure (SSIM): SSIM assesses the similarity between two images by comparing local patterns of pixel intensities that have been normalized for luminance and contrast. It is defined as:

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{11}$$

  where μ_x and μ_y are the local means, σ_x^2 and σ_y^2 are the local variances, and σ_{xy} is the local covariance between the predicted and reference image patches. C_1 and C_2 are small constants that stabilize the division. SSIM values range from 0 to 1, with higher values indicating greater structural similarity.
- Learned Perceptual Image Patch Similarity (LPIPS): LPIPS is a perceptual metric that compares the deep feature representations of two images as extracted by a pretrained convolutional neural network (e.g., VGG). Formally, LPIPS is defined as:

$$\mathrm{LPIPS}(x,y) = \sum_l \frac{1}{H_l W_l}\sum_{h,w}\left\| w_l \odot \left(f_l^x(h,w) - f_l^y(h,w)\right)\right\|_2^2, \tag{12}$$

  where f_l^x and f_l^y are the activation features at layer l of the network for images x and y, respectively, and w_l are learned weights for each channel. Lower LPIPS values indicate greater perceptual similarity.

These 2D metrics are specifically targeted at evaluating visual realism and texture fidelity.

For the assessment of geometric and structural integrity of the generated 3D models, we employed the following metrics, widely used in the literature [NSRR20; KPZK17; XLX*24], comparing the generated mesh against the ground truth 3D reference:

- Chamfer Distance (CD): CD measures the average closest-point distance between two point sets P and Q, sampled from the predicted and ground-truth surfaces, respectively. It is defined as:

$$\mathrm{CD}(P,Q) = \frac{1}{|P|}\sum_{p\in P}\min_{q\in Q}\|p-q\|_2^2 + \frac{1}{|Q|}\sum_{q\in Q}\min_{p\in P}\|q-p\|_2^2, \tag{13}$$

  Lower values indicate a closer alignment between the two point sets.
- F-score: The F-score evaluates the geometric overlap between two point clouds by computing the harmonic mean of precision and recall, given a distance threshold τ:

$$\text{F-score} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{14}$$

  where:

$$\mathrm{Precision} = \frac{1}{|P|}\sum_{p\in P}\mathbb{1}\left[\min_{q\in Q}\|p-q\|_2 < \tau\right], \tag{15}$$

$$\mathrm{Recall} = \frac{1}{|Q|}\sum_{q\in Q}\mathbb{1}\left[\min_{p\in P}\|q-p\|_2 < \tau\right], \tag{16}$$

  The threshold τ controls the tolerance for considering two points as matching. Higher F-scores indicate better alignment and completeness.
- Accuracy: Accuracy measures the signed point-to-surface distance between the predicted mesh and the reference mesh. Given a set of vertices v_i from the predicted mesh and a reference surface S, the accuracy is:

$$\mathrm{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} d(v_i, S). \tag{17}$$

  where $d(v_i,S)$ denotes the signed Euclidean distance from vertex v_i to the closest point on the surface S. Mean and standard deviation of these distances are reported to quantify both bias and dispersion. Lower absolute accuracy values and standard deviations indicate higher geometric fidelity.

This evaluation approach ensures a comprehensive assessment of both the visual realism and structural fidelity of the generated models, enabling a robust comparison across different generation modalities.

4. Experiments and Results

4.1. Experimental Settings

The experiments were conducted on a system running Ubuntu 22.04, equipped with an NVIDIA A6000 GPU with 48 GB of VRAM. For the image generation, we used Blender 4.4. To automate the generation of the diverse object views, we employed the BlenderNeRF add-on[‡], which allows automatic camera positioning and rendering around the target object. We employed an Axis-Aligned Bounding Box (AABB) of 8, a camera radius of 8.5m and a camera radius of 50mm. For the 3D generation, we employed the `TRELLIS-image-large` model, which is the image-to-3D

