npj | heritage science

Review

https://doi.org/10.1038/s40494-025-02164-1

A review of the development and application of generative technology in digital museums

Jingfan Xu[1,2], Longquan Yan[3], Ruichao Zhang[4] & Mingquan Zhou[1,3] ✉

Digital museums replicate and preserve cultural relics in digital form, protecting them from natural disasters and human damage while supporting future research. With advances in interactive generative technology, the digital display and reconstruction of artifacts have greatly improved. This paper reviews progress in AI-generated content, including text-to-image, 3D modeling, and large-scale scene generation. It further explores future applications of generative technologies in innovative cultural heritage display.

A digital museum is a museum established in digital space[1], which utilizes technologies such as the internet, virtual reality, augmented reality, and other digital tools, as well as standardized digital resources, to process, organize, and manage museum collections, scientific research, and educational dissemination. It provides a platform for disseminating knowledge related to natural or cultural heritage to various segments of the public through online access or interactive experiences[2]. This model offers personalized and precise resources and services tailored to the diverse needs of visitors[3]. Digital museums may serve as a digital extension of traditional physical museums[4] or as entirely virtual cultural spaces[5]. Their forms include virtual exhibitions, online artifact repositories, 3D scene reconstructions, and multimedia interactive experiences. As a new format for cultural heritage preservation and digital display, the digital museum has become a focal point of societal attention, with the rapid development of generative technologies providing robust technical support for the digital recreation and innovative preservation of cultural artifacts[6]. Through text-driven generative technologies, digital heritage creation, display, and interaction offer entirely new experiences, making exhibitions more vivid, realistic, and expressive in showcasing cultural heritage. Generative technologies play a crucial role in digital museum construction and cultural heritage preservation, not only in restoring damaged artifacts and historical scenes but also in creating immersive virtual exhibitions that enhance the interactivity and appeal of cultural dissemination[7]. Additionally, these technologies support data augmentation and intelligent translation, promoting the sharing and reuse of cultural resources, empowering the innovative expression and global dissemination of traditional culture, and serving as a vital technological bridge between the past and the future.

In recent years, the application of generative technologies in the field of digital museums has expanded continuously, encompassing areas such as intelligent guides, artifact restoration, virtual exhibitions, and cultural dissemination. With advancements in natural language processing and pre-trained language models, large language models based on the Transformer architecture have been incorporated into digital guide systems to enhance the comprehension of visitor inquiries and the quality of generated content. For instance, the Digital Palace Museum utilizes basic AI algorithms to provide fundamental guiding services[8]. AIGC and deep learning algorithms are now widely applied in the development of predictive and restoration models for artifacts, enabling the reconstruction and repair of damaged objects. For example, the Sichuan Provincial Institute of Cultural Relics and Archaeology and Tencent SSV Digital Culture Lab have explored human-AI collaboration in artifact restoration, focusing on the Sanxingdui site[9]. Methods such as generative adversarial networks (GANs) have enabled exhibitions to adopt more personalized and creative approaches, allowing users to engage in secondary curation and narrative expression based on digital artifacts. An example is the "Everyone is a Curator" project at the Hangzhou Museum, where users, empowered with agency, curate, extract, and recreate digital exhibits, planning cloud exhibitions that reflect their personal narrative style[10]. Furthermore, generative technologies, by integrating text, images, animations, and other media, facilitate multimodal fusion in cultural dissemination. Customizable services tailored to different audience types further enrich the ways museum content is expressed and communicated.

With continuous advancements in AIGC technology, the Diffusion Model[11]—recognized as a core image generation technique—has emerged as a focal point in digital museum research. Progressively denoising data to generate high-quality images, it demonstrates significant advantages in both fine detail generation and overall stylistic consistency. Among its applications, 3D modeling of cultural relics

[1]School of Journalism and Communication, Shaanxi Normal University, West Chang'an Street, Xi'an, 710119 Shaanxi Province, China. [2]Department of Language and Literature, Hetao College, Yunzhong Street, Bayannur, 015000 Inner Mongolia, China. [3]National and Local Joint Engineering Research Center for Cultural Heritage Digitization, Northwest University, Xuefu Street, Xi'an, 710127 Shaanxi Province, China. [4]Institute for Advanced Study in History of Science, Xuefu Street, Xi'an, 710127 Shaanxi Province, China. ✉e-mail: mqzhou@bnu.edu.cn

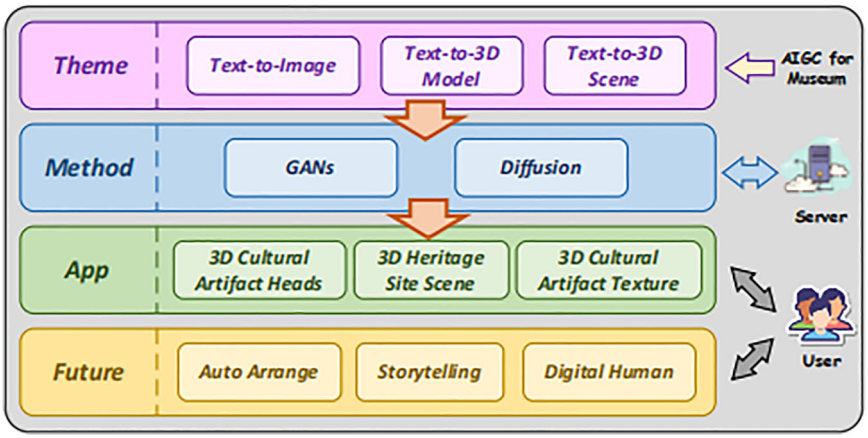

Fig. 1 | Overall framework of the generative digital museum research process.

plays a crucial role in digital museums. Although 3D generation still faces challenges such as limited data and poor generalization, integrating 2D diffusion mechanisms can enhance the fidelity and diversity of 3D outputs. Recent studies have further introduced Generative Adversarial Networks (GANs)[12] into diffusion-based models, significantly improving the realism and interactivity of virtual museums, thereby enhancing exhibition flexibility and visitor engagement. However, according to the 2021 Museum Innovation Barometer, fewer than 20% of surveyed museums worldwide have adopted generative technologies in areas such as collections, management, education, and finance[13].

Despite notable advancements in the application of generative technologies within digital museums, significant challenges remain. In the realm of virtual artifact reconstruction, most generative models rely on limited structural cues or image fragments to restore missing parts, while current acquisition technologies fail to provide sufficient physical data, making it difficult to accurately reproduce original appearances and historical styles. As a result, questions persist regarding the interpretability and historical authenticity of reconstructions[14]. Furthermore, the three-dimensional forms and decorative details of digitally generated artifacts often fail to reflect the craftsmanship and technological standards of their corresponding historical periods, leading to stylistic and temporal inconsistencies. Although such digital products may offer immersive experiences, they fall short in ensuring authenticity and rigor, thereby impeding the effective dissemination of accurate educational content and cultural meaning in virtual exhibitions. This may result in semantic drift, historical misinterpretation, and the potential fabrication of facts[15,16]. In addition, multimodal generation methods often overlook the semantic constraints and regulatory mechanisms caused by missing modality features, compromising the continuity and stability of outputs and leading to perceptible distortions during visualization[17].

In the digitization of cultural heritage, generative technologies are emerging as a pivotal force driving the intelligent and digital transformation of various functions within digital museums[18]. Leveraging pre-trained algorithms and large model architectures, generative models have enhanced the capacity of GANs and diffusion models to integrate traditional cultural content and interpret complex scenarios. By synthesizing incomplete spatial, textural, and semantic information, these models generate restoration proposals that improve accuracy and enable visual representation. Moreover, in the currently underdeveloped areas of guided tours and narrative construction, large models' language comprehension capabilities facilitate more natural and personalized human-computer interaction[19], enabling customized virtual guide services and online exhibition experiences[20]. Finally, through large-scale multimodal prediction, these technologies enrich the representation of cultural artifacts across formats—such as textual descriptions, video demonstrations, and 3D models—thereby supporting the construction of multimodal narratives in cultural communication.

