Diffusion-Based Visual Art Creation: A Survey and New Perspectives

BINGYUAN WANG, Computational Media and Arts, The Hong Kong University of Science and Technology - Guangzhou Campus, Guangzhou, China
QIFENG CHEN, Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, Hong Kong
ZEYU WANG, Computational Media and Arts, The Hong Kong University of Science and Technology - Guangzhou Campus, Guangzhou, China and Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, Hong Kong

The integration of generative AI in visual art has revolutionized not only how visual content is created but also how AI interacts with and reflects the underlying domain knowledge. This survey explores the emerging realm of diffusion-based visual art creation, examining its development from both artistic and technical perspectives. We structure the survey into three phases: data feature and framework identification, detailed analyses using a structured coding process, and open-ended prospective outlooks. Our findings reveal how artistic requirements are transformed into technical challenges and highlight the design and application of diffusion-based methods within visual art creation. We also provide insights into future directions from technical and synergistic perspectives, suggesting that the confluence of generative AI and art has shifted the creative paradigm and opened up new possibilities. By summarizing the development and trends of this emerging interdisciplinary area, we aim to shed light on the mechanisms through which AI systems emulate and, possibly, enhance human capacities in artistic perception and creativity.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; *Computer vision*; *Computer graphics*; • **Applied computing** → Arts and humanities;

Additional Key Words and Phrases: AI-generated content, diffusion model, visual art, creativity, human-AI collaboration

ACM Reference Format:
Bingyuan Wang, Qifeng Chen, and Zeyu Wang. 2025. Diffusion-Based Visual Art Creation: A Survey and New Perspectives. *ACM Comput. Surv.* 57, 10, Article 268 (May 2025), 37 pages. https://doi.org/10.1145/3728459

This work was supported by Guangzhou Basic Research Project #2024A04J4229 and Guangzhou Education Bureau Project #2024312075.

Authors' Contact Information: Bingyuan Wang, Computational Media and Arts, The Hong Kong University of Science and Technology—Guangzhou Campus, Guangzhou, Guangdong, China; e-mail: bwang667@connect.hkust-gz.edu.cn; Qifeng Chen, Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, Hong Kong; e-mail: cqf@ust.hk; Zeyu Wang (Corresponding author), Computational Media and Arts, The Hong Kong University of Science and Technology—Guangzhou Campus, Guangzhou, Guangdong, China and Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, Hong Kong; e-mail: zeyuwang@ust.hk.

ACM 0360-0300/2025/05-ART268
https://doi.org/10.1145/3728459

1 Introduction

As an emerging concept and evolving field, **Artificial Intelligence Generated Content (AIGC)** has made significant progress and impact over the past several years, especially since the diffusion model was proposed [61]. However, visual art, encompassing a wide variety of genres, media, and styles, possesses high artistic value and diverse creativity, sparking widespread interest. However, compared to general method innovations [62, 138] and specific model designs [44, 123], relatively limited research focuses on diffusion-based methods for visual art creation. Fewer works thoroughly examine the problem, summarize frameworks, or provide trends and insights for future research.

Relevant surveys approach this problem from both technical and artistic perspectives. Some recent surveys focus on the intersection of artificial intelligence with content generation, examining data modalities and tasks [47, 83] to methodological progressions and applications [10, 16]. These surveys reviewed a series of work on artistic stylization [79], appearance editing [130], text-to-image transitions [168], and the newfound applications of AI across multiple data modalities [176]. Methodologically, they span **Neural Style Transfer (NST)** [70] and GAN inversion [158] to attention mechanisms [55] and diffusion models [23]—each contributing to the state of the art in their own right. From an application perspective, they explore the transformative integration of AIGC across various domains, and while remarkable, they also highlight challenges that call for further development and ethical consideration [96, 160]. Meanwhile, surveys with an artistic focus unravel the interplay between arts and humanities within the AIGC era, probing into the processing and understanding of art through advanced computational methods [8, 18, 92], the generative potential of AI in creating novel art forms [38, 171], and the applicability of its integration in enhancing educational and therapeutic experiences [25, 103]. We noticed a lack of surveys that specifically focus on combining diffusion-based models with visual art creation, and aim to fill this gap with our work.