‡ https://github.com/maximeraafat/BlenderNeRF

Scene	2D Evaluation						3D Evaluation					
	PSNR ↑		SSIM ↑		LPIPS ↓		CD ↓		F-score ↑		Accuracy ↓	
	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi
Greek Vase	20.68	**22.72**	0.915	**0.929**	0.057	**0.047**	**0.002**	**0.002**	**0.996**	0.993	**0.001(0.03)**	-0.005(0.032)
Nefertiti Bust	20.38	**24.26**	0.882	**0.904**	0.103	**0.072**	0.011	**0.001**	0.805	**1.000**	0.026(0.077)	**0.005(0.025)**
Minareto	26.88	**34.12**	0.974	**0.995**	0.014	**0.003**	**0.001**	**0.001**	0.997	**0.998**	0.003(0.016)	**-0.001(0.014)**
Castle	14.58	**14.88**	**0.795**	0.793	0.171	**0.167**	0.020	**0.016**	0.744	**0.773**	**0.003(0.016)**	-0.019(0.066)
Hermitage	16.37	**18.88**	0.747	**0.767**	0.225	**0.199**	0.018	**0.006**	0.792	**0.965**	-0.022(0.076)	**0.01(0.05)**

Table 2: *Comprehensive evaluation metrics for both 2D visual quality and 3D structural fidelity on the 5 selected CH scenes. PSNR, SSIM and LPIPS are used for 2D evaluation. CD, F-score and Accuracy, in terms of Mean(STD), are used for 3D evaluation.*

1.2B parameter implementation of the TRELLIS framework. All hyperparameters were maintained consistent with the original work [XLX*24] to ensure comparability with the state-of-the-art results. In particular, CFG strength was set to 3 and sampling steps to 50. For the evaluation phase, we leveraged the Torchmetrics[§] Python library for 2D evaluation (PSNR, SSIM, LPIPS) and a combination of the Pytorch3D[¶] Python library (CD, F-score) and the CloudCompare[‖] software (Mean Accuracy, STD Accuracy) for 3D evaluation.

4.2. Results

Two distinct generation scenarios were systematically evaluated to assess the model's performance under different input conditions: single-view generation and multi-view generation. In the single-view generation scenario, the model was conditioned using a single randomly selected image from the training set, representing the most challenging case with minimal input information. Conversely, the multi-view generation scenario utilized the complete set of 100 training images, providing the model with comprehensive visual information about the target object or scene. For evaluation, we rendered 25 test images from both the ground-truth 3D models and the generated counterparts, ensuring identical camera viewpoints to enable direct comparisons.

4.2.1. 2D Visual Quality Evaluation

The quantitative evaluation of visual fidelity, presented in Table 2, reports the results in terms of PSNR, SSIM, and LPIPS, providing a comprehensive assessment of image quality, structural similarity, and perceptual differences respectively.

The results demonstrate a clear distinction between the model's performance on individual objects versus complex scenes. For single-object reconstructions such as the Minareto, Greek Vase, and Nefertiti Bust, the model achieved exceptional performance metrics, with the Minareto reaching a PSNR of 34.12 in multi-view mode. This superior performance can be attributed to the relatively simpler geometry and more uniform material properties of these individual artifacts. In contrast, the more complex Castle and Hermitage scenes presented greater challenges, as evidenced by their lower metric scores, particularly in the single-view condition where the PSNR for the Castle dropped to 14.58. Across all test cases, the multi-view generation consistently outperformed the single-view approach, though the degree of improvement varied significantly depending on the complexity of the subject. The most substantial gains were observed in single object scenes, suggesting that additional visual information becomes increasingly valuable if there is an already good single-view representation. For the more complex scenes, increases in performances can be seen between the single-view and multi-view generation, but on a much lower rate. This can be attributed to the high complexity of the scene, which hinders the generation quality.

