Evaluating Generative AI for Museum Artifacts Documentation

Elisa Mariarosaria Farella, Simone Rigon, Gianluca Bertolasi, Fabio Remondino

3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy
Email: (elifarella, srigon, gbertolasi, remondino)@fbk.eu

Keywords: heritage digitization, generative AI, single-image 3D reconstruction, geometric assessment, texture fidelity

Abstract
In recent years, the European Commission (EC) identified the 3D digitization of cultural heritage sites and artifacts as one of its priorities and promoted numerous initiatives and recommendations to accelerate documentation campaigns. However, current digitization targets remain far from being achieved, and heritage institutions have been increasingly encouraged to explore faster and cost-effective 3D documentation solutions. Moreover, traditional image- and range-based 3D surveying techniques frequently struggle when reconstructing objects featuring non-collaborative surfaces (such as reflective or transparent objects), are time-consuming, and require specialized knowledge. Generative AI methods, able to generate 3D models also from a single input image, have recently emerged as a potentially faster alternative, yet their performance on heritage assets remains mostly unexplored. This paper evaluates three state-of-the-art and recent single-image GenAI frameworks - SAM3D, Tripo3D and Trellis2 - on several museum artifacts featuring diffuse, reflective, transparent, and mixed-material surfaces of varying scale and geometric complexity, for which accurate ground truth is available. The aim is to analyze whether these frameworks can act as complementary or alternative solutions for fast heritage documentation.

Figure 1. Museum heritage artifacts used in this study for the GenAI frameworks evaluation.

1. Introduction

The European Commission (EC) has recently published a report (European Commission, 2024) assessing the progress made during the 2021–2023 two-year period in implementing its Recommendation on a Common European Data Space for Cultural Heritage (CNECT/LUX/2021/OP/0070). The Recommendation fosters Member States to enhance heritage access, protection, and preservation through vast and ambitious digitization campaigns of sites, monuments, libraries, archives, and museums across Europe. The report highlights that the goal of digitizing 40% of Europe’s heritage by 2025 and 50% by 2030 remains far from being achieved and urges Member States to intensify digitization efforts, fostering the adoption of advanced solutions such as Artificial Intelligence (AI) and Virtual Reality (VR). In support of heritage institutions and professionals, the same Commission has established and spread guidelines to standardize digitization practices and ensure the high-quality of digitized assets. From basic principles to workflows, tips, and handbooks on best practices, these guidelines cover a broad range of 2D and 3D well-established digitization technologies and encourage the adoption of standardized metadata schemas for the proper reuse and integration of the digitized sources into different platforms. While these recommendations are useful references for consolidated image- and range-based 3D surveying techniques, they do not cover emerging methods.
In the last few years, new AI-based rendering and reconstruction techniques have captured attention for their ability to generate highly realistic 3D models of complex heritage assets. Despite their promising results in several fields, from gaming to VR applications, the adoption of these techniques in the cultural heritage (CH) domain remains experimental, as standardized workflows, quality control procedures, and integration guidelines are still lacking.
Neural Radiance Field (NeRF) (Mildenhall et al., 2021; Remondino et al., 2023; Cai and Lu, 2024) and Gaussian Splatting (GS) techniques (Kerbl et al., 2023; Fei et al., 2024) have emerged, respectively, as advanced AI-based 3D reconstruction and rendering solutions. While NeRF exploits neural networks to synthesize realistic 3D scenes from a set of 2D images, GS allows for real-time rendering of novel views even in the case of scenes with non-collaborative surfaces. Concurrently, Generative AI (GenAI) techniques (Sengar et al., 2025; Elhanashi et al., 2026) are attracting attention in the image-to-3D domain. By leveraging neural networks and advanced Deep Learning (DL) algorithms, these techniques allow for creating realistic 3D models (Liu et al., 2024a) from minimal input (often just a single image) and/or text prompts, opening up new opportunities for faster and cost-effective content creation. Unless NeRF and GS, GenAI solutions do not require a set of oriented images for creating 3D assets, but the models learn the underlying 3D structure, geometry, and appearance from large-scale datasets, commonly compressing this information into latent or volumetric representations. During inference, some methods can generate 3D outputs in a feed-forward manner (e.g., generating reconstructions directly from an input in a single pass and without additional refinements), or can rely on iterative optimization techniques to refine the geometry in multiple stages using loss functions or diffusion models. Despite the evident convenience of reducing documentation time and effort, the quality of AI-generated assets in the heritage sector has only recently begun to be investigated (Jaramillo and Sipiran, 2024;

Balloni et al., 2025; Wang et al., 2025). Furthermore, since consolidated 3D surveying techniques are already effective in delivering high-quality 3D models but struggle when reconstructing objects featuring non-collaborative surfaces (Farella et al., 2022; Yan et al., 2025) and are often time-consuming, the potential of generative methods - as an alternative or complementary solutions for rapid documentation - needs to be analyzed, taking into account the geometric and textural heterogeneity and variability of heritage assets.