This survey aims to provide a comprehensive review of the intersection of diffusion-based generative methods and visual art creation. We define the research scope through two independent taxonomies from technical and artistic perspectives, identifying diffusion-based generative techniques as one of the key methods and art as a significant application scenario. Our research goals are to *analyze how diffusion models have revolutionized visual art creation* and to *offer frameworks and insights for future research in this area*. We address four main research questions that explore the trending topics, current challenges, employed methods, and future directions in diffusion-based visual art creation.

We first conduct a structural analysis of diffusion-based visual art creation, which highlights current hot topics and evolving trends. Through categorizing data into application, understanding, and generation, we find a concentration of research on generation, specifically in controllable, application-oriented, and historically or genre-specific art creation [28, 91, 154]. Furthermore, we present a new analytical framework that aligns artistic scenarios with data modalities and generative tasks, allowing for a structured approach to the research questions. Temporal analysis suggests a post-diffusion boom in visual art creation, with a steady rise in diffusion-based methods [49, 124, 127]. Generative methods have shifted from traditional rule-based simulations to diffusion model modifications, along with a progression from image-only inputs to more controllable conditional formats, and an increase in dataset generality and process complexity [7, 109, 125, 174]. The emerging trends point toward a technical evolution from basic model frameworks to interactive systems and a shift in focus toward user requirements for creativity and interactivity. These findings set the stage for our survey, which aims to bridge the gap between technological advancements and artistic creation, fostering a synergy that can lead to a new wave of innovation in AIGC.

The exploration from artistic requirements to technical problems forms a cornerstone of our investigation in diffusion-based visual art. We scrutinize the symbiotic relationship between application domains, artistic genres, and their correspondence with data modality and generative tasks. Supported by a robust body of work [34, 66, 88, 119, 181], we delve into how different visual art forms and domains drive the development of technical solutions. From complex artistic scenarios to specific genres like traditional Chinese painting [48, 84, 101, 144, 150], our approach deciphers the intersection of AI and art, translating artistic goals into computational tasks. We establish a framework that defines these relationships via data modalities—randing from brush strokes [85] to 3D models [40]—and generative tasks like quality enhancement [57] and controllable generation [73]. These tasks are then meticulously tied to artistic objectives, with corresponding evaluation metrics such as **Contrastive Language-Image Pretraining (CLIP)** Score for controllability [117, 121] and **Fréchet Inception Distance (FID)** for fidelity [59, 128]. This multifaceted evaluation system ensures that the generated art not only meets technical standards but also fulfills the nuanced demands of artistic creation, aligning with the evolving trends in the post-diffusion era.

We also investigate the intricate designs and applications of diffusion-based methods to explore how these methods enhance the generative process in visual art. We offer a detailed classification of tasks such as controllable generation, content editing, and stylization, each bolstered by novel diffusion-based approaches that prioritize user input and artistic integrity [28, 49]. Innovations like ILVR [28] and ControlNet [174] exemplify the strides made in achieving precise control over image attributes, whereas advances in methods like GLIDE [113] and InstructPix2Pix [13] showcase the growing sophistication in content editing and the ability to adaptively respond to textual prompts. Stylization techniques, such as InST [178] and DiffStyler [67], demonstrate the nuanced application of artistic styles, whereas quality enhancement tools like eDiff-I [6] and PIXART-α [22] push the limits of image resolution and fidelity. Furthermore, we categorize these methods based on a unified diffusion model structure, highlighting advancements in individual modules such as encoder-decoders, denoisers, and noise predictors [21, 94, 98]. These developments manifest in trends that emphasize attention mechanisms, personalization, control, quality, modularity, multitasking, and efficiency [15, 78, 125]. The synthesis of these trends reflects a dynamic evolution in diffusion-based generative models, marking a transformative era in visual art creation.