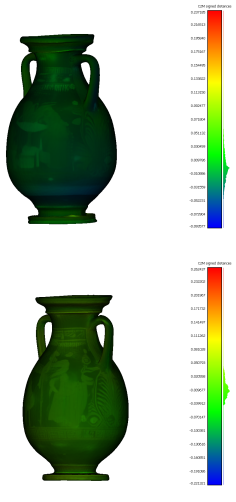
4.2.2. 3D Structural Fidelity Evaluation

The assessment of 3D structural accuracy, detailed in Table 2, provides crucial insights into the geometric fidelity of the generated models. The Chamfer Distance and F-score metrics offer complementary perspectives on the accuracy and completeness of the reconstructed geometries. The 3D evaluation results strongly correlate with the 2D findings, with individual objects again demonstrating superior reconstruction quality. The Minareto and Greek Vase achieved near-perfect F-scores in both generation modes, with correspondingly minimal CDs. The Nefertiti Bust showed particularly notable improvement in the multi-view condition, achieving a perfect F-score compared to 0.805 in single-view mode. For architectural scenes, while the absolute metrics were lower, they still show a good structural fidelity compared to the original 3D objects, with the multi-view approach providing substantial benefits, such as reducing the CD for the Hermitage by a factor of three (0.006) compared to single-view generation (0.018). As for the Accuracy, all the scenes presented very low mean values and low standard deviations, aligning with the other metrics. Accuracy results are also visually presented in Fig. 2 (distance threshold: 0.1). Qualitative results of the generated 3D models are presented in Fig. 3, providing visual confirmation of the quantitative findings. The single-object reconstructions exhibit good fidelity in both geometry and texture, while the multi-view versions show particularly sharp details and accurate material representation.

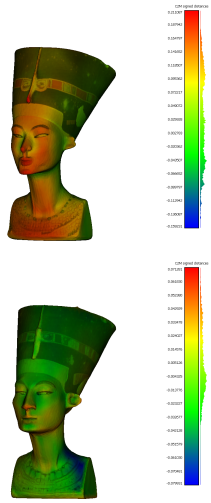
4.3. Results Discussion

The comprehensive evaluation reveals several important insights about the capabilities and limitations of genAI for CH preservation. The outstanding performance on individual objects suggests that current generative approaches have reached a level of maturity suitable for documenting and reconstructing discrete artifacts. The consistently high scores across both 2D and 3D metrics for objects like the Minareto, Greek Vase and Nefertiti Bust indicate

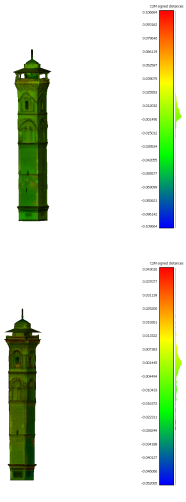
[§] https://github.com/Lightning-AI/torchmetrics
[¶] https://pytorch3d.org/
[‖] https://www.cloudcompare.net/



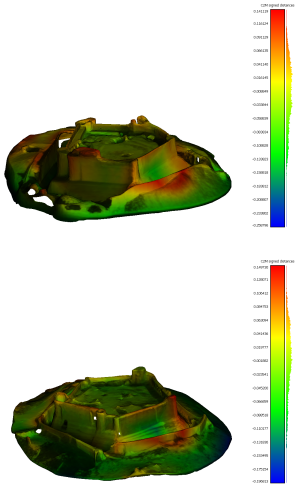
(a) *Single-view (top) and multi-view (bottom) Accuracy on the Greek Vase scene*



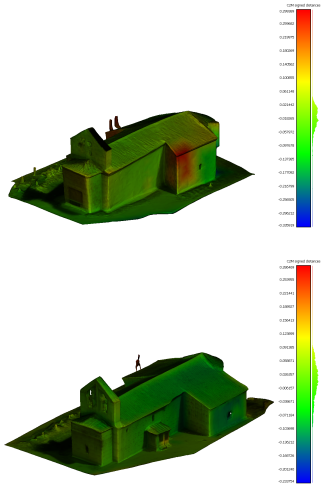
(b) *Single-view (top) and multi-view (bottom) Accuracy on the Nefertiti Bust scene*