1.1 Aim of the work

The article investigates the capability of GenAI methods to reconstruct 3D heritage assets from a single input image (since currently only a few methods support multi-view data). Pushed by the EC recommendations for vast and ambitious digitization campaigns (European Commission, 2024), the goal is to assess GenAI performance with heritage museum artifacts, featuring diverse and challenging attributes, such as non-collaborative surfaces (e.g., reflective, uniform textures, or transparent objects), symmetrical, or complex shapes.
The work, in particular, aims to assess whether these methods can serve as an effective, rapid, and metric documentation tool, especially in contexts where heritage is at risk (e.g., conflict areas or emergency situations) or where museums and institutions lack technical expertise or resources to conduct a full digitization campaign. While previous studies in the CH domain primarily focused on restoration of heritage objects via diffusion models (Jaramillo and Sipiran, 2024), painted pottery (Wang et al., 2025), or the evaluation of a single method with limited sets of museum artifacts (Balloni et al., 2025), the contribution of our work is:

- to extend the evaluation of GenAI 3D reconstruction to a large variety of CH artifacts, featuring different materials, shapes, dimensions, and levels of complexity, for which accurate photogrammetric ground truth data are available;
- to assess the performance across a broader set of conditions, e.g., mixed materials, non-collaborative, reflective, transparent and complex surfaces;
- to compare different state-of-the-art and recent GenAI frameworks by testing their capabilities with single-image processing.

2. Related works

Early approaches for 3D content generation based on deep generative models include Generative Adversarial Networks (GANs) (Wu et al., 2016) and Variational Autoencoders (VAEs) (Kingma and Welling, 2014), which, despite promising results, have often faced scalability and reconstruction challenges for detailed topologies.
The limited performance of these approaches encouraged the development of diffusion models, originally developed for image generation (Ho et al., 2020; Rombach et al., 2022), for 3D tasks. These generative frameworks learn to iteratively transform random noise into structured data by reversing a gradual noising process and modelling complex data distribution. 2D-to-3D approaches, such as DreamFusion (Poole et al., 2023), leverage pre-trained 2D diffusion models to generate 3D assets from text. In such cases, a 3D representation is optimized through score distillation in order to match synthesized 2D views of the object (He et al., 2023; Liu et al., 2023).
Native 3D diffusion approaches, which train directly on 3D data and limit reliance on 2D priors, such as Michelangelo (Zhao et al., 2023) or TripoSG (Li et al., 2025), promise to further improve the model's quality.
Building on these advancements, methods such as Trellis (Xiang et al., 2025a) combine structured 3D latent representations with rectified-flow transformers to produce multiple 3D outputs (meshes, radiance fields, Gaussian Splats). These approaches integrate sparse 3D grids and multi-view visual features extracted from a foundational model.
More recent developments, like Trellis2 (Xiang et al., 2025b), introduce latent representations and new geometric formulations (i.e., O-Voxel structures) to improve the representation of complex and non-manifold geometries. By leveraging transformer-based architectures with flow-matching objectives, these models produce 3D textured assets with physically based rendering (PBR) properties, featuring higher geometric fidelity and faster inference. Specifically, material opacity is integrated to better handle translucent materials.
A further class of methods exploits feedforward inference for single-view 3D reconstruction (Liu et al., 2024b). One implementation in this field is Triposr (Tochilkin et al., 2024), which predicts watertight meshes from a single RGB image. Unlike diffusion-based approaches, which rely on iterative sampling and implicit multi-view consistency, TripoSR directly predicts 3D geometry from a single RGB image, leveraging a transformer architecture for fast feed-forward 3D generation. Based on this implementation, a powerful and commercial solution, Tripo3D, has been recently released.
Concurrently, approaches based on 3D visual grounding emerged. A representative example is SAM3D (Chen et al., 2025), which builds upon the Segment Anything Model (Kirillov et al., 2023) to associate objects in 2D images with corresponding 3D representations. Unlike traditional segmentation-based workflows, SAM3D employs a human-and-model-in-the-loop pipeline for annotating shapes, textures, poses, and positions. The model is trained in multiple stages, combining synthetic pretraining with real-world alignment to improve 3D visual grounding accuracy.

3. Methods and evaluation

3.1 Selected GenAI methods

Trellis2, SAM3D, and Tripo3D are examined in this work as state-of-the-art and recent solutions spanning structured latent generative models, 3D visual grounding, feedforward reconstruction and triplane-NeRF transformer-based frameworks. While structured latent generative models encode 3D objects into a compact latent space that captures essential geometry, topology, and material properties, ensuring efficient processing (Trellis2), feed-forward and multi-stage reconstruction models with 3D visual grounding predict object shape and pose directly from images, associating 2D objects with their 3D counterparts in multi-object scenes (SAM3D). Triplane-NeRF transformer-based frameworks leverage features extracted by the transformers and NeRF representations on three orthogonal planes to reconstruct geometry and textures (Tripo3D). In particular:

- *Trellis2* (Xiang et al., 2025b)[1]: it is a recent evolution of Trellis for creating high-resolution 3D content with complex topologies and flexible material properties. It introduces the O-Voxel, a native 3D representation encoding each voxel with both geometric and material information. Arbitrary topologies, such as non-watertight or self-intersecting areas, are better modelled through a Flexible Dual Grid. Material properties are effectively represented through physically-

[1] https://microsoft.github.io/TRELLIS.2/

based rendering (PBR) parameters (such as color, opacity, metallicity, and roughness), which enable a higher realistic object rendering. A fully sparse convolutional network (Sparse Compression VAE) is exploited for computational efficiency at high resolution. Geometry and materials are modelled separately while maintaining their spatial alignment.