The frontiers of diffusion-based visual art creation are seen through the lens of technical evolution and human-AI collaboration. Technically, we are witnessing a leap into higher dimensions and more diverse modalities, transcending traditional boundaries to create immersive experiences [169, 172]. A synergistic perspective reveals a future where human and AI collaboration is seamless, allowing for interactive systems that augment human creativity and facilitate a deeper reception and alignment with content [29, 35, 54]. These approaches range from the use of human concepts as task inspiration to the generation of content that resonates emotionally and models that encapsulate the essence of creativity [129, 155, 163]. This multidimensional approach is shifting paradigms, enabling a greater understanding and co-creation between humans and AI [126, 184, 185]. It paints a future where the boundaries between human and AI creativity become blurred, leading to a new era of digital artistry.

In summary, our literature review yields the following contributions:

— A comprehensive dataset and taxonomy of AIGC techniques in visual art creation, coded with multidimensional, fine-grained labels.
— A framework for analyzing and categorizing the relationship between diffusion-based generative methods and their applications in visual art creation, with multifaceted features and relationships as key findings.

— A summary of frontiers, trends, and future outlooks from multiple interdisciplinary perspectives.

2 Background

Prior to the emergence of diffusion models, the field of machine learning in visual art creation had already gone through several significant developments. These stages were marked by various generative models that opened new chapters in image synthesis and editing. One of the earliest pivotal advancements was **Generative Adversarial Networks (GANs)** by Goodfellow et al. [53], which introduced a new framework where a generator network learned to produce data distributions through an adversarial process. Following closely, CycleGAN by Zhu et al. [188] overcame the need for paired samples, enabling image-to-image translation without paired training samples. These models gained widespread attention due to their potential in a variety of visual content creation tasks. Simultaneously with the development of GANs, another important class of models was the **Variational Autoencoder (VAE)** introduced by Kingma and Welling [74], which offered a method to generate continuous and diverse samples by introducing a latent space distribution. This laid the groundwork for controllable image synthesis and inspired a series of subsequent works. With enhanced computational power and innovation in model design, Karras et al. [72] pushed the quality of image generation further with StyleGAN, a model capable of producing high-resolution and lifelike images, driving more personalized and detailed image generation. The incorporation of attention mechanisms into generative models significantly improved the relevance and detail of generated content. The Transformer by Vaswani et al. [143], with its powerful sequence modeling capabilities, influenced the entire field of machine learning, and in visual art generation, the successful application of Transformer architecture to image recognition with **Vision Transformer (ViT)** by Dosovitskiy et al. [41], and further for high-resolution image synthesis with Taming Transformers by Esser et al. [44], showed the immense potential of Transformers in visual generative tasks. Subsequent developments like SPADE by Park et al. [115] and the time-lapse synthesis work by Nam et al. [112] marked significant steps toward more complex image synthesis tasks. These methods provided richer context awareness and temporal dimension control, offering users more powerful creative expression capabilities. The introduction of **Denoising Diffusion Probabilistic Models (DDPMs)** by Ho et al. [61] and the subsequent showcase by Ramesh et al. [123] of DALL·E, which could create images from textual prompts based on such models, marked another leap forward for generative models, adding a new chapter to the history of model development. These models achieved breakthroughs in image quality and also demonstrated new possibilities in terms of controllability and diversity. These developments constitute a rich history of visual art creation in the field of machine learning, laying a solid foundation for the arrival of the diffusion era. In this survey, we will delve deeper into how diffusion models inherit and transcend the boundaries of these prior technologies, opening a new chapter in creative generation.

From an artistic perspective, the advancements in machine learning and generative models have intersected intriguingly with the domain of visual arts, which encompasses a wide variety of genres, media, and styles. Artists have traditionally held the reins of creative power, with the ability to produce works that carry significant artistic value and cultural resonance. The introduction of sophisticated generative algorithms offers a new toolkit for artists, potentially expanding the boundaries of their creativity [104]. As these technological tools become more accessible and integrated into artistic workflows [31], they present an opportunity for artists to experiment with novel forms of expression, blending traditional techniques with computational processes [43]. This fusion sparks widespread interest not only within the tech community but also among art enthusiasts who are curious about the new creative possibilities [107]. Machine learning models, especially those capable of generating high-quality visual content, are increasingly seen as

collaborators in the artistic process. Rather than replacing human creativity, they are enhancing it, enabling artists to explore complex patterns, intricate details, and conceptual depths that were previously difficult or impossible to achieve manually [50]. This symbiotic relationship between artist and algorithm is transforming the landscape of visual art. Artists are beginning to harness these models to create works that challenge our understanding of art and authorship [106]. As a result, the dialogue between technology and art is becoming richer, with machine learning models contributing to the creation of art that offers greater creative freedom and artistic value. This evolving dynamic prompts both excitement and philosophical reflection on the nature of creativity and the role of artificial intelligence in the future of artistic expression [89].