(c) *Single-view (top) and multi-view (bottom) Accuracy on the Minareto scene*



(d) *Single-view (top) and multi-view (bottom) Accuracy on the Castle scene*



(e) *Single-view (top) and multi-view (bottom) Accuracy on the Hermitage scene*

Figure 2: *Accuracy results on the 5 dataset scenes, for both single-view (top) and multi-view (bottom) generations.*

that these methods can reliably capture both the visual appearance and geometric form of such subjects. This is mainly due to a combination of object structure and model training data. The 3 single objects are mostly symmetric, presenting an easier reconstruction scenario than the Castle and Hermitage scenes, which present different shape and more complex structures, resulting in a less precise reconstruction. Furthermore, the datasets used to train the TRELLIS model do not include large scenes tied specifically to CH, while presenting a large amount of single objects. This disparity is reflected in the results, as the reconstruction struggled more with the Castle and Hermitage scenes, while generalizing well on the other scenes. Furthermore, the Castle and Hermitage scenes present an increased number of surfaces and potential occlusions and a larger physical scale, requiring more extensive context understanding. These findings have significant implications for practical applications in CH. For individual artifacts, the results support the use of generative approaches even with limited input imagery, while for architectural heritage, they suggest the need for more comprehensive documentation efforts or potentially hybrid approaches combining generative AI with traditional techniques. The consistent advantage of multi-view generation across all test cases underscores the importance of thorough documentation when creating digital preservation records.

(a) *Greek Vase ground truth* **(b)** *Greek Vase single-view generation* **(c)** *Greek Vase multi-view generation*

(d) *Nefertiti Bust Ground truth* **(e)** *Nefertiti Bust single-view generation* **(f)** *Nefertiti Bust multi-view generation*

Figure 3: *Qualitative examples from the generation results from the Greek Vase and Nefertiti Bust scenes.*

5. Conclusion and Future works

This work introduced a novel evaluation framework for assessing generative AI models in CH, with a specific focus on the TRELLIS model. Our experimental results demonstrate that modern 3D generative approaches achieve remarkable fidelity in reconstructing individual cultural artifacts. The quantitative metrics revealed consistent superiority of multi-view generation across all test cases. The framework successfully identified key performance patterns: single-object generation consistently outperformed architectural scenes in both visual fidelity (average SSIM 0.943 vs. 0.780) and geometric accuracy (average CD of 0.0013 m vs. 0.011 m). This performance gap highlights the current limitations of generative approaches when dealing with complex, large-scale heritage sites containing diverse materials and occluded structures. These findings provide valuable insights for CH professionals, suggesting that, while genAI is already viable for artifact digitization, architectural heritage preservation may require hybrid approaches combining AI with traditional techniques for more complex environments. The proposed evaluation metrics offer a standardized methodology for comparing different reconstruction techniques in heritage applications. Based on the limitations identified, we propose several directions for future research. First, hybrid approaches combining generative AI with traditional photogrammetric methods could address the current limitations in architectural reconstruction, particularly for complex structural elements. Second, the development of sparse-view optimization techniques could maintain reconstruction quality while significantly reducing documentation requirements, potentially operating effectively with just 5-20 input images. Finally, comparing with other Neural Rendering representations could help in addressing the unique needs of heritage conservation, potentially enabling more accurate digital preservation of vulnerable artifacts and sites. These advancements could further bridge the gap between experimental results and practical heritage documentation needs.

References

[AXF*23] ANCIUKEVIČIUS, TITAS, XU, ZEXIANG, FISHER, MATTHEW, et al. "Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, 12608–12618.

[BCP*24] BALLONI, EMANUELE, CEKA, DAVID, PIERDICCA, ROBERTO, et al. "Comparative assessment of Neural Rendering methods for the 3D reconstruction of complex heritage sites in the inner areas of the Marche region-Italy". *Digital Applications in Archaeology and Cultural Heritage* 35 (2024), e00371.

[BGP*23] BALLONI, E, GORGOGLIONE, L, PAOLANTI, M, et al. "Few shot photogrametry: A comparison between nerf and mvs-sfm for the documentation of cultural heritage". *INTERNATIONAL ARCHIVES OF THE PHOTOGRAMMETRY, REMOTE SENSING AND SPATIAL INFORMATION SCIENCES* 48 (2023).

[CFB*24] CROCE, V, FORLEO, G, BILLI, D, et al. "Neural Radiance Fields (Nerf) For Multi-Scale 3D Modeling Of Cultural Heritage Artifacts". *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 48 (2024), 165–171.

[CGD*22] COLLINS, JASMINE, GOEL, SHUBHAM, DENG, KENAN, et al. "Abo: Dataset and benchmarks for real-world 3d object understanding". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, 21126–21136.