- *Tripo3D*[2]: it is a commercial solution based on Triposr (Tochilkin et al., 2024) implementation. It is a transformer-based framework for single-image 3D reconstruction. The model takes a single RGB image as input and leverages a triplane-based Neural Radiance Field (NeRF) for generating the 3D output. The image decoder uses a pre-trained DINOV1 vision transformer to extract global and local features, while an image-to-triplane decoder maps latent vectors into a triplane-NeRF representation. The NeRF model encompasses multi-layer perceptrons (MLPs) in charge of predicting the color and density of 3D points in space.
- *SAM3D* (Chen et al., 2025)[3]: it is a single-image 3D reconstruction framework released by META, able to predict object shape, texture, pose, and scale. Unlike most methods, it can generate coherent multi-object scenes. The model builds on a two-stage latent flow architecture. For modelling the geometry and predicting the object layout (rotation, translation, scale), it leverages a Mixture-of-Transformer (MoT) flow transformer. For texture and model refinement, it synthetizes textures using a sparse latent flow transformer. The implemented multi-stage training strategies enable scaling from synthetic isolated objects to complex real-world scenes, handling occlusions efficiently. The outputs are meshes or 3D Gaussian splats.

3.2 Datasets and test objects

The evaluation of the aforementioned GenAI frameworks is conducted on seven diverse heritage artefacts (Figure 1, Table 1) for which an accurate photogrammetric ground truth is available. Additionally, to assess their performance on transparent assets, which are commonly part of museum collections and yet difficult to be digitized with optical techniques, a glass object (Figure 3) from a benchmark dataset[4] is included in the experiments. In particular, the test objects (Figures 1, 2, 3) include:

- a pottery (ceramic vessel), with a matte surface and regular geometry, used for evaluating the reconstruction of diffuse and non-reflective materials featuring some decorations;
- a wooden bowl exhibiting textural irregularities and fine handcrafted details, used for assessing the reconstruction with organic materials and small-scale geometries;
- two small bronze statuettes with highly reflective metallic surfaces, which typically complicates 3D reconstruction processes due to specular reflections;
- a prehistoric saw tool composed of wood and flint, exhibiting distinct reflectance properties;
- a prehistoric wooden helmet, featuring an intertwined structure with cavities, used to test the model's capability to infer occluded or complex topologies;
- a marble metope fragment, depicting two large figures and characterized by a light color and a relatively matte surface, providing an example of a stone material with a mild reflectivity;
- a transparent glass (ground truth – photogrammetry on powdered surfaces), included to evaluate the behavior on transparent and reflective surfaces. For this case, only the reconstructed geometry and not the texture quality is assessed.

The characteristics of the datasets, the number of faces of the available ground truth, and the corresponding reconstructed geometries are reported in Table 1. For methods that did not allow direct control over the number of reconstructed faces, the generated 3D outputs are decimated to ensure consistency across all models for a fair comparison.

Object	Dimensions [mm]	Faces GT	Faces AI
Pottery	289x87x315	1M	712k
Bowl	206x197x122	2.3M	488k
Bronze_1	65x22x44	429k	156k
Bronze_2	120x60x50	211k	136k
Helmet	290x247x239	7.2M	858k
Saw	283x87x315	402k	101k
Metope	834x1011x512	1.6M	405k
Glass	80x80x236	948k	140k

Table 1. Datasets description.

3.3 Methods evaluation - metrics

For the evaluation, a Hausdorff signed distance is used to quantify geometric discrepancies between the reconstructed GenAI meshes and the photogrammetric or synthetic ground truth. Artifacts dimensions are also reported in Table 1 to better interpret the significance of errors and deviations, although scale is not explicitly estimated by the methods. On the other hand, the texture quality is assessed using multiple image-based metrics (Table 2). These metrics capture not only pixel-level differences but also structural, perceptual, and color-based discrepancies between the reconstructed textures and the ground truth images.

- MSE (Mean Squared Error) and PSNR (Peak Signal-to Noise Ratio) quantify low-level pixel differences. Despite being widely used for image quality assessment, they cannot fully capture human perception, particularly for structural or semantic differences.
- SSIM (Structural Similarity Index) and MS-SSIM (Multi-Scale SSIM) evaluate structural similarities at single and multiple scales, respectively. These metrics better correlate with perceptual quality, as they look at patterns of intensity, contrast, and texture to define how humans perceive image quality. While SSIM compares local regions of the image and can highlight structural errors, being sensitive to distortions in edges, shapes, and textures, MS-SSIM extends the evaluation on multiple scales, capturing differences that are indistinct at a single resolution but affect the overall perception of the image. Both, focusing on structural information, are less sensitive to color shifts.
- LPIPS (Learned Perceptual Image Patch Similarity) is a learned perceptual similarity metric that outperforms traditional metrics in modelling human perception. Unlike pixel-based metrics, it compares deep feature representations of image patches and is sensitive to differences in texture, style, and fine-grained perceptual details. However, its sensitivity depends on the underlying neural network used to extract features, and it can overemphasize differences that are perceptually minor.
- ΔE (CIELAB color difference) specifically measures perceptual color discrepancies in the CIELAB space. It is highly related to human vision and perception, but it ignores structural differences.