As illustrated in Fig. 1, this paper adopts a research framework centered on the empowerment of digital museums through generative artificial intelligence (AIGC), encompassing four key dimensions: research themes, technical methodologies, application scenarios, and future directions. Based on this framework, the paper is structured as follows:

Section "Generative approaches for digital museums" focuses on three predominant application themes—text-to-image, text-to-3D model, and text-to-3D scene—tracing the evolution of relevant research, background context, technological development paths, and representative applications within cultural heritage settings. Text-to-Image guides the generation of planar visual perceptions of artifacts within a two-dimensional color space, enriching digital displays for online museums. Text-to-3D model enables the geometric, structural, textural, and semantic-consistent reconstruction and restoration of cultural artifacts, offering enhanced restoration accuracy and interactive experiences. Text-to-3D scene facilitates the recreation of spatial layouts and historical contexts, supporting the development of immersive environments and large-scale digital museums. Section "Research on generative models" investigates the application of mainstream generative methods—generative adversarial networks (GANs) and diffusion models—in digital museums, analyzing their development trajectories and conducting a detailed comparative analysis. It further discusses the strengths and limitations of these models in terms of image and 3D generation authenticity, semantic consistency, and text-guidance performance, while outlining their adaptability and future prospects based on recent advancements. Section "Generative museum applications" explores three major application scenarios of these technologies in digital museums: facial reconstruction of historical figures or lost artifacts using 3D portrait generation; visualization of large-scale heritage sites, architecture, and historical spaces via 3D scene reconstruction; and the restoration of missing artifact textures through text-guided voxel generation and texture mapping, thereby expanding virtual artifact databases. Section "Generative technologies empowering the future of museums" presents forward-looking concepts for future applications, including automated artifact curation, narrative integration, and personalized virtual guide systems. Finally, the section "Conclusions" offers a comprehensive review and conclusion of the study.

In summary, this paper reviews the current state of AIGC technologies and digital museum development, constructing a research framework and outlining application scenarios for a "generative digital museum," with the aim of providing theoretical foundations, methodological guidance, and scenario-based insights for future research.

Generative approaches for digital museums

Text-to-artifact image generation

Text-to-image (Text-to-image, T2I) refers to the generation of two-dimensional images that accurately reflect textual descriptions through advanced computational techniques[21]. This technology demonstrates significant potential in the digital restoration and innovative preservation of cultural artifacts. By producing visually authentic images based on textual input, it opens new avenues for the digital creation[22], display, and interaction of cultural heritage. T2I not only brings greater vividness to the digital presentation of traditionally static artifacts but also enables more expressive and artistic representations that convey deeper cultural meanings. Particularly in online virtual exhibitions within digital museums, the application of this technology greatly enhances the visitor experience, immersing audiences in highly realistic historical and cultural environments that vividly evoke the essence of the past.

Deng et al.[23] proposed a novel approach for artifact image generation based on a low-rank adaptive diffusion model. By integrating low-rank adapter (LoRA) techniques into stable diffusion and fine-tuning on historical cultural heritage images, the method enables image generation and style transfer specific to Chinese cultural artifacts. The model significantly enhances the cultural attributes and historical texture of generated images while maintaining the expressive power of the original diffusion framework. Moreover, by leveraging low-rank approximation, it adaptively controls model capacity, enabling more efficient and lightweight fine-tuning. Hsieh et al.[24] introduced a state-of-the-art text-image matching technique for cultural heritage digitization. Employing web crawlers to automatically collect artifact resources, their method uses text-driven image retrieval and generation to promote the digital reproduction of cultural artworks. The study also addresses technological practices in digital reconstruction, online presentation, and artifact preservation, while emphasizing the importance of data ethics, privacy protection, and fairness. Liu et al.[25] conducted a focused analysis of the performance variability of current Text-to-Image (T2I) models across diverse cultural contexts. They proposed and constructed the C^3 (challenging cross-cultural) benchmark, comprising a large set of challenging Chinese cultural prompts and evaluation criteria to assess generative models' capabilities in responding to culturally specific inputs. A multimodal filtering metric was introduced, integrating text-text, image-text, and object-text alignment to select high-quality, culturally relevant training samples for model fine-tuning. Models fine-tuned using this filtered data outperformed baseline and comparative models in terms of cultural appropriateness, object completeness, and overall esthetic quality, as demonstrated in the C^3 benchmark evaluation.

In future research, we will further enhance the generation quality and efficiency of diffusion models to meet the demands of large-scale artifact image generation. Furthermore, integrating artificial intelligence technologies from multiple domains may diversify digital representation methods for cultural artifacts, thereby providing comprehensive technical support for artifact preservation and innovative development.

Text-to-3D artifact model generation

The rapid advancement of text-to-image diffusion models has propelled progress in text-to-3D generation technology[26,27], introducing novel possibilities for 3D artifact model generation in digital museums. This technology enables the generation of highly detailed and complex 3D models based on textual descriptions, offering unprecedented technical pathways for digital preservation and exhibition of cultural artifacts. Nevertheless, the field of 3D model generation continues to face significant challenges, primarily including the scarcity of diverse 3D datasets and limitations in generated object categories[28]. These constraints impede the generalization capabilities of generative models, making it difficult to encompass the diverse and structurally complex nature of artifact objects[29].

Existing research demonstrates that DreamFusion[30] substantially enhances 3D generation performance by leveraging image distribution scores from pre-trained 2D diffusion models, thereby enabling the generation of high-fidelity and diverse 3D artifact models. Current research focuses on improving the fidelity and stability of generative models to expand their expressiveness across diverse scenarios, providing robust technical support for digital museum exhibitions, education, and dissemination. Nevertheless, the 2D-to-3D generation process remains subject to inherent uncertainties. To address this issue and ensure generative model coherence and consistency, certain studies have incorporated 3D guidance mechanisms combined with large-scale 2D diffusion models[31] to optimize generation outcomes and mitigate potential model biases. In the optimization process of 3D generation based on 2D diffusion models, techniques such as textual inversion, CLIP loss enhancement[32], and LoRA[33] fine-tuning have been extensively employed. These methods facilitate better alignment between textual descriptions and generated results, improving generation accuracy and expressiveness. Furthermore, the emergence of 3D diffusion models such as Point-E[34] and Shape-E[35] has made the generation of complex structures more feasible, providing substantial support for artifact 3D modeling and accommodating the modeling requirements of diverse artifact types. By integrating these technologies with digital museums, digital modeling and multidimensional exhibition of artifacts can be achieved, significantly enhancing cultural heritage preservation standards while enriching public interaction and experience.

In the field of artifact digital modeling, text-to-3D generation technology demonstrates considerable application potential. Leveraging digital museum technologies, this innovative generative approach can substantially elevate digital preservation standards for cultural artifacts[36] while delivering enriched, authentic, and immersive cultural experiences to the public. Future research directions will focus on further enhancing generative model authenticity, improving stability, and expanding application scope to more effectively address various challenges encountered in the 2D-to-3D generation process, thereby providing expanded possibilities and opportunities for cultural heritage preservation, intelligent museum construction, and public cultural dissemination.

Text-to-3D large-scale museum scene generation

In digital museum construction, 3D large-scale scene generation plays a pivotal role[37], serving as both the carrier for artifact exhibition and the virtual space for visitor exploration. Currently, 3D large-scale scene generation faces numerous challenges, primarily including the difficulty of acquiring high-quality 3D data and the complexity of the generation process.

In virtual museum scene generation, methods such as GAN and InfiniCity[38] demonstrate the potential of integrating 2D and 3D models, particularly excelling in complex scene modeling. By employing multi-dimensional modeling, these approaches enable the creation of richer and more refined virtual spatial models. Meanwhile, diffusion models, as an emerging generative technology, further enhance the realism of museum scenes by jointly predicting depth information[39], resulting in more nuanced and lifelike 3D visual effects. This technology extends beyond image generation, advancing the development and application of virtual museum technologies.

Layout generation techniques based on Transformer[40] architectures, such as LayoutTransformer[41] and ATISS[42], offer diverse solutions for digital museum exhibition design. These methods automatically generate optimal layouts tailored to varying exhibition needs and spatial characteristics, enhancing the flexibility and efficiency of museum hall planning. They also serve as practical tools for curators, improving the visual impact and interactivity of exhibitions[43]. In panoramic image generation, models like COCO-GAN[44] and InfinityGAN[45] capture the spatial relationships among exhibits, providing visitors with a more immersive and intuitive virtual tour experience. Although current approaches still face challenges in object-level editability and multi-view consistency, continued technological advancements are expected to address these issues, further improving the presentation quality of digital museums.