[Dan*24] DANG, X. et al. "Bridging the past and present: AI-driven 3D restoration of degraded artefacts for museum digital display". *Journal of Cultural Heritage* 69 (2024), 18–26.

[DLW*23] DEITKE, MATT, LIU, RUOSHI, WALLINGFORD, MATTHEW, et al. "Objaverse-xl: A universe of 10m+ 3d objects". *Advances in Neural Information Processing Systems* 36 (2023), 35799–35813.

[DPN*22] DEVAGIRI, JEEVAN S, PAHEDING, SIDIKE, NIYAZ, QUAMAR, et al. "Augmented Reality and Artificial Intelligence in industry: Trends, tools, and future challenges". *Expert Systems with Applications* 207 (2022), 118002.

[DSG*24] DAHAGHIN, M., SATTLER, T., GARBIN, S., et al. "Gaussian Heritage: 3D Digitization of Cultural Heritage with Integrated Object Segmentation". *arXiv preprint arXiv:2409.19039* (2024).

[FJG*21] FU, HUAN, JIA, RONGFEI, GAO, LIN, et al. "3d-future: 3d furniture shape with texture". *International Journal of Computer Vision* 129 (2021), 3313–3337.

[FWM*17] FANGI, GABRIELE, WAHBEH, WISSAM, MALINVERNI, EVA SAVINA, et al. "Archaeological Syrian Heritage memory safeguard by low cost geomatics techniques". *IMEKO International Conference on Metrology for Archaeology and Cultural Heritage, Lecce, Italy, October*. 2017, 23–25.

[HBM*24] HÖLLEIN, LUKAS, BOŽIČ, ALJAŽ, MÜLLER, NORMAN, et al. "Viewdiff: 3d-consistent image generation with text-to-image models". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, 5043–5052.

[HKP*22] HOU, YUMENG, KENDERDINE, SARAH, PICCA, DAVIDE, et al. "Digitizing intangible cultural heritage embodied: State of the art". *Journal on Computing and Cultural Heritage (JOCCH)* 15.3 (2022), 1–20.

[HZG*23] HONG, YICONG, ZHANG, KAI, GU, JIUXIANG, et al. "Lrm: Large reconstruction model for single image to 3d". *arXiv preprint arXiv:2311.04400* (2023).

[JS24] JARAMILLO, P. and SIPIRAN, I. "Cultural Heritage 3D Reconstruction with Diffusion Networks". *arXiv preprint arXiv:2410.10927* (2024).

[Kan*24] KANNEN, N. et al. "Beyond Aesthetics: Cultural Competence in Text-to-Image Models". *arXiv preprint arXiv:2407.06863* (2024).

[KMJ*24] KHANNA, MUKUL, MAO, YONGSEN, JIANG, HANXIAO, et al. "Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, 16384–16393.

[KPZK17] KNAPITSCH, ARNO, PARK, JAESIK, ZHOU, QIAN-YI, and KOLTUN, VLADLEN. "Tanks and temples: Benchmarking large-scale scene reconstruction". *ACM Transactions on Graphics (ToG)* 36.4 (2017), 1–13.

[LGL*24] LONG, XIAOXIAO, GUO, YUAN-CHEN, LIN, CHENG, et al. "Wonder3d: Single image to 3d using cross-domain diffusion". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, 9970–9980.

[LHH*24] LIU, JIAN, HUANG, XIAOSHUI, HUANG, TIANYU, et al. "A Comprehensive Survey on 3D Content Generation". *arXiv preprint arXiv:2402.01166* (2024). URL: https://arxiv.org/abs/2402.01166.

[LLL*24] LI, PENG, LIU, YUAN, LONG, XIAOXIAO, et al. "Era3d: high-resolution multiview diffusion using efficient row-wise attention". *Advances in Neural Information Processing Systems* 37 (2024), 55975–56000.

[LTGR23] LLULL, PATRICK, THOMPSON, EMILY, GARCIA, MIGUEL, and ROSSI, LAURA. "Evaluation of 3D Reconstruction Techniques for Cultural Heritage Preservation". *Journal of Cultural Heritage* 58 (2023), 123–134. DOI: 10.1016/j.culher.2023.01.012.