[2] https://www.tripo3d.ai/
[3] https://ai.meta.com/research/sam3d/
[4] https://github.com/3DOM-FBK/NeRFBK

Metric	Formula	Typical Range	Interpretation
MSE (Mean Squared Error)	$MSE = \frac{1}{N}\sum_{i-1}^{N}\left(I_1(i) - I_2(i)\right)^2$	$0 \rightarrow \infty$	Average pixel-wise squared error. 0 = identical Higher values = larger differences
PSNR (Peak Signal-to-Noise Ratio)	$PSNR = 10 \times log_{10}\frac{MAX^2}{MSE}$	0 → high dB	Higher values → better quality > 30 dB = very similar < 20 dB = very different
SSIM (Structural Similarity Index)	$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xY} + C_2)}{(\mu_x{}^2 + \mu_Y{}^2 + C_1)(\sigma_x{}^2 + \sigma_Y{}^2 + C_2)}$	$0 \rightarrow 1$	Measures structural similarity. 1 = identical <0.6 = large structural differences
MS – SSIM (Multi-Scale SSIM)	$MS - SSIM = \prod_{j=1}^{M} SSIM_j^{w_j}$	$0 \rightarrow 1$	Similar to SSIM, but evaluates multiple scales. Lower values indicate differences in structure across resolutions.
LPIPS (Learned Perceptual Image Patch Similarity)	$LPIPS\ (x, x_0) = \sum_l \frac{1}{H_l W_l}\sum_l \left\| w_l \odot \left(\widehat{y_{hw}}^{(l)} - \widehat{x_{0hw}}^{(l)}\right)\right\|_2^2$	$0 \rightarrow 1$	Learned perceptual similarity. 0 = identical >0.4 = strong perceptual difference
$\Delta E(CIELAB)$	$\Delta E = \sqrt{(L_1 - L_2)^2 + (a_1 - a_2)^2 + (b_1 - b_2)^2}$	$0 \rightarrow \infty$	Perceptual color difference in CIELAB space. <1 = imperceptible 2–3 = noticeable >10 = strong color difference

Table 2. Texture metrics definition. Each metric provides complementary details on the quality and fidelity of the generated textures.

4. Experiments and results

4.1 Geometric assessment

The metric assessment of the reconstructed geometries for each method is reported in Table 3. Missing values (“-”) in Table 3 indicate cases where the reconstruction failed or could not be properly aligned and co-registered due to too large geometric discrepancies. Figure 2 shows examples of processing failures, while Figure 3 gives some visual results for successful single-image 3D generation with Tripo3D and different objects’ views for inspecting the quality of also hidden-occluded regions.

From the geometric assessment (Table 3) emerges that:

- SAM3D achieves competitive mean errors in some cases but also features high deviations and maximum absolute errors. These results indicate that the method is able to approximate the global object shape, but with relevant local deviations impeding the achievement of a uniform geometric fidelity. The best method’s performance is observed on the Bronze_1 statuette, benefiting from its compact geometry despite its reflective metallic surface. The highest standard deviation among the methods on the bowl suggests that fine details intensify reconstruction discrepancies. Very high deviations and errors are also evident on the matte surface and the regular geometry of the pottery, despite the low mean values. Another problematic case is represented by the helmet, confirming that SAM3D is not able to reliably infer occluded and complex topologies. The metope returns the highest standard deviations for this method. No valid reconstruction is obtained for the saw, as for all the other frameworks.
- Tripo3D provides the best medians across all datasets. On the two bronze statuettes, the method achieves its best results, and it surpasses the other frameworks in almost all metrics on the bowl and Bronze_2 statuette. The results suggest that the underlying TripoSR-based feed-forward architecture generalises effectively with most of the datasets. A degradation of the performance is visible on the pottery despite its rather simple and matte surface, with the worst metrics among all methods. The worst results of the entire experiments with this framework occur for the metope, suggesting some difficulties of the feed-forward architecture with quite homogeneous materials and textures. The negative mean error on the helmet indicates an issue with the method when inferring the occluded regions. A similar offset is observed on the glass.
- Trellis2 results demonstrate strong and competitive performance on geometrically complex and optically challenging objects. On the Bronze_1 statuette, Metope, and glass datasets, it records the lowest errors among the methods, suggesting that its O-Voxel representation can improve the modelling of different material objects. The explicit modelling of material properties, such as the opacity, integrated into the latent representations, seems to support the geometrically plausible inference of completely transparent objects. The lowest standard deviation across all methods is observed in most of the datasets. The lowest mean error is also recorded for the helmet, demonstrating the high capacity of handling occluded regions compared to the competing architectures. Performance is instead decreasing on some diffuse surfaces, such as the bowl.

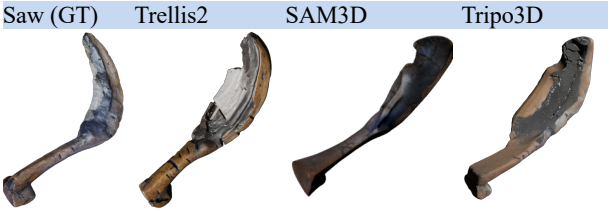

Figure 2. Failure of the three methods for the saw datasets: the reconstructed geometry differs significantly from the original.