3 Related Work

In this section, we provide an overview of the scope of AIGC and contributions of pertinent surveys that concentrate on fields and topics relevant to diffusion-based visual art creation. We first collected 42 surveys and filtered out 30 by relevance. These surveys are primarily categorized by their focus on either technical (17) or artistic (13) aspects. Collectively, they establish the paradigm of this interdisciplinary field and create a platform for our discussion.

3.1 Relevant Surveys with Technical Focus

From a technical view, a tier of surveys focus on the advancements and implications of artificial intelligence in content generation. For example, Cao et al. [16] provide a detailed review of the history and recent advances in AIGC, highlighting how large-scale models have improved the extraction of intent information and the generation of digital content such as images, music, and natural language. We further break down this view into data and task, method, and application perspectives.

3.1.1 Data and Task Perspectives. A series of surveys inspect AIGC from data and modality and highlight the evolution and challenges in various tasks, including artistic stylization, appearance editing, text-to-image generation, text-to-3D transformation, and AI-generated content across multiple modalities. Prior to the diffusion era, the survey by Kyprianidis et al. [79] delves into the field of **Nonphotorealistic Rendering (NPR)**, presenting a comprehensive taxonomy of artistic stylization techniques for images and video. It traces the development of NPR from early semiautomatic systems to the automated painterly rendering methods driven by image gradient analysis, ultimately discussing the fusion of higher-level computer vision with NPR for artistic abstraction and the evolution of real-time stylization techniques. Schmidt et al. [130] review the state of the art in the artistic editing of appearance, lighting, and material, essential for conveying information and mood in various industries. The survey categorizes editing approaches, interaction paradigms, and rendering techniques while identifying open problems to inspire future research in this complex and active area. In the era of large generative models, the survey by Bie et al. [10] on text-to-image generation explores how the integration with large language models and the use of diffusion models have revolutionized text-to-image generation, bringing it to the forefront of machine learning research and greatly enhancing the fidelity of generated images. The review provides a critical comparison of existing methods and proposes potential improvements and future pathways, including video and 3D generation. Li et al. [83] conducted the first comprehensive survey on text-to-3D, an active research field due to advancements in text-to-image and 3D modeling technologies. The work introduces 3D data representations and foundational technologies, summarizing how recent developments realize satisfactory text-to-3D results and are used in various applications like avatar and scene generation. Finally, the survey by Foo et al. [47] on AI-generated content spans a plethora of data modalities, from images and videos

to 3D shapes and audio. It reviews single-modality and cross-modality AIGC methods, discusses the representative challenges and works in each modality, and suggests future research directions.

3.1.2 Method Perspective. A main body of recent surveys in generative AI and computer vision has been on the evolution of methodologies for style transfer, GAN inversion, attention mechanisms, and diffusion models, which have been instrumental in driving forward the state of the art. NST has evolved into a field of its own, with a variety of algorithms aimed at improving or extending the seminal work of Gatys et al. [50]. Jing et al. [70] provide a taxonomy of NST algorithms and compares them both qualitatively and quantitatively, also highlighting the potential applications and future challenges in the field. In the realm of GANs, the survey on GAN inversion [158] details the process of inverting images back into the latent space to enable real image editing and interpreting the latent space of GANs. It outlines representative algorithms, applications, and emerging trends and challenges in this area. The survey on attention mechanisms in computer vision [55] categorizes them based on their approach, including channel, spatial, temporal, and branch attention. This comprehensive review links the success of attention mechanisms in various visual tasks to the human ability to focus on salient regions in complex scenes, and it suggests future research directions. Diffusion-based image generation models have seen significant progress, paralleling advancements in large language models like ChatGPT. Zhang et al. [176] examine the issues and solutions associated with these models, particularly focusing on the stable diffusion framework and its implications for future image generation modeling. Text-to-image diffusion models are also reviewed [168], offering a self-contained discussion on how basic diffusion models work for image synthesis. This includes a review of state-of-the-art methods on text-conditioned image synthesis, applications beyond, and existing challenges. **Retrieval-Augmented Generation (RAG)** for AIGC is discussed in a survey that classifies RAG foundations and suggests future directions by illuminating advancements and pivotal technologies [182]. The survey provides a unified perspective encompassing all RAG scenarios, summarizing enhancement methods, and surveying practical applications across different modalities and tasks. Finally, an overview of diffusion models addresses their applications, guided generation, statistical rates, and optimization [23]. It reviews emerging applications and theoretical aspects of diffusion models, exploring their statistical properties, sampling capabilities, and new avenues in high-dimensional structured optimization.