Overall, the integration and advancement of these technologies provide multidimensional support for large-scale 3D scene generation in virtual museums, significantly enhancing the realism, interactivity, and immersive quality of virtual exhibitions. In the future, digital museums will be able to

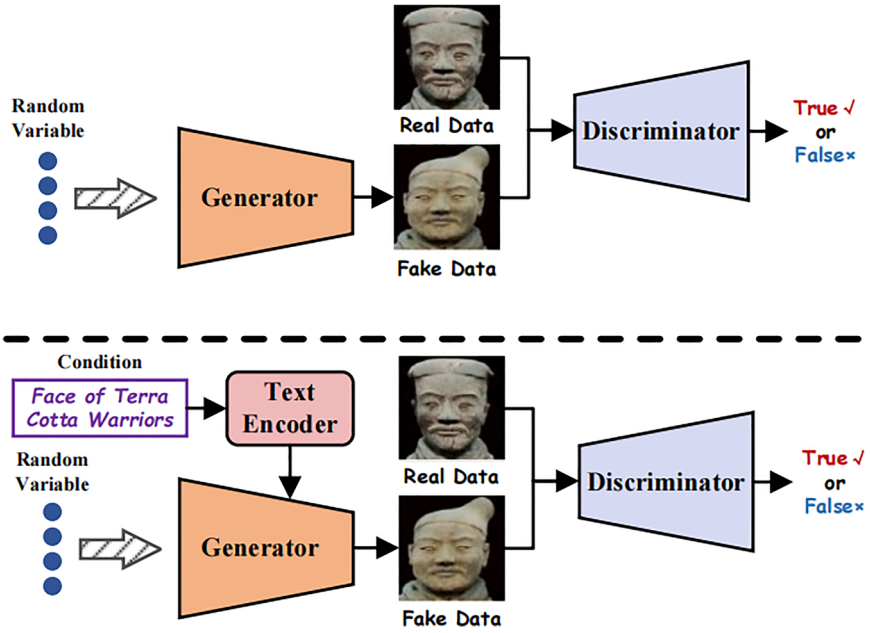

Fig. 2 | Generative adversarial networks and conditional generative adversarial networks.

more accurately reconstruct large-scale museum environments and cultural artifacts, offering audiences richer interactive and educational experiences.

Research on generative models

Generative adversarial networks (GANs)

Generative adversarial networks (GANs) are a pivotal technique in image synthesis and currently represent the most widely used approach for text-to-image generation tasks[46]. A GAN consists of two core components: a generator and a discriminator, which engage in an adversarial process. The generator attempts to produce realistic data samples, while the discriminator seeks to distinguish between generated and real samples. As illustrated in Fig. 2, the generator transforms textual features into images, while the discriminator evaluates both the authenticity and semantic consistency of the generated images. Through iterative training, the generator progressively improves its ability to produce realistic images, while the discriminator enhances its accuracy, collectively resulting in a significant improvement in image generation quality.

Early research adopted Reed et al.'s model as the backbone, supported by deep convolutional generative adversarial networks, and differentiated through constraints into conditional generative adversarial networks (CGAN) and unconditional generative adversarial networks (GAN)[47], as illustrated in Fig. 2. The primary distinction between these approaches lies in CGAN's incorporation of textual features to constrain both the generator and discriminator, enabling the model to extract semantic information from text during training and generate images aligned with textual descriptions, thereby establishing the foundation for subsequent text-to-image generation research. Correspondingly, unconditional generative models such as CycleGAN and DCGAN, lacking conditional encoders, exhibit isotropic randomness in generation results, which impedes control and integration of generated outcomes.

With the introduction of multiple GAN models, including StackGAN[48], AttnGAN[49], and LAFITE[50], increasingly sophisticated generation mechanisms and techniques have been incorporated to further enhance image quality and text-alignment effectiveness. StackGAN[47] employs a multi-stage generation mechanism, progressively transforming low-resolution images into high-resolution counterparts, thereby continuously enriching image detail. AttnGAN[49] incorporates attention mechanisms to better capture crucial details from text and reflect them in generated images. LAFITE[51] integrates CLIP models, further improving semantic consistency between generated images and textual descriptions. DF-GAN enhances semantic consistency between text and images through a matching-aware gradient penalty, strengthening qualitative guidance of text on generation outcomes[52], enabling rapid image generation. LEGAN addresses inter-class content imbalance through local outlier factor and information entropy, employing joint training with augmented sparse and dense samples[53], typically applied to data augmentation tasks in medical imaging[54], establishing balanced inter-class data relationships. MARS leverages large language models and integrated expert modules to decouple and comprehend data in textual, visual, and video formats[55], achieving higher training efficiency and precise, realistic multimodal content generation. Table 1 presents a detailed comparison of the aforementioned GAN models. Through adversarial training, GANs have significantly enhanced both the quality and semantic consistency of generated images in text-to-image tasks. By integrating these technologies, image resolution and detail have been substantially improved, achieving greater realism while enhancing textual description consistency. The development of generative adversarial networks has established a solid foundation for future research, and with continued technological advancement and refinement, the quality of generated images and textual consistency will further improve.

Diffusion models

Diffusion model[11] encompasses two core processes: the forward diffusion process (FDP)[56] and the reverse diffusion process (RDP)[57]. The forward diffusion process constitutes a "blurring" procedure wherein data progressively loses its original characteristics through a parameterized Markov chain that gradually adds noise to the original data until it ultimately becomes pure noise. This generative process demonstrates strong interpretability and enables diverse sample generation. As illustrated in Fig. 3, diffusion models are generally categorized into denoising diffusion probabilistic models (DDPM)[58] and conditional diffusion probabilistic models[59]. Denoising diffusion probabilistic models learn posterior distributions to progressively denoise and restore data, handling both continuous and discrete data. In continuous data spaces, models approximate posterior probabilities to eliminate noise; in discrete spaces, models utilize state transition matrices to capture inter-data dependencies, thereby facilitating data restoration. Conditional diffusion models integrate conditional information into the generation process, achieving conditional generation. These models typically employ three guidance strategies: explicit classifier guidance, implicit classifier guidance, and CLIP-based multimodal guidance. Explicit classifier guidance requires additional classifier training to provide

Table 1 | Development history and method comparison of GANs-related research

Methods	Year	Mode	Advantages	Disadvantages	Application
GAN	2014	Image generation	This method builds a generator network and a discriminator network that supervise each other through a competitive learning process.	The training process is unstable and prone to mode collapse.	Data augmentation, image restoration, and super-resolution reconstruction.
DCGAN	2016	Image generation	Improves the quality and performance of image generation by building an adversarial generation process through convolutional neural networks.	It is suitable for relatively simple image generation tasks and generates images with lower resolution.	Low-resolution image generation.
CycleGAN	2020	Image generation	It does not require paired data and ensures the consistency of images before and after transformation through cycle consistency.	The generated images may have some distortion, especially in complex image transformation tasks.	Unsupervised image transformation tasks such as style transfer and domain adaptation.
CGAN	2014	Text-to-image	Introducing additional conditional information as guidance in GAN to achieve the generation process of text-to-image.	The training process is unstable, prone to mode collapse, and relies on manual hyperparameter adjustment.	Image synthesis, image transformation, and 3D model generation.
StackGAN	2016	Text-to-image	A two-stage conditional generation pipeline is constructed.	The model has low fault tolerance and relies on the results of the first stage, so it is prone to distortion and blur.	Standard resolution text-to-image generation, style transfer, and image super-resolution reconstruction.
AttnGAN	2017	Text-to-image	Focus on keywords in a given description and perform fine-grained reconstruction of different sub-regions.	High training complexity and slow convergence; attention focus in sub-regions is prone to imbalance.	High-quality, high-resolution image generation, and multimodal generation.
LAFITE	2022	Text-to-image	Language-independent text-to-image generation is achieved, reducing dependence on large-scale image and text datasets.	It relies on CLIP pre-training, has weak generalization ability in specific fields, and small objects are easily deformed.	Zero-shot image conversion and generation, video generation.
DF-GAN	2022	Text-to-image	Semantic consistency between text and image is enhanced by a matching-aware gradient penalty.	The generation process is too simplified, which limits the diversity of generated images.	Fast image generation supports artistic creation.
LEGAN	2024	Class to image	Solve the problem of content imbalance between classes through local anomaly factors and information entropy.	Applied to t specific fields, the generalization ability is poor.	Establish a balanced relationship between data categories.
MARS	2025	Text-to-image	Use the semantic-visual language integrated expert module to decouple text and image processing.	Limited by the generalization ability of the underlying language model.	Fast text-to-image joint generation.

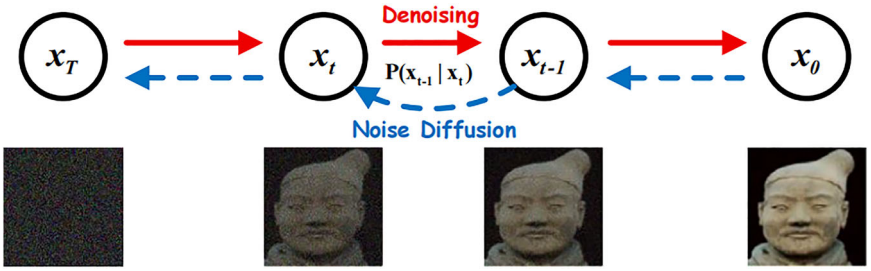

Fig. 3 | Diffusion model working principle diagram.