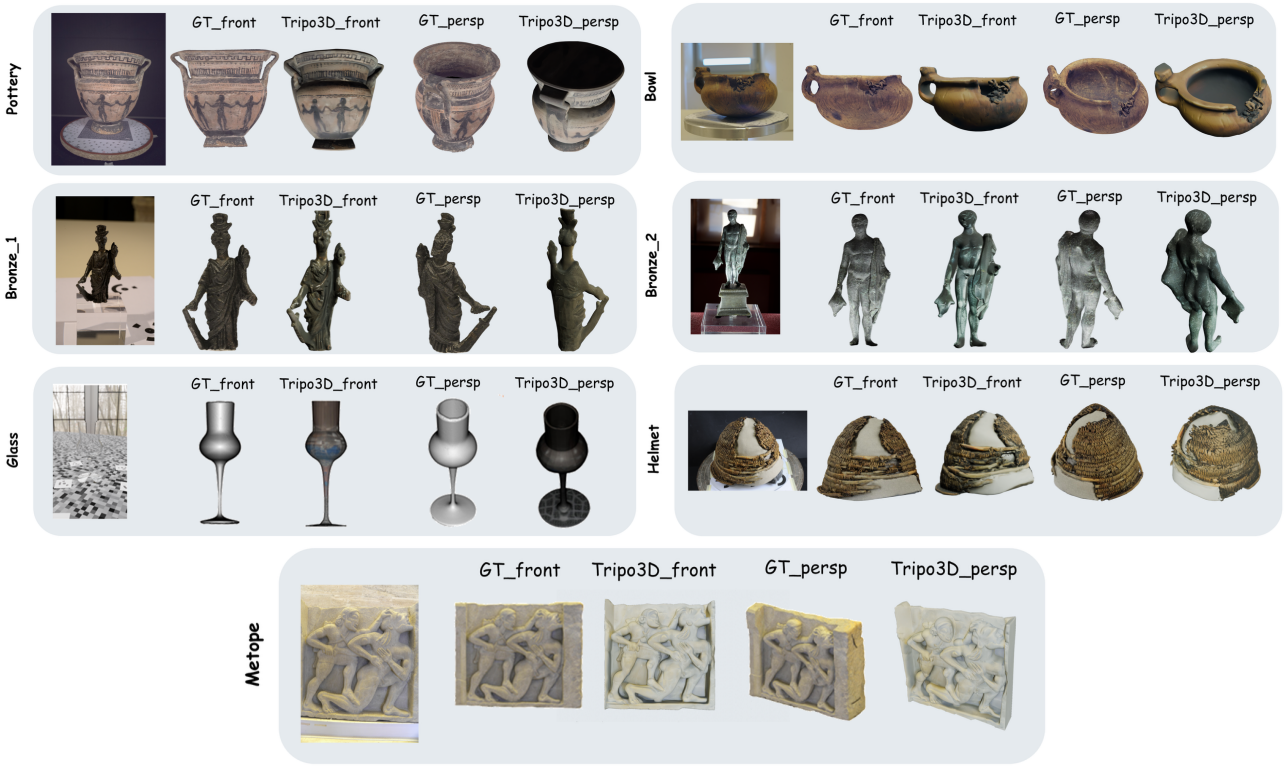

Figure 3. Different object’s view of the reconstructed geometries with Tripo3D which shows, more than the other methods, reliable capability of generating 3D geometries.

Object	SAM3D			Tripo3D			Trellis2		
	Mean	St. dev.	Max abs. dist.	Mean	St. dev.	Max abs. dist.	Mean	St. dev.	Max abs. dist.
Pottery	**0.787**	14.796	49.013	7.262	15.166	67.732	-3.979	**7.436**	**36.317**
Bowl	2.763	7.949	35.960	**1.796**	**4.502**	**22.795**	2.999	6.789	34.270
Bronze_1	0.469	1.918	9.521	0.443	1.50	5.167	**0.351**	**1.431**	**5.011**
Bronze_2	1.545	3.530	18.153	**0.503**	**2.308**	10.091	1.064	2.491	**9.618**
Helmet	7.954	12.781	44.678	-1.361	**8.314**	26.416	**0.343**	9.255	**23.726**
Saw	-	-	-	-	-	-	-	-	-
Metope	5.257	26.518	99.059	18.170	29.664	116.962	**3.196**	**14.673**	**90.499**
Glass	2.223	4.662	15.024	-0.908	4.565	23.252	**0.633**	**2.580**	**9.585**
Median	*2.223*	*7.949*	*35.96*	***0.503***	***4.565***	***23.252***	*0.633*	*6.789*	*23.726*

Table 3. Metric assessment of derived geometries against photogrammetric and synthetic ground truth.

Overall, the results indicate that no unique framework achieves uniform performance across the diverse datasets, but Tripo3D and Trellis2 show superior capabilities of delivering faithful geometries and sufficient generalization capabilities with different material datasets. Trellis2, in particular, emerges as one of the most stable and generally performs better in cases involving specular, transparent, and topologically complex objects. Tripo3D, in most cases, exhibits competitive performance and demonstrates the capability of generating reliable geometries (Figure 3). SAM3D shows high deviations, making the method less valid when geometric consistency across the entire object’s surface is required. None of the frameworks effectively handled mixed-material surfaces, highlighting an unsolved challenge for current single-image GenAI methods with datasets featuring heterogeneous radiometric properties.