3.1.3 Application Perspective. From an application perspective, recent surveys have explored the integration and impact of AIGC across different domains such as brain-computer interfaces, education, and mobile networks, emphasizing its transformative potential. The survey by Mai et al. [103] introduces the concept of brain-conditional multimodal synthesis within the AIGC framework, termed *AIGC-Brain*. This domain leverages brain signals as a guiding condition for content synthesis across various modalities, aiming to decode these signals back into perceptual experiences. The survey provides a detailed taxonomy for AIGC-Brain decoding models, task-specific implementations, and quality assessments, offering insights and prospects for research in brain-computer interface systems. The systematic literature review by Chen et al. [25] addresses AIGC's application in education, highlighting the profound impact of technologies like ChatGPT. The review identifies key themes such as performance assessment, instructional applications, and the advantages and risks of AIGC in education. It delves into the research trends, geographical distribution, and future agendas to integrate AI more effectively into educational methods, tools, and innovation. Xu et al. [160] survey the deployment of AIGC services in mobile networks, focusing on providing personalized and customized content while preserving user privacy. The survey examines the lifecycle of AIGC services, collaborative cloud-edge-mobile infrastructure, creative applications, and the associated challenges of implementation, security, and privacy. It also outlines future research directions for enhancing mobile AIGC networks.

These technically oriented surveys characterize remarkable advancements in the field of generative AI, emphasizing the innovative algorithms and interaction paradigms that enable the creation of diverse content across various data modalities. However, they also point out the existing challenges, including the need for further technical development, the consideration of ethical issues, and the imperative to address potential negative impacts on society.

3.2 Relevant Surveys with Artistic Focus

Another tier of work adopts an artistic view by specifically focusing on arts and humanities in the AIGC era. For example, Liu et al. [96] explore the transformational impact of artificial general intelligence on the arts and humanities, addressing critical concerns related to factuality, toxicity, biases, and public safety, and proposing strategies for responsible deployment. We further break the view into processing and understanding, generation, and application perspectives.

3.2.1 Processing and Understanding Perspectives. The surveys with an artistic focus shed light on the intersection of art and technology, where advanced processing techniques and computational methods are employed to understand and enhance the appreciation of visual arts. The review by DePolo et al. [38] discusses the mechanical properties of artists' paints, emphasizing the importance of understanding paint material responses to stress through tensile testing data and other innovative techniques. The study highlights how new methods allow for the investigation of historic samples with minimal intervention, utilizing techniques such as nanoindentation, optical methods like laser shearography, computational simulations, and noninvasive approaches to predict paint behavior. Castellano and Vessio [18] provide an overview of deep learning in pattern extraction and recognition within paintings and drawings, showcasing how these technological advances paired with large digitized art collections can assist the art community. The goal is to foster a deeper understanding and accessibility of visual arts, promoting cultural diffusion. The comprehensive survey by Zhang et al. [171] on the computational aesthetic evaluation of visual art images tackles the challenge of quantifying aesthetic perception. It reviews various approaches, from handcrafted features to deep learning techniques, and explores applications in image enhancement and automatic generation of aesthetic-guided art, while addressing the challenges and future directions in this field. The review by Liu et al. [92] on neural networks for hyperspectral imaging of historical paintings details the application of neural networks for pigment identification and classification. By focusing on processing large spectral datasets, the review contributes to the application of these networks, enhancing artwork analysis and preservation of cultural heritage. Last, the survey by Bengamra et al. [8] on object detection in visual art offers a taxonomy of the methods used in the analysis of artwork images, proposing a classification based on the degree of learning supervision, methodology, and style. It outlines challenges and future directions for improving object detection performance in visual art, contributing to the overall understanding of human history and culture through art.