[LZW*23a] LIU, ZHEN, ZHANG, YUXIN, WANG, YINDA, et al. "One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization". *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, 12345–12354. DOI: 10.1109/ICCV.2023.12345.

[LZW*23b] LIU, ZHEN, ZHANG, YUXIN, WANG, YINDA, et al. "One-2-3-45++: Enhancing Single Image to 3D Mesh Conversion with Improved Fidelity". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, 23456–23465. DOI: 10.1109/CVPR.2023.23456.

[MGF*23] MARCHELLO, G, GIOVANELLI, R, FONTANA, E, et al. "Cultural heritage digital preservation through ai-driven robotics". *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 48 (2023), 995–1000.

[MLR*24] MELAS-KYRIAZI, LUKE, LAINA, IRO, RUPPRECHT, CHRISTIAN, et al. "Im-3d: Iterative multiview diffusion and reconstruction for high-quality 3d generation". *arXiv preprint arXiv:2402.08682* (2024).

[NSRR20] NOCERINO, ERICA, STATHOPOULOU, ELISAVET KONSTANTINA, RIGON, SIMONE, and REMONDINO, FABIO. "Surface reconstruction assessment in photogrammetric applications". *Sensors* 20.20 (2020), 5863.

[QWZ*24] QIAN, CHEN-HSUAN, WANG, YIFAN, ZHANG, ZEXIANG, et al. "Magic123: One Image to High-Quality 3D Model Using Both 2D and 3D Diffusion Priors". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, 34567–34576. DOI: 10.1109/CVPR.2024.34567.

[RSL*24] RIBEIRO, MARCOS, SANTOS, JOANA, LOBO, JOÃO, et al. "VR, AR, gamification and AI towards the next generation of systems supporting cultural heritage: Addressing challenges of a museum context". *Proceedings of the 29th International ACM Conference on 3D Web Technology*. 2024, 1–10.

[Shi25] SHIH, NAI-JIE. "AI-assisted 3D Preservation and Reconstruction of Temple Arts". *arXiv preprint arXiv:2503.10031* (2025).

[Spe24] SPENNEMANN, DIRK HR. "Generative artificial intelligence, human agency and the future of cultural heritage". *Heritage* 7.7 (2024), 3597.

[SWY*23] SHI, YICHUN, WANG, PENG, YE, JIANGLONG, et al. "Mvdream: Multi-view diffusion for 3d generation". *arXiv preprint arXiv:2308.16512* (2023).

[TWZ*23a] TANG, JUNSHU, WANG, TENGFEI, ZHANG, BO, et al. "DreamGaussian: Generative 3D Gaussian Splatting for Efficient Text-to-3D Synthesis". *arXiv preprint arXiv:2308.12345* (2023). URL: https://arxiv.org/abs/2308.12345.

[TWZ*23b] TANG, JUNSHU, WANG, TENGFEI, ZHANG, BO, et al. "Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior". *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, 22819–22829.

[TZC*25] TANG, ZHENYU, ZHANG, JUNWU, CHENG, XINHUA, et al. "Cycle3d: High-quality and consistent image-to-3d generation via generation-reconstruction cycle". *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 7. 2025, 7320–7328.

[WLW*24] WANG, ZHEN, LI, DONGYUAN, WU, YAOZU, et al. "Diffusion models in 3d vision: A survey". *arXiv preprint arXiv:2410.04738* (2024).

[WYZ*24] WU, TONG, YUAN, YU-JIE, ZHANG, LING-XIAO, et al. "Recent advances in 3d gaussian splatting". *Computational Visual Media* 10.4 (2024), 613–642.

[XLX*24] XIANG, JIANFENG, LV, ZELONG, XU, SICHENG, et al. "Structured 3d latents for scalable and versatile 3d generation". *arXiv preprint arXiv:2412.01506* (2024).

[YPW23] YANG, JIAWEI, PAVONE, MARCO, and WANG, YUE. "Freenerf: Improving few-shot neural rendering with free frequency regularization". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, 8254–8263.