4.2 Texture assessment

Texture quality results are reported in Tables 4,5,6, while visual comparisons are provided in Figure 4. Texture metrics are not reported for cases where geometric reconstruction failed to return plausible results. In particular:

- SAM3D (Table 4) shows the weakest performance among the frameworks in the texture assessment, with its overall best performance on the bowl. The high color deviation is, however, confirmed also in the other methods with this dataset. Although being the worst-performing framework on most of the pixel-based and perceptual metrics, it achieves a competitive SSIM on most of the objects, suggesting that structural information is still preserved with sufficient results in terms of luminance, contrast, and pattern similarity.
- Tripo3D (Table 5) excels for median results across almost all metrics, with surpassing results on most of the pixel-based and multi-scale structural similarity metrics. The most relevant outcome is the distinction of the framework on perceptual metrics (best LPIPS values across almost all the objects). It is worth noting that Tripo3D is the only method that produced the correct decoration on the pottery dataset, with better reconstruction quality and perceptual fidelity.

- Trellis2 (Table 6) surpasses SAM3D on several metrics, but it is still inferior to Tripo3D. Its best performance is on the metope, with overall quite balanced behavior across the datasets. It struggles mainly on perceptual metrics compared to Tripo3D, suggesting some limitations in capturing fine texture details.

Overall, Tripo3D exhibits the best performance across the datasets and among the methods, with a superior capability of preserving meaningful details, like in the pottery decoration. Trellis2 shows stable behavior with generated textures that are structurally plausible, despite being perceptually less accurate. SAM3D confirms its limitations in reproducing good-quality and visually realistic textures.

Object	SAM3D					
	MSE↓	PSNR (dB)↑	SSIM↑	MS-SSIM↑	LPIPS↓	ΔE↓
Pottery	0.029	15.36	**0.715**	0.597	0.434	11.914
Bowl	**0.057**	**12.38**	0.592	**0.560**	0.509	**19.483**
Bronze_1	0.008	20.68	0.754	0.722	0.240	4.463
Bronze_2	0.017	17.49	**0.766**	**0.749**	0.241	6.798
Helmet	0.058	12.35	**0.595**	0.582	0.412	15.730
Metope	0.101	9.92	**0.678**	0.547	0.521	20.912
Median	*0.043*	*13.87*	*0.696*	*0.589*	*0.423*	*13.822*

Table 4. SAM3D texture assessment.

Object	Tripo3D					
	MSE↓	PSNR (dB)↑	SSIM↑	MS-SSIM↑	LPIPS↓	ΔE↓
Pottery	**0.022**	16.50	0.712	**0.627**	**0.300**	10.457
Bowl	0.061	12.13	**0.577**	0.541	**0.478**	19.751
Bronze_1	0.008	20.76	**0.799**	**0.747**	**0.184**	4.230
Bronze_2	**0.013**	**18.84**	0.752	0.744	**0.199**	**5.713**
Helmet	**0.027**	**15.63**	0.572	**0.589**	**0.305**	**11.776**
Metope	**0.037**	**14.23**	0.567	0.465	0.542	17.046
Median	*0.024*	*16.065*	*0.644*	*0.608*	*0.302*	*11.111*

Table 5. Tripo3D texture assessment.

Object	Trellis2					
	MSE↓	PSNR (dB)↑	SSIM↑	MS-SSIM↑	LPIPS↓	ΔE↓
Pottery	0.023	**16.27**	**0.715**	0.613	0.329	**9.757**
Bowl	0.075	11.23	0.556	0.514	0.498	21.265
Bronze_1	**0.006**	**21.81**	0.779	0.739	0.236	4.220
Bronze_2	0.024	16.14	0.750	0.731	0.240	7.335
Helmet	0.033	14.73	0.573	0.565	0.349	13.123
Metope	0.060	12.20	0.677	**0.576**	**0.414**	**15.212**
Median	*0.028*	*15.435*	*0.696*	*0.594*	*0.339*	*11.440*

Table 6. Trellis2 texture assessment.

5. Conclusions

This work investigates the capability of three state-of-the-art and recent GenAI frameworks (SAM3D, Tripo3D, and Trellis2) to reconstruct 3D heritage artifacts from a single input image, evaluating both geometric accuracy and texture fidelity against photogrammetric and synthetic ground truth.

The results show a certain variability of the performance of the frameworks across the heterogeneous datasets, demonstrating that the reconstruction accuracy is primarily conditioned by material properties, geometric object complexity, and view-dependent ambiguity.

Overall, Tripo3D emerges as the best-performing method in terms of geometric accuracy and texture quality, with the best median values across all the metrics used in the evaluation. These results indicate stronger coherence between geometry and appearance compared to the other frameworks, despite some limitations and an increased ambiguity when handling homogeneous surfaces with limited distinctive visual features and low-contrast textures.

Trellis2 has also stable and consistent performance, particularly evident with specular, transparent, and geometrically complex shapes, demonstrating the high capabilities of its material-aware representation to deliver rather robust results. Overall, it remains slightly inferior to Tripo3D, especially in terms of perceptual texture fidelity.

SAM3D exhibits the worst performance among the evaluated datasets. It demonstrates to approximate the global shape sufficiently and preserve some structural information, but it also shows the highest variability and local inconsistencies in terms of geometry and texture representation.