3.2.2 Generation Perspective. Surveys focused on the generation of art through AI technologies underscore the transformative role that AI plays in both understanding and creating visual arts. Cetinic and She [19] offer an integrated review of AI's dual application in art analysis and creation, including an overview of artwork datasets and recent works tackling various tasks such as classification and computational aesthetics, as well as practical and theoretical considerations in the generation of AI art. Shahriar [133] examines the potential of GANs in art creation, exploring their use in generating visual arts, music, and literary texts. This survey highlights the performance and architecture of GANs, alongside the challenges and future recommendations in the field of computer-generated arts. The overview of AI in painting by Liu [93] reveals the field's current status and future direction, discussing how AI algorithms can produce unique art forms and automate

tasks in traditional painting, thereby promising a revolution in the digital art world and traditional painting processes. Ko et al. [75] delve into large-scale text-to-image generation models like DALL-E, discussing their potential to support visual artists in creative works through automation, exploration, and mediation. The study includes an interview and literature review, offering design guidelines for future intelligent user interfaces using large-scale text-to-image generation models. Last, the review by Maerten and Soydaner [102] on deep neural networks in AI-generated art examines the evolution of these architectures, from classic convolutional networks to advanced diffusion models, providing a comparison of their capabilities in producing AI-generated art. This review encapsulates the rapid progress and interaction between art and computer science.

3.2.3 Application Perspective. The surveys with an artistic focus on application delve into the transformative potential of integrating art with other disciplines, particularly science education and therapy, to foster holistic learning and healing experiences. Turkka et al. [142] investigate how art is integrated into science education, revealing through a qualitative e-survey of science teachers (n=66) that while the incorporation of art can enhance teaching, it is infrequently applied in classroom practices. The study presents a pedagogical model for art integration, which characterizes integration through content and activities, and suggests that teacher education should provide more consistent opportunities for art integration to enrich science teaching. The study on art therapy [65] surveys the clinical applications and outcomes of art therapy as a nonpharmacological intervention for mental disorders. The systematic review of 413 literature pieces underscores the clinical effectiveness of art therapy in alleviating symptoms of various mental health conditions, such as depression, anxiety, and cognitive impairments, including Alzheimer's and autism. It emphasizes the therapeutic power of art in assisting patients to express emotions and providing medical specialists with complementary diagnostic information.

The artistically oriented surveys reveal how technological advancements—specifically in AI—have revolutionized not only the analysis and preservation of visual arts but also enabled the active creation of innovative art forms. These studies underscore the potential of AI to deeply understand artistic nuances and contribute creatively, thus enriching the artistic domain with new tools and methodologies. However, we observe that while there is existing literature surveying diffusion models and visual art creation individually, there is a gap in research that synthesizes both perspectives. This opens up an opportunity to explore how diffusion models can be specifically applied to the domain of visual art creation, potentially leading to innovative approaches that could transform content production and understanding. We aim to bridge this gap by merging the technical intricacies of generative AI with the creative process of art. By doing so, we seek to contribute to the ongoing dialogue between art and technology, enhancing the creative process and expanding the scope of possibilities in visual art creation.

4 Research Scope and Concepts

In this section, we first define the survey's research scope and explain relevant concepts. Then, we summarize our research goals and target questions. Together, they establish a coherent context and lay a foundation for the following sections.