Table 2 | The development history and method comparison of diffusion model-related research

Methods	Year	Mode	Advantages	Disadvantages	Application
DDPM	2020	Image generation	The pioneering work of diffusion models, establishing the foundational framework for diffusion-based generation.	The generated quality is poor, and the unconditional guidance leads to uncontrollable generation outcomes.	Simple image generation, low-quality image generation.
Imagen	2022	Text-to-image	Extremely high image quality; strong natural language understanding capabilities.	Large training data, with relatively low training efficiency.	High-precision, high-quality content generation.
Stable Diffusion	2022	Text-to-image	Through a less effective attention mechanism, it supports controlling the image generation process using text, bounding boxes, and other inputs.	It requires substantial hardware resources and extensive prior knowledge tuning.	Applicable to tasks such as image restoration, image generation, and image editing.
GLIDE	2022	Text-to-image	Achieving stronger text comprehension, generating highly realistic images with strong semantic alignment.	Excessive reliance on guidance strategies leads to a loss of generation diversity, with a large model parameter size.	Text-to-image generation, text-driven image restoration.
DALLE-2	2022	Text-to-image	Supports image editing, transformation, and generation, with more accurate semantic consistency.	Relies on complex pipeline structures, resulting in high training costs.	High-fidelity image generation, artistic creation.
Muse	2023	Text-to-image	High image quality, more stable training.	Low inference efficiency.	High-quality image generation.

conditional constraints, while implicit classifier guidance achieves guidance directly through model parameters without additional training, resulting in lower computational costs. CLIP-based guidance significantly enhances text-image alignment through multimodal alignment but incurs higher computational costs. The advantage of implicit classifier guidance lies in its elimination of additional training requirements, making it applicable to multiple generative models, including stable diffusion[60], GLIDE[61], and DALLE-2[62], with its lower computational costs establishing it as a widely adopted guidance strategy. Conversely, while CLIP-based guidance demonstrates exceptional performance in text-image matching[63], its substantial computational costs limit its broader application potential. Diffusion models generate data through progressive denoising, effectively adapting to both continuous and discrete data denoising mechanisms, demonstrating excellence across various data types. Conditional diffusion models achieve conditional generation through multiple guidance strategies[64], particularly implicit classifier guidance[65], which offers wide applicability and cost-effectiveness due to its elimination of additional training requirements. In contrast, CLIP-based guidance significantly improves generation quality and text-image alignment[66] but incurs higher computational costs. Table 2 summarizes the technical comparisons and application scenarios of the aforementioned diffusion models. Future research directions may focus on improving generation quality while reducing the computational costs of CLIP-based guidance strategies, thereby promoting broader application of diffusion models.

Generative museum applications

Artifact portrait generation

Three-dimensional cultural artifact portrait generation represents a pivotal research domain within cultural heritage digitization and virtual preservation. By leveraging computer vision and generative models, this approach enables efficient restoration or creation of historically significant artifact portraits[67], supporting applications in artifact restoration, exhibition, and digital museums. The skinned multi-person linear model (SMPL)[68] constitutes a widely adopted skeletal-driven parametric human body model for 3D human model generation. This model decomposes the human body into morphological and postural components, facilitating flexible 3D mesh deformation to precisely generate diverse human models. It supports gender-specific template meshes, enabling distinct gender differentiation in male and female model generation, demonstrating substantial application value in virtual reality, animation, and game development domains. In early 3D character generation research, scholars attempted to integrate text-based 2D prior information with neural field generation techniques[69], exploring methodologies for generating complex 3D characters from textual descriptions. These investigations established foundational frameworks for subsequent technological innovations, concentrating on algorithmic automation of character morphology, posture, and motion attribute generation while utilizing image generation techniques to enhance output quality. Recent years have witnessed rapid advancement in generative model technologies, yielding novel approaches such as DreamAvatar[70] and AvatarCraft[71], which combine diffusion models with SMPL to further enhance 3D avatar generation precision and flexibility. These methodologies not only enable precise control over character shape generation but also facilitate personalized design of character identity and style, rendering generated 3D characters increasingly diverse and distinctive. In motion synthesis and editing domains, approaches such as MotionCLIP[72] and AvatarCLIP[73] utilize 3D motion autoencoders and CLIP to achieve text-driven animation generation. These methodologies enable automatic synthesis or editing of character motion, providing more efficient animation generation paradigms.

In the field of 3D facial generation, models such as DreamFace[74] have employed CLIP and neural rendering techniques to achieve text-driven 3D facial design from scratch[75]. This technology enables users not only to generate high-quality facial images but also to personalize and adjust them based on individual facial features. It lays a robust foundation for character creation in gaming, avatar design in virtual social settings, and interactive roles within virtual reality environments. As illustrated in Fig. 4, the 3D cultural heritage avatars reconstructed using this technology demonstrate the effective transformation of textual descriptions into high-fidelity

Fig. 4 | Results of text-driven 3D cultural relic portrait generation using AIGC.

three-dimensional forms with detailed facial features. These outputs exhibit a remarkable degree of resemblance to actual artifacts and maintain stylistic consistency with the original cultural objects.

With ongoing advancements in multimodal-guided techniques and neural representation methods, text-driven 3D generation has become increasingly flexible, efficient, and precise across various applications, including virtual reality, animation, and gaming. These innovations not only accelerate the creation of 3D virtual characters but also expand the creative potential of text-based input, demonstrating significant promise, particularly in virtual reality and animated media. Looking ahead, continued technological progress is expected to enable the development of more personalized, content-rich, and interactive digital characters.

3D heritage site scene generation

3D scene reconstruction involves generating a three-dimensional model of a scene from multiple viewpoint images[76]. In the context of digital museum development, heritage site scene generation represents a key direction in digital cultural heritage preservation[77]. Leveraging modern computer graphics and artificial intelligence, it enables the restoration of historical sites, supporting conservation, virtual exhibition, and cultural tourism industries. Large-scale 3D scene reconstruction is computationally intensive, typically requiring the processing of numerous images and multi-view data. Effective model generation demands highly efficient methods capable of rapidly converting this data to ensure the timely reconstruction of expansive environments[78]. Another major challenge lies in achieving a high degree of realism and coherence across the scene. Maintaining consistency in details and lighting from varying perspectives remains a significant technical obstacle. To address these issues, scene modeling often relies on the manual work of skilled 3D designers or the application of sophisticated computational techniques for post-processing. While such approaches ensure high quality and consistency, their heavy dependence on manual effort and computational resources poses challenges for large-scale deployment.

To address the challenges mentioned above, recent years have seen the development of several emerging methods, yielding significant advancements. Text2Room[29] and Text2NeRF[79] are innovative approaches that combine text with 3D scene generation, enabling the creation of corresponding environmental layouts from natural language descriptions. Text2Room focuses on structured modeling of environments, emphasizing the inference of spatial distribution and relationships between elements from textual descriptions. In contrast, Text2NeRF utilizes neural radiance fields (NeRF) technology[79], concentrating on multi-view scene modeling to ensure consistency in lighting and details from different perspectives, enhancing the realism of the generated scenes. These technologies are particularly well-suited for rapidly creating panoramic scenes and have opened new avenues for large-scale scene reconstruction.

Additionally, Roomdreaming[80] and Ctrl-Room[81] integrate geometry-guided and layout learning techniques to significantly improve the visual coherence and editability of 3D scenes. These tools leverage learned geometric data to guide the generation process, ensuring harmony in both structure and appearance while facilitating subsequent adjustments. These methods not only accelerate scene generation but also allow even non-experts to create personalized scenes through simple modifications. From an editing perspective, FastScene[82] further enhances the speed of scene generation. Using advanced algorithms and efficient data processing, it can create high-quality 3D scenes in relatively short timeframes, providing substantial practical value for applications that require instant content generation. This method reduces computational costs and optimizes the generation process, facilitating large-scale and multi-scene production. In the study of 3D scene generation, various perspective strategies have gradually demonstrated their unique advantages. External perspective strategies, exemplified by Text2Room and Text2NeRF, aim to construct a comprehensive environment from a macro perspective. Another category of methods, represented by MAV3D[83] and GALA3D[84], takes an object-centered approach, focusing on the object itself and its interactions with the surrounding environment. This approach is particularly suitable for creating scenes involving complex interactions between objects and their surroundings, such as character-object interactions. Meanwhile, persistent view methods like SceneScape[85] and VividDream[86] are ideal for dynamic and interactive scenarios, especially in open-world scene construction for applications such as gaming and virtual reality. Furthermore, object composition-based approaches, such as CompoNeRF[87] and set-the-scene[88], focus on the combination and arrangement of multiple objects. These techniques are frequently used in the creation of scenes where multiple elements coexist and influence one another, simulating complex interactions and combinations between objects, which is especially useful for dynamic and intricate scene creation. Currently, most NeRF methods concentrate on multi-view static scene reconstruction, achieving high-quality representations through volumetric rendering. They provide fine-grained 3D representations and high-fidelity synthesis of complex lighting, materials, and details. Figure 5 illustrates the general architecture for the reconstruction of ancient architecture using NeRF[89], which includes steps such as image capture, sparse point cloud reconstruction, NeRF modeling, and multi-view synthesis. Multi-view images are first captured with camera poses, followed by volumetric representation learning within the NeRF network. This process renders the geometry, texture, and mesh details of ancient architecture in 3D space. By aligning and merging with sparse point clouds, it enables the precise restoration of the complex structures, intricate decorations, and realistic lighting effects of historical buildings.