All methods failed in reconstructing a dataset featuring heterogeneous materials and different reflectance properties (the saw), highlighting some limitations of all the evaluated frameworks.

As a general conclusion, analyzed single-image GenAI methods are currently not yet sufficiently accurate to meet the standards of the heritage 3D metric documentation field, despite the results of these implementations being promising as fast and effective solutions for approximate 3D reconstruction and visualization.

Acknowledgments

We acknowledge Cassa di Risparmio di Trento e Rovereto (CARITRO) for supporting the development of the ground truth models used in this work within the JUDIT project (https://judit.fbk.eu/). The presented research activities are partially developed as part of the EU projects “3D-4CH - Online competence centre in 3D for Cultural Heritage” (GA 101195149), co-funded under the Digital Europe Programme and “XRCulture - eXplore & Reuse 3D cultural heritage within the Data Space” (GA 101174317), co-funded under the Digital Europe Programme.

References

Balloni, E., Paolanti, M., Uggeri, J., Zingaretti, P., Pierdicca, R., 2025. Enhancing Cultural Heritage with Generative AI: A Comparative Framework for the Evaluation of 3D Model Accuracy and Visual Fidelity. *Proc. Digital Heritage 2025*, Eurographics Association.

Cai, J. and Lu, H., 2024. Nerf-based multi-view synthesis techniques: A survey. Proc. *Int. Wireless Communications and Mobile Computing (IWCMC)*, pp. 208-213.

Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., Lin, A., 2025. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*.

CNECT/LUX/2021/OP/0070 - Deployment of a common European data space for cultural heritage: *PM. Annual report M12. Europeana Foundation*. September 2023. URL: https://pro.europeana.eu/page/data-space-deployment. CC BY-SA.

Elhanashi, A., Essahraui, S., Dini, P., Paolini, D., Zheng, Q., Saponara, S., 2026. Generative AI and the Foundation Model Era: A Comprehensive Review. *Big Data and Cognitive Computing*, 10, 94.

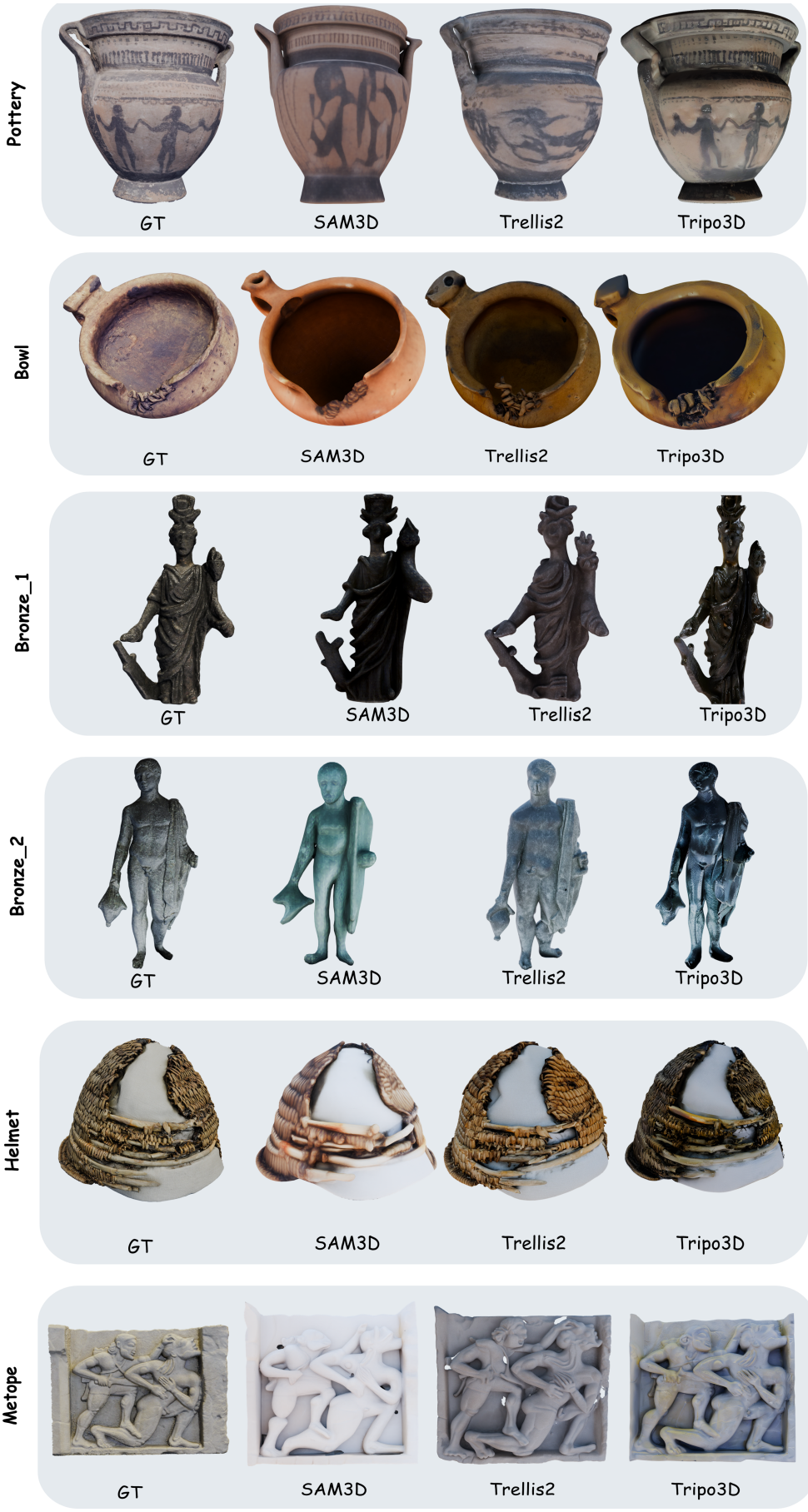

Figure 4. Rendered objects views used for the texture assessment.