4.1 Research Scope

Based on the surveys discussed in the previous section, we identify two independent taxonomies in the technical and artistic realms. The first taxonomy, typical in surveys with a technical focus, categorizes diffusion-based generative techniques as one of the generative methods and art as an application scenario [16, 47, 83, 96]. However, surveys with an artistic stance commonly adopt historical or theoretical perspectives, categorize relevant research by application scenarios and

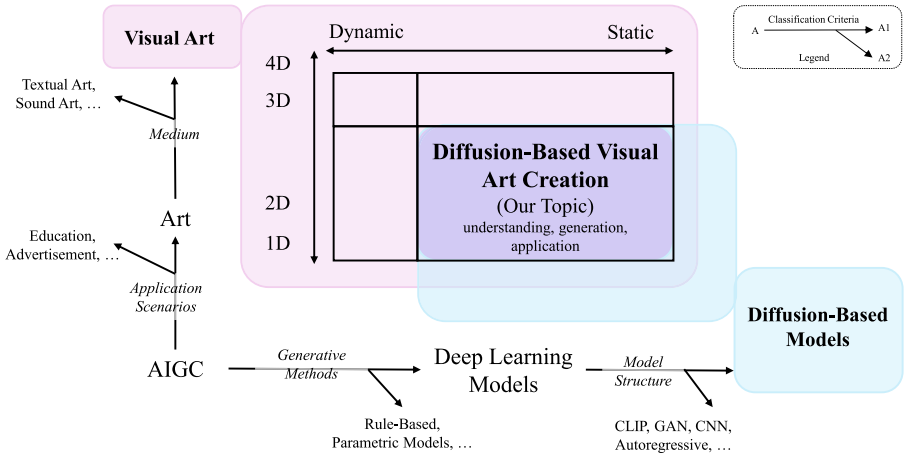

Fig. 1. Identifying the scope of this survey. We adopt two independent taxonomies to determine the research scope. For visual arts (creative targets), we primarily include 2D static visual content, supplemented by a small amount of animation, 3D, and cartoons. Regarding diffusion models (generative methods), we mainly cover aspects such as model design, task applications, and human-computer interaction.

artistic categories (in Section 6.1, we correspond them to different data modalities or applications), and focus more on the implications of generated results [17, 38, 92, 93, 171]. We display the two taxonomies and our research scope in Figure 1. The independent taxonomies are represented as perpendicular axes. Following our motivation, this survey lies in the intersection of these two axes.

4.2 Relevant Concepts

To clearly define our research scope and differentiate it from similar work, we provide an explanation and categorization method for the two most relevant concept realms and their subconcepts.

4.2.1 Diffusion Model. In January 2020, Ho et al. [61] proposed the DDPM and tested its performance on multiple image synthesis tasks, proclaiming the advent of the post-diffusion era. Ten months later, Song et al. [138] adapted the denoising process to the latent space and significantly improved the generative performance, which is called denoising diffusion implicit models. In 2021, different researchers optimized the method by integrating advanced text-image encoders (e.g., CLIP [121]) and conditioning methods (e.g., ILVR [28]). Another series of work systematically framed the generative task [113] and established relevant benchmarks [39], demonstrating surpassing performance over previous state-of-the-art methods.

In early 2022, many technical companies released respective diffusion-based generative frameworks, including DALL·E-2 [122], Imagen [127], and Stable Diffusion [124], among others. These methods feature extensive training and can generate high-quality, artistic images to meet commercial needs. From late 2022, the field has shifted from a common focus to different subtracks and downstream applications, by diversifying multiple tasks, introducing different methods, and adapting to various scenarios. Meanwhile, within the AIGC framework, the field of NLP (Natural Language Processing) has also witnessed significant breakthroughs. Researchers proposed foundational models (e.g., the GPT series [114]), designed adaptation methods (e.g., LoRA [64]), and achieved comparable performance with humans in NLP tasks [14]. Combined with these advancements, the field of Diffusion Model increased in both width and inclusiveness, becoming more expanded and more interconnected with other fields.