Overall, current 3D scene generation technologies have seen significant improvements in speed, quality, and editability. By adopting various generation strategies and optimizing perspectives, these methods can produce more efficient and realistic results across a range of application scenarios. With ongoing technological advancements, it is likely that these methods will increasingly integrate deep learning and generative adversarial network (GAN) techniques, further enhancing the quality and flexibility of 3D scene generation in the future.

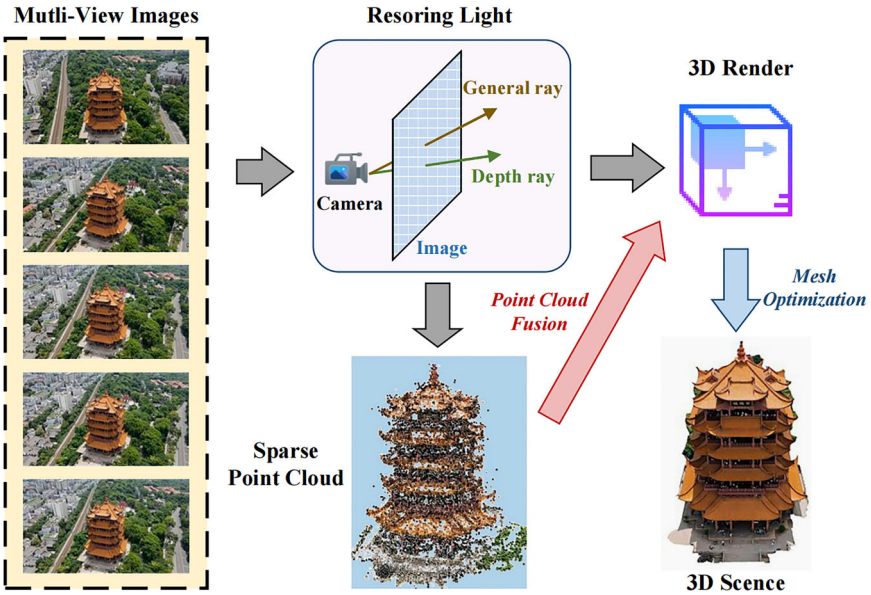

Fig. 5 | The process of ancient building scene generation based on NeRf.

3D cultural relic texture generation

3D cultural relic texture generation is a crucial component of cultural heritage digitization, enabling the realistic reproduction of artifact appearances through texture generation and optimization[83]. This process provides essential technical support for digital exhibition, restoration, and preservation. Although recent advancements in text-guided 3D texture generation have yielded notable progress, the technology still faces significant challenges—particularly in achieving high-quality and consistent texture outputs. A central issue lies in aligning textual descriptions with the physical morphology of artifact surfaces. To produce textures that meet practical demands, it is necessary to possess a deep understanding of complex geometries and ensure that the generated textures are coherent with both the surface shapes and material properties. Moreover, in the automated design of richly detailed patterns, current technologies often fall short of replicating the esthetic perception and creative judgment of human designers, limiting their broader applicability. Current research primarily focuses on automating texture design through natural language prompts, aiming to generate textures that are both semantically aligned and visually accurate. These approaches rely on deep learning models to translate textual descriptions into corresponding 3D texture images. However, ensuring smooth transitions, structural coherence, and the avoidance of visual artifacts during texture generation remains a persistent technical challenge.

To address these challenges, researchers have proposed a range of innovative techniques and methods, including TANGO[90], TexFusion[91], TEXTure[92], Text2Tex[93], and X-Mesh[94], all of which play critical roles in enhancing the quality of texture generation. TANGO improves alignment between textual input and geometric models, thereby increasing the accuracy of texture generation. TexFusion and Text2Tex leverage deep learning techniques to enhance the rendering of fine details and ensure visual coherence. X-Mesh focuses on generating more complex textures and morphological variations[95]. While these approaches have achieved notable results in certain domains, challenges remain in generalizing their effectiveness across broader applications[96].

Thanks to the continued integration of advanced technologies such as deep learning models, generative adversarial networks (GANs), and diffusion models, 3D texture generation has seen notable improvements in both accuracy and efficiency. Deep neural networks, trained on large-scale datasets, can produce highly realistic and detailed texture maps, significantly reducing the manual workload traditionally required in texture design. Figure 6 illustrates the outcomes of various methods on 3D models and their corresponding texture maps. Panels (a–d) display results generated using GAN-based techniques, which show moderate success in reconstructing object structures and capturing semantic texture features. However, these methods often struggle with texture fidelity, geometric consistency, and continuity—resulting in issues such as discontinuities, blurring, or repetitive patterns, especially on complex surfaces or in high-frequency regions. Such limitations can lead to unrealistic and incoherent texture applications. In contrast, panels (e, f) demonstrate the advantages of diffusion-based methods, which achieve superior performance in texture reconstruction accuracy and stylistic consistency. These models produce smoother, more natural texture transitions and effectively align with 3D geometries, enhancing overall visual realism and perceptual quality. Nevertheless, fully automating the texture generation process still requires further optimization of existing techniques to improve their robustness and adaptability across diverse and complex scenarios. Future research is likely to focus on multimodal learning, model scalability, and diversity in texture synthesis, paving the way for broader applications in cultural heritage preservation, virtual reality, and beyond.

Generative technologies empowering the future of museums

Looking ahead, generative technologies are poised to exert a profound influence on digital museums, opening new avenues for the preservation and presentation of cultural heritage. In this new era, it is essential to fully recognize the application value of these emerging technologies and harness their potential to advance the intelligent development of China's museum sector. First, the multimodal generative capabilities of these technologies will allow cultural digitization to transcend traditional two-dimensional and static displays. Through high-precision 3D modeling and dynamic scene construction, audiences will be able to explore historical cultures more immersively, fostering deeper cultural understanding and exchange. Second, interaction design based on virtual experiences is expected to become a primary mode for digital museum exhibits. By leveraging virtual reality (VR), augmented reality (AR), and related technologies, visitors will be able to intuitively perceive the rich historical and cultural information embedded within artifacts. This shift from "static display" to "dynamic interaction" not only stimulates public interest in culture but also deepens their historical awareness and engagement. Moreover, as artificial intelligence continues to be widely adopted, museums face challenges such as insufficient digital resources and slow data updates. Within the framework of generative technology empowerment, museums—as key venues for cultural transmission and exhibition—still hold vast potential for future development.

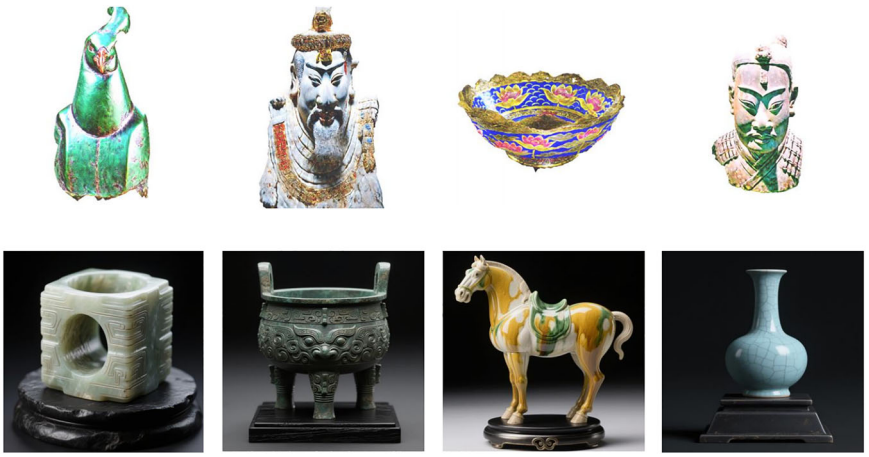

Fig. 6 | 3D cultural relic model and corresponding texture generation.