https://doi.org/10.5194/isprs-archives-XLIX-B2-2026-833-2026

European Commission, 2024. European Commission: Directorate-General for Communications Networks, Content and Technology, The future of Europe’s past – Why Member States must do more to advance digitisation of our cultural heritage – Implementation of the 2021 Commission recommendation on a common European Data Space for Cultural Heritage – *Progress report 2021-2023, Publications Office of the European Union*, 2024, https://data.europa.eu/doi/10.2759/5921820.

Farella, E.M., Morelli, L., Rigon, S., Grilli, E., Remondino, F., 2022. Analysing Key Steps of the Photogrammetric Pipeline for Museum Artefacts 3D Digitisation. *Sustainability*, 14, 5740.

Fei, B., Xu, J., Zhang, R., Zhou, Q., Yang, W., He, Y., 2024. 3D gaussian splatting as new era: A survey. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 31(8), pp. 4429-4449.

He, L., Yan, H., Luo, M., Luo, K., Wang, W., Du, W., Chen, H., Yang, H., Zhang, Y., 2023. Iterative reconstruction based on latent diffusion model for sparse data reconstruction. *arXiv:2307.12070*.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems, 33*, pp.6840-6851.

Jaramillo, P. Sipiran, I., 2024. Cultural heritage 3d reconstruction with diffusion networks. Proc. *ECCV,* pp. 104-117.

Kerbl, B., Kopanas, G., Leimkuhler, T., Drettakis, G., 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. Proc. *SIGGRAPH*, 42(4).

Kingma, D.P. and Welling, M., 2014. Auto-encoding variational bayes. Proc. *ICLR*.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y. and Dollár, P., 2023. Segment anything. Proc. *ICCV*, pp. 4015-4026.

Li, Y., Zou, Z.X., Liu, Z., Wang, D., Liang, Y., Yu, Z., Liu, X., Guo, Y.C., Liang, D., Ouyang, W. and Cao, Y.P., 2025. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *arXiv preprint arXiv:2502.06608*.

Liu, J., Huang, X., Huang, T., Chen, L., Hou, Y., Tang, S., Liu, Z., Ouyang, W., Zuo, W., Jiang, J. and Liu, X., 2024a. A comprehensive survey on 3D content generation. *arXiv preprint arXiv:2402.01166*.

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

Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C., 2023. Zero-1-to-3: Zero-shot one image to 3D object. In Proc. *ICCV*, pp. 9298–9309.

Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R. and Ng, R., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM, 65(1)*, pp.99-106.

Poole, B., Jain, A., Barron, J.T., Mildenhall, B., 2023. Dreamfusion: Text-to-3d using 2D diffusion. Proc. *ICLR.*

Remondino, F., Karami, A., Yan, Z., Mazzacca, G., Rigon, S., Qin, R., 2023. A critical analysis of NeRF-based 3D reconstruction. *Remote Sensing, 15*(14), p.3585.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. Proc. *CVPR,* pp. 10684-10695.

Sengar, S.S., Hasan, A.B., Kumar, S., Carroll, F., 2025. Generative artificial intelligence: a systematic review and applications. *Multimedia Tools and Applications, 84*(21), pp.23661-23700.

Tochilkin, D., Pankratz, D., Liu, Z., Huang, Z., Letts, A., Li, Y., Liang, D., Laforte, C., Jampani, V. and Cao, Y.P., 2024. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*.

Wang, Y., Cui, H., Du, J., Di, H., Yan, L., Xu, F., Cui, K., Liu, Y., Tu, D., 2025. Single-image 3D reconstruction of painted potteries using AI diffusion and feedforward models. *Heritage Science, 13*(1), p.545.

Wu, J., Zhang, C., Xue, T., Freeman, B., Tenenbaum, J., 2016. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems, 29*.

Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J., 2025a. Structured 3D latents for scalable and versatile 3d generation. Proc. *CVPR,* pp. 21469-21480.

Xiang, J., Chen, X., Xu, S., Wang, R., Lv, Z., Deng, Y., Zhu, H., Dong, Y., Zhao, H., Yuan, N.J., Yang, J., 2025b. Native and compact structured latents for 3D generation. *arXiv preprint arXiv:2512.14692*.

Yan, Z., Padkan, N., Trybała, P., Farella, E.M. and Remondino, F., 2025. Learning-Based 3D Reconstruction Methods for Non-Collaborative Surfaces - A Metrological Evaluation. *Metrology, 5*(2), p.20.

Zhao, Z., Liu, W., Chen, X., Zeng, X., Wang, R., Cheng, P., Fu, B., Chen, T., Yu, G. and Gao, S., 2023. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. *Advances in neural information processing systems, 36*, pp.73969-73982.