Automated display of virtual artifacts

Time and space are fundamental dimensions of movable cultural relics and represent key expressions of real-world data. In digital museums, the chronological sequencing of artifacts is emerging as a promising direction for both curation and education. By arranging artifacts according to their historical periods, cultural phases, or significant events, museums can implement more scientifically grounded and pedagogically valuable timelines. From the perspective of historical research, such temporal sequencing offers a clearer representation of the trajectory of cultural development. For instance, when showcasing ancient civilizations, artifacts may be arranged in a progression—from the earliest handmade pottery, to the complex casting of bronze ware, and eventually to the refined craftsmanship of porcelain. This approach not only supports deeper academic inquiry but also helps the public better understand the stages and milestones of cultural evolution. Extensive analyses have been conducted on the metadata of numerous artifacts, including their excavation dates, cultural contexts, and manufacturing techniques. With the advancement of digital humanities, various platforms and systems have been developed to visualize and analyze the spatiotemporal evolution of cultural heritage. Notable examples include the Chronological Map of Tang and Song Literature developed by South-Central Minzu University and the China Historical Geographic Information System (CHGIS), jointly developed by Fudan University and Harvard University. These tools facilitate a dynamic understanding of cultural change across time and space, enriching both scholarly research and public engagement in digital heritage.

Giving artifacts stories

The importance of narrative has long been self-evident, and through generative technologies, artifacts are imbued with stories, bringing them to life. In future museums, large models will leverage their exceptional linguistic capabilities to craft rich, compelling narratives based on an artifact's uniqueness, historical context, and related cultural stories. For the audience, these stories transform artifacts from mere static exhibits into vibrant, culturally rich carriers of meaning. For instance, a piece of ancient bronze can be brought to life through generative technology, offering detailed descriptions of its manufacturing process, its use in ancient society, the ritualistic culture behind it, and the legendary stories it has carried through the ages. These narratives not only captivate the audience's attention but also deepen their insights and memories regarding the artifact and its historical significance. From the perspective of cultural heritage, the storytelling capabilities of generative technologies help museums transcend the limitations of time and space in cultural dissemination. Through platforms like the internet, these stories can be shared globally, attracting a wider audience to engage with and explore the museum's cultural offerings. Furthermore, with ongoing advancements in generative technologies, the content of these stories will become increasingly rich and accurate, positioning storytelling as a key driver of both cultural preservation and innovation.

Virtual digital guide tour

Traditional museum tours typically rely on live explanations from professional guides. However, through the use of virtual digital avatars, museums can offer an interactive, personalized visiting experience, enabling visitors to engage in self-directed learning and exploration. The generation and application of voice-driven digital avatars has become a prominent area of research. Creating a virtual digital guide requires the use of modeling and rendering technologies to produce avatars with realistic appearances and natural movements[97]. Skeletal animation systems are employed to precisely control the avatar's body movements, ensuring that its behavior during the tour aligns with natural human actions. Additionally, multimodal interaction technologies based on deep learning enable interaction between the virtual guide and the audience. Through voice, gesture, and facial expression recognition, the digital avatar can accurately perceive the visitor's intentions and respond accordingly.

To interact effectively with the audience, the virtual digital guide must possess a comprehensive knowledge system related to the museum's exhibits. This involves structuring artifact information, historical and cultural knowledge, and exhibition layouts into a knowledge graph. The virtual guide can then use this knowledge graph to perform inference, providing accurate, insightful, and logically coherent explanations in response to visitors' inquiries. Beyond content, elements such as voice, tone, and emotional expression must also be focal points in the development of the virtual digital guide. By optimizing emotional models and affective computing techniques, adjustments can be made to the pacing, language, and tone of the narration, thereby enhancing visitor engagement and satisfaction.

Conclusions

This paper provides a comprehensive study and review of the application of generative technologies in digital museum environments, along with their long-term impact on the digital preservation of cultural heritage. Digital museums have made significant progress in artifact image generation, 3D modeling, and the creation of large-scale virtual environments using generative technologies, greatly optimizing methods for the virtual display and preservation of cultural relics. With the ongoing advancement of cutting-edge technologies such as diffusion models and generative adversarial networks, the digital reconstruction of artifacts will increasingly incorporate higher levels of authenticity and rich detail, thereby enhancing the interactivity and immersive experience of digital museums.

The construction of smart museums leverages AIGC technologies to generate high-quality digital artifacts and vivid historical scenes through text-to-image, text-to-3D model, and text-to-3D scene generation, breaking the constraints of time and space and enhancing the breadth and depth of cultural dissemination. This paper systematically organizes the main technologies used in the artifact image generation process, focusing on GANs and diffusion models. First, we compare the different modes, frameworks, and advantages and disadvantages of various GAN architectures,

highlighting their application scenarios in generative fields and offering a cross-disciplinary description and comparison in artifact generation. Additionally, diffusion models have demonstrated their progressive denoising capabilities. When combined with text and 3D modeling technologies, they offer richer methods for digital exhibitions. We systematically discuss and compare the development of diffusion models, from DDPM to Muse, providing a structured reference for future artifact digitization research. In the context of generative museums, we describe the three most popular application scenarios: artifact avatar generation, 3D heritage site scene generation, and 3D artifact texture generation. Through the latest innovative technologies and methods, we present corresponding application examples. These results showcase the powerful capabilities of generative models in visual restoration, spatial reconstruction, and detail synthesis, while offering new pathways for the virtual display, interactive experiences, and educational dissemination of digital artifacts.

At the same time, we clearly distinguish between the limitations at the theoretical and practical levels: Theoretically, current models still face issues such as training instability and low precision in generation control. Practically, the deployment of these models in real-world scenarios continues to encounter challenges such as difficulty in data acquisition, high computational requirements, and insufficient cross-domain adaptability.

Finally, this paper highlights the significant reference value of research in this field for cultural policy. The continuous maturation of generative technologies not only provides new tools for the preservation and dissemination of cultural heritage but also has a profound impact on the standardization of digital museum construction, educational outreach, and the expansion of social awareness. It offers technical support and pathway references for the formulation and implementation of relevant policies.

In the future, the integration of generative technologies with immersive technologies such as augmented reality (AR) and virtual reality (VR) will create unprecedented opportunities for innovation in museums. Visitors will be able to deeply engage with cultural experiences in multidimensional spaces through virtual tours, interactive learning, and digital artifact deconstruction, fostering a strong cultural resonance. Future research should focus more on enhancing the accuracy and diversity of generative models, while addressing the challenges of generalizing small sample data, to promote the intelligent and sustainable development of generative technologies in the digital museum sector. With ongoing technological advancements, we firmly believe that generative technologies will become a key driving force in the development of digital museums, injecting new vitality into cultural heritage and providing the public with more captivating cultural experiences.

Data availability

No datasets were generated or analysed during the current study.

Received: 25 February 2025; Accepted: 7 November 2025;
Published online: 20 November 2025

References

1. Srinivasan, R., Boast, R., Furner, J. & Becvar, K. M. Digital museums and diverse cultural knowledges: moving past the traditional catalog. *Inf. Soc.* **25**, 265–278 (2009).
2. Zheng, F., Wu, S., Liu, R. & Bai, Y. What influences user continuous intention of digital museum: integrating task-technology fit (ttf) and unified theory of acceptance and usage of technology (utaut) models. *Herit. Sci.* **12**, 253 (2024).
3. Parry, R. Digital heritage and the rise of theory in museum computing. *Mus. Manag. Curatorship* **20**, 333–348 (2005).
4. Marty, P. F. Museum websites and museum visitors: digital museum resources and their use. *Mus. Manag. Curatorship* **23**, 81–99 (2008).
5. Styliani, S., Fotis, L., Kostas, K. & Petros, P. Virtual museums, a survey and some issues for consideration. *J. Cult. Herit.* **10**, 520–528 (2009).
6. Zhou, Y., Liu, Y., Shao, Y. & Chen, J. Fine-tuning diffusion model to generate new kite designs for the revitalization and innovation of intangible cultural heritage. *Sci. Rep.* **15**, 7519 (2025).
7. Niu, M. & Zhou, Y. The optimization of illustration design in cultural and creative products for liaoning region under intelligent generative adversarial network. *IEEE Access* **13**, 114746–114755 (2025).
8. Dalong, D. The construction of the virtual museum in the forbidden city of china. *Inf. Cult.* **59**, 246–265 (2024).
9. Huang, X., Li, Y. & Tian, F. Enhancing user experiecne in interactive virtual museums for cultural heritage learning through extended reality: the case of Sanxingdui bronzes. *IEEE Access* **13**, 59405–59421 (2025).
10. Guo, X. et al. Specialized or general ai? a comparative evaluation of llms' performance in legal tasks. *Artif. Intell. Law* **33**, 1–37 (2025).
11. Croitoru, F.-A., Hondru, V., Ionescu, R. T. & Shah, M. Diffusion models in vision: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 10850–10869 (2023).
12. Creswell, A. et al. Generative adversarial networks: an overview. *IEEE Signal Process. Mag.* **35**, 53–65 (2018).
13. Sokolovskiy, K., Vershitskaya, E. & Latysheva, V. The efficient use of new cultural management tools in the adaptation of public heritage during the covid-19 pandemic. *Mus. Manag. Curatorship* **38**, 593–615 (2023).
14. Garstki, K. Virtual representation: the production of 3d digital artifacts. *J. Archaeol. Method Theory* **24**, 726–750 (2017).
15. Collomosse, J. & Parsons, A. To authenticity, and beyond! building safe and fair generative ai upon the three pillars of provenance. *IEEE Comput. Graph. Appl.* **44**, 82–90 (2024).
16. Brüns, J. D. & Meißner, M. Do you create your content yourself? using generative artificial intelligence for social media content creation diminishes perceived brand authenticity. *J. Retail. Consum. Serv.* **79**, 103790 (2024).
17. Mai, S., Zeng, Y. & Hu, H. Multimodal information bottleneck: learning minimal sufficient unimodal and multimodal representations. *IEEE Trans. Multimed.* **25**, 4121–4134 (2022).
18. Wang, H., Song, C. & Li, H. Application of social media communication for museum based on the deep mediatization and artificial intelligence. *Sci. Rep.* **14**, 28661 (2024).
19. Benyon, D. & Murray, D. Applying user modeling to human-computer interaction design. *Artif. Intell. Rev.* **7**, 199–225 (1993).
20. Pasquinelli, C., Trunfio, M., Punziano, G. & Del Chiappa, G. Online tourism experiences: exploring digital and human dimensions in in-remote destination visits. *J. Hosp. Mark. Manag.* **32**, 385–409 (2023).
21. Cheng, J. et al. Generative adversarial networks: a literature review. *KSII Trans. Internet Inf. Syst.* **14**, 4625–4647 (2020).
22. Lyu, Y., Wang, X., Lin, R. & Wu, J. Communication in human–AI co-creation: perceptual analysis of paintings generated by text-to-image system. *Appl. Sci.* **12**, 11312 (2022).
23. Deng, J., Cao, X. & Cheng, B. Research on generating cultural relic images based on a low-rank adaptive diffusion model. In *Proc. 2024 Guangdong–Hong Kong–Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence* (*DEAI 2024*) 629–634 (IEEE, 2024).
24. Hsieh, K. et al. Cultural heritage meets AI: advanced text-to-image models for digital reconstruction and preservation. In *Proc. 2024 6th International Conference on Control and Robotics*(*ICCR 2024*) 265–269 (IEEE, 2024).
25. Liu, B. et al. On the cultural gap in text-to-image generation. Preprint at arxiv:2307.02971 (2023).
26. Kumari, N. et al. Ablating concepts in text-to-image diffusion models. In *Proc. IEEE/CVF International Conference on Computer Vision* (*ICCV 2023*) 22691–22702 (IEEE, 2023).
27. Ruiz, N. et al. Dreambooth: fine tuning text-to-image diffusion models for subject-driven generation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR 2023*) 22500–22510 (IEEE, 2023).
28. Liu, Q., Zhang, Y., Bai, S., Kortylewski, A. & Yuille, A. Direct-3d: learning direct text-to-3d generation on massive noisy 3d data. In

Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (*CVPR 2024*) 6881–6891 (IEEE, 2024).
29. Höllein, L., Cao, A., Owens, A., Johnson, J. & Nießner, M. Text2room: extracting textured 3d meshes from 2d text-to-image models. In *Proc. IEEE/CVF International Conference on Computer Vision* (*ICCV 2023*) 7909–7920 (IEEE, 2023).
30. Poole, B., Jain, A., Barron, J. T. & Mildenhall, B. Dreamfusion: text-to-3d using 2d diffusion. Preprint at arXiv:2209.14988 (2022).
31. Li, K. et al. Learning pseudo 3D guidance for view-consistent texturing with 2D diffusion. In *Proc. European Conference on Computer Vision* (*ECCV 2024*) 18–34 (Springer, 2024).
32. Chen, Y., Qi, X., Wang, J. & Zhang, L. Disco-clip: a distributed contrastive loss for memory efficient clip training. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR 2023*) 22648–22657 (2023).
33. Sundaram, J. P. S., Du, W. & Zhao, Z. A survey on lora networking: research problems, current solutions, and open issues. *IEEE Commun. Surv. Tutor.* **22**, 371–388 (2019).
34. Nichol, A., Jun, H., Dhariwal, P., Mishkin, P. & Chen, M. Point-e: a system for generating 3D point clouds from complex prompts. Preprint at arXiv:2212.08751 (2022).
35. Shah, A. & Patel, P. Broadband coplanar waveguide-fed stub loaded pot shape e-textile antenna equipped with perfect electric conductor. *Int. J. RF Microw. Comput. -Aided Eng.* **31**, e22591 (2021).
36. Kantaros, A., Ganetsos, T. & Petrescu, F. I. T. Three-dimensional printing and 3d scanning: emerging technologies exhibiting high potential in the field of cultural heritage. *Appl. Sci.* **13**, 4777 (2023).
37. Chen, Z., Wang, G. & Liu, Z. Scenedreamer: unbounded 3d scene generation from 2D image collections. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 15562–15576 (2023).
38. Deng, J. et al. Citycraft: a real crafter for 3D city generation. Preprint at arXiv:2406.04983 (2024).
39. Yang, R., Ota, K., Dong, M. & Wu, X. Semantic layout-guided diffusion model for high-fidelity image synthesis in 'the thousand li of rivers and mountains'. *Expert Syst. Appl.* **263**, 125645 (2025).
40. Guerra, F.dC. F. & Mota, W. S. Current transformer model. *IEEE Trans. Power Deliv.* **22**, 187–194 (2006).
41. Gupta, K. et al. Layouttransformer: layout generation and completion with self-attention. In *Proc. IEEE/CVF International Conference on Computer Vision* (*ICCV 2021*) 1004–1014 (IEEE, 2021).
42. Paschalidou, D. et al. Atiss: autoregressive transformers for indoor scene synthesis. *Adv. Neural Inf. Process. Syst.* **34**, 12013–12026 (2021).
43. Trichopoulos, G., Konstantakis, M., Alexandridis, G. & Caridakis, G. Large language models as recommendation systems in museums. *Electronics* **12**, 3829 (2023).
44. Struski, Ł, Knop, S., Spurek, P., Daniec, W. & Tabor, J. Locogan-locally convolutional gan. *Comput. Vis. Image Underst.* **221**, 103462 (2022).
45. May, C. & Aliaga, D. Cubegan: omnidirectional image synthesis using generative adversarial networks. *Comput. Graph. Forum* **42**, 213–224 (2023).
46. Goodfellow, I. et al. Generative adversarial networks. *Commun. ACM* **63**, 139–144 (2020).
47. Zhang, H. et al. Stackgan++: realistic image synthesis with stacked generative adversarial networks. *IEEE Trans. pattern Anal. Mach. Intell.* **41**, 1947–1962 (2018).
48. Yan, L. et al. Digan: distillation model for generating 3d-aware terracotta warrior faces. *Herit. Sci.* **12**, 317 (2024).
49. Naveen, S., Kiran, M. S. R., Indupriya, M., Manikanta, T. & Sudeep, P. Transformer models for enhancing attngan based text to image generation. *Image Vis. Comput.* **115**, 104284 (2021).
50. Zhou, Y. et al. Towards language-free training for text-to-image generation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR 2022*) 17907–17917 (IEEE, 2022).
51. Hou, Y., Zhang, W., Zhu, Z. & Yu, H. Clip-gan: stacking clips and gan for efficient and controllable text-to-image synthesis. *IEEE Trans. Multimed.* **27**, 3702–3715 (2025).
52. Tao, M. et al. Df-gan: a simple and effective baseline for text-to-image synthesis. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR 2022*) 16515–16525 (IEEE, 2022).
53. Ding, H., Huang, N., Wu, Y. & Cui, X. Legan: addressing intra-class imbalance in gan-based medical image augmentation for improved imbalanced data classification. *IEEE Trans. Instrum. Meas.* **73**, 2517914 (2024).
54. Tao, Y. et al. Lagan: lesion-aware generative adversarial networks for edema area segmentation in sd-oct images. *IEEE J. Biomed. Health Inform.* **27**, 2432–2443 (2023).
55. Wang, C., Zhang, Y., Zhang, Y., Tian, R. & Ding, M. Mars image super-resolution based on generative adversarial network. *IEEE Access* **9**, 108889–108898 (2021).
56. Gilboa, G., Sochen, N. & Zeevi, Y. Y. Forward-and-backward diffusion processes for adaptive image enhancement and denoising. *IEEE Trans. Image Process.* **11**, 689–703 (2002).
57. Lee, J., Nguyen, D., Kim, J., Kang, J. & Lee, S. Double reverse diffusion for realistic garment reconstruction from images. *Eng. Appl. Artif. Intell.* **127**, 107404 (2024).
58. Huberman-Spiegelglas, I., Kulikov, V. & Michaeli, T. An edit friendly ddpm noise space: Inversion and manipulations. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR 2024*) 12469–12478 (IEEE, 2024).
59. Yang, X., Zhou, D., Feng, J. & Wang, X. Diffusion probabilistic model made slim. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR 2023*) 22552–22562 (IEEE, 2023).
60. Fernandez, P., Couairon, G., Jégou, H., Douze, M. & Furon, T. The stable signature: rooting watermarks in latent diffusion models. In *Proc. IEEE/CVF International Conference on Computer Vision* (*ICCV 2023*) 22466–22477 (IEEE, 2023).
61. Halgren, T. A. et al. Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *J. Med. Chem.* **47**, 1750–1759 (2004).
62. Park, D., Na, H. & Choi, D. Performance comparison and visualization of ai-generated-image detection methods. *IEEE Access* **12**, 62609–62627 (2024).
63. Zhang, C., Hu, M., Li, W. & Wang, L. Adversarial attacks and defenses on text-to-image diffusion models: a survey. *Inf. Fusion* **114**, 102701 (2025).
64. Zhu, Y., Li, Z., Wang, T., He, M. & Yao, C. Conditional text image generation with diffusion models. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR 2023*) 14235–14245 (IEEE, 2023).
65. Booker, L. B., Goldberg, D. E. & Holland, J. H. Classifier systems and genetic algorithms. *Artif. Intell.* **40**, 235–282 (1989).
66. Luo, H. et al. Clip4clip: an empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing* **508**, 293–304 (2022).
67. Pei, H., Zhang, C., Zhang, X., Liu, X. & Ma, Y. Recognizing materials in cultural relic images using computer vision and attention mechanism. *Expert Syst. Appl.* **239**, 122399 (2024).
68. Chen, X., Wang, G., Xu, X., Torr, P. & Lin, L. Parametric linear blend skinning model for multiple-shape 3d garments. *IEEE Trans. Vis. Comput. Graph.* **31**, 5935–5947 (2024).
69. Nie, W., Chen, R., Wang, W., Lepri, B. & Sebe, N. T2td: Text-3d generation model based on prior knowledge guidance. *IEEE Trans. Pattern Anal. Mach. Intell.* **47**, 172–189 (2024).
70. Cao, Y., Cao, Y.-P., Han, K., Shan, Y. & Wong, K.-Y. K. Dreamavatar: text-and-shape guided 3D human avatar generation via diffusion models. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR 2024*) 958–968 (IEEE, 2024).
71. Jiang, R. et al. Avatarcraft: transforming text into neural human avatars with parameterized shape and pose control. In *Proc. IEEE/CVF*

International Conference on Computer Vision (*ICCV 2023*) 14371–14382 (IEEE, 2023).
72. Tevet, G., Gordon, B., Hertz, A., Bermano, A. H. & Cohen-Or, D. Motionclip: exposing human motion generation to clip space. In *Proc. European Conference on Computer Vision* (*ECCV 2022*) 13688, 358–374 (Springer, 2022).
73. Hong, F. et al. AvatarCLIP: zero-shot text-driven generation and animation of 3D avatars. *ACM Trans. Graph* **41**, 1–14 (2022).
74. Zhang, L. et al. Dreamface: progressive generation of animatable 3d faces under text guidance. Preprint at arXiv:2304.03117 (2023).
75. Kindlmann, G. & Scheidegger, C. An algebraic process for visualization design. *IEEE Trans. Vis. Comput. Graph.* **20**, 2181–2190 (2014).
76. Denninger, M. & Triebel, R. 3D scene reconstruction from a single viewport. In *Proc. European Conference on Computer Vision* (*ECCV 2020*) 51–67 (Springer, 2020).
77. Wang, X. & Ren, J. Controllable diffusion models for hazardous construction site scene generation. *Appl. Soft Comput.* **181**, 113446 (2025).
78. Snavely, N., Simon, I., Goesele, M., Szeliski, R. & Seitz, S. M. Scene reconstruction and visualization from community photo collections. *Proc. IEEE* **98**, 1370–1390 (2010).
79. Zhang, J., Li, X., Wan, Z., Wang, C. & Liao, J. Text2nerf: Text-driven 3d scene generation with neural radiance fields. *IEEE Trans. Vis. Comput. Graph.* **30**, 7749–7762 (2024).
80. Wang, S.-Y. et al. Roomdreaming: generative-AI approach to facilitating iterative, preliminary interior design exploration. In *Proc. 2024 CHI Conference on Human Factors in Computing Systems* (*CHI 2024*) 1–20 (Association for Computing Machinery (ACM), 2024).
81. Fang, C. et al. Ctrl-Room: controllable text-to-3D room meshes generation with layout constraints. In *Proc. International Conference on 3D Vision (3DV)* (2025).
82. Ma, Y., Zhan, D. & Jin, Z. FastScene: text-driven fast 3D indoor scene generation via panoramic Gaussian splatting. In *Proc. International Joint Conference Artificial Intelligence (IJCAI)*, 1173–1181 (2024).
83. Singer, U. et al. Text-To-4D Dynamic Scene Generation. In *Proc. International Conference on Machine Learning,* Vol. 202, 31915–31929 (2023).
84. Zhou, X. et al. GALA3D: towards text-to-3D complex scene generation via layout-guided generative gaussian splatting. In *Proc. International Conference on Machine Learning,* Vol. 235, 62108–62118 (2024).
85. Silver, D., Silva, T. H. & Adler, P. Changing the scene: applying four models of social evolution to the scenescape. arXiv:2209.10665 (2022).
86. Lee, Y.-C. et al. Vividdream: generating 3D scene with ambient dynamics. Preprint at arXiv:2405.20334 (2024).
87. Bai, H. et al. Componerf: text-guided multi-object compositional nerf with editable 3D scene layout. Preprint at arXiv:2303.13843 (2023).
88. Cohen-Bar, D., Richardson, E., Metzer, G., Giryes, R. & Cohen-Or, D. Set-the-scene: global-local training for generating controllable nerf scenes. In *Proc. IEEE/CVF International Conference on Computer Vision* (*ICCV 2023*) 2920–2929 (IEEE, 2023).
89. Ge, Y. et al. 3d reconstruction of ancient buildings using uav images and neural radiation field with depth supervision. *Remote Sens.* **16**, 473 (2024).
90. Tango, T. A test for spatial disease clustering adjusted for multiple testing. *Stat. Med.* **19**, 191–204 (2000).
91. Cao, T., Kreis, K., Fidler, S., Sharp, N. & Yin, K. Texfusion: synthesizing 3D textures with text-guided image diffusion models. In *Proc. IEEE/CVF International Conference on Computer Vision* (*ICCV 2023*) 4169–4181 (IEEE, 2023).
92. Heckbert, P. S. Survey of texture mapping. *IEEE Comput. Graph. Appl.* **6**, 56–67 (2007).
93. Chen, D. Z., Siddiqui, Y., Lee, H.-Y., Tulyakov, S. & Nießner, M. Text2tex: text-driven texture synthesis via diffusion models. In *Proc. IEEE/CVF International Conference on Computer Vision* (*ICCV 2023*) 18558–18568 (IEEE, 2023).
94. Ma, Y. et al. X-mesh: towards fast and accurate text-driven 3d stylization via dynamic textual guidance. In *Proc. IEEE/CVF International Conference on Computer Vision* (*ICCV 2023*) 749–2760 (IEEE, 2023).
95. Zhang, Q. et al. A mamba based vision transformer for fine grained image segmentation of mural figures. *npj Herit. Sci.* **13**, 1–9 (2025).
96. Benyon, D., Quigley, A., O'keefe, B. & Riva, G. Presence and digital tourism. *AI Soc.* **29**, 521–529 (2014).
97. Kotler, N. New ways of experiencing culture: the role of museums and marketing implications. *Mus. Manag. Curatorship* **19**, 417–425 (2001).

Acknowledgements

This research was supported by the Technology Innovation Leading Project of Shaanxi Province (Grant No. 2024QY-SZX-11, Research and Application of New VR Technologies for Public Cultural Services) and the 2025 Key Project of the Research Base for Cultural Communication and Innovation in the J-shaped Bend of the Yellow River, Prominent Research Base for Humanities and Social Sciences in Higher Education Institutions of Inner Mongolia (Grant No. JZW2025001, Research on the Communication Strategy of the Yellow River Bend Culture Based on Digital Technology).

Author contributions

J.X.: Conceptualization, methodology, data curation, and writing— original draft and editing. L.Y.: Literature review, algorithm validation, formal analysis, and editing. R.Z.: 3D modeling technology development, core algorithm optimization, and writing—review and editing. M.Z.: Supervision, research framework design, funding acquisition, and writing—final approval. All authors have read and approved the final version of the manuscript to be published.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Mingquan Zhou.