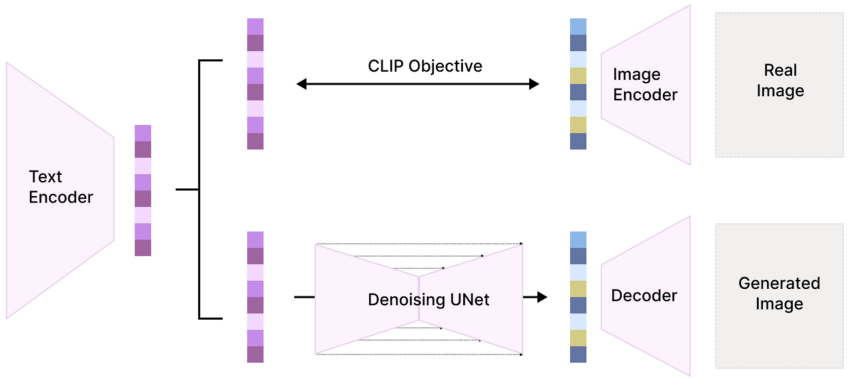

Fig. 2. Diffusion-based generative structures suggested by Stable Diffusion [124] and DALL·E-2 [122]. The image illustrates how the diffusion model integrates with the CLIP model to form the pipeline for generative tasks. The upper half shows the training process, and the lower half shows the inference process with the internal mechanism of diffusion models.

Figure 2 illustrates the basic structure of how a diffusion model is used as the core structure for a complete generative process (see Figure 2(a)), and how the model is combined with other pretrained foundational models (e.g., CLIP as text-image encoder, see Figure 2(b)) to accomplish typical text-to-image generative tasks. From a technical perspective, the structure of a diffusion-based generative model typically consists of the following five parts:

— *Encoder and decoder*: In image generation, the encoder and decoder connect the pixel space and latent space. During the generation process, the encoder compresses the input data into a latent representation, and the decoder subsequently reconstructs the output from this compressed form.
— *Denoiser*: As a core component, the denoiser works to remove noise from the latent, in a step-by-step manner, by a set of learned Gaussian noises. Researchers design both new model structures and denoising processes for better performance.
— *Noise predictor*: This module predicts key parameters of noise distribution, which is learned during the training process. Setting proper noise can guide the generation process toward intended targets.
— *Post-processor*: After the initial output is generated, this module refines the results by enhancing its resolution and final quality.
— *Additional modules*: These may include any extra components that supplement the core functionality, such as modules to improve controllability or fulfill specific tasks.

In Section 6.4, we will adhere to this framework to categorize different generative methods.

4.2.2 Visual Art. We break up the topic by three different perspectives of visual art (1) as a conceptual realm under *art*, (2) as visual contents created by *artists*, and (3) as generated results with the quality of *artistic*. These perspectives will be revisited in the Discussion section (Section 7) and are mentioned briefly next:

— *What is visual art?* Art is broadly understood to encompass a variety of creative expressions that convey ideas, emotions, and concepts through various media. Among them, the field of Visual Art features different visual forms as media [80]. This field can be further split into different artistic categories, including painting, sketch, and sculpture, among others (where painting is the most common category), and different dimensions, including 1D

brush stroke, 2D images, 3D scenes, and so on (where 2D images are the most common representation).

— *Who is the artist?* Throughout our history, different schools of criticism have adopted different views on the subject and creation of artwork. One perspective of AIGC is to mimic the role of artists with advanced generative models, which provides a possible future framework for creativity [105].

— *How is visual art created?* In visual art generation, a common way is to incorporate human aesthetics and expertise into both the generative and evaluation processes. Such a system includes data choice, task definition, and model design, which is parallel to the artist's role as both creators and receivers (e.g., [43]). However, the process is data driven and rule based, which is different from perception, emotion, and creativity as artists' driving forces.

— *What is artistic?* In the artistic realm, the definition of artistic features more psychological and philosophical elements and is also under debate [42]. However, scientific researchers often adopt a common ground and propose more acceptable standards and more approachable metrics. As will be discussed in Section 6.3, the commonly referred terms include high quality, stylistic/realistic, and controllable, among others.

4.3 Research Goals and Questions

Following the previous discussion and prior work, we summarize two research goals of our work:

–G1 *Analyze how diffusion-based methods have facilitated and transformed visual art creation.* How are diffusion-based generative systems and models used for different visual art applications?

–G2 *Provide frameworks, trends, and inspirations for future research in relevant fields.* How may human and generative AI inspire each other in diffusion-based visual art creation?

Based on the two research goals, we further propose four research questions as the basis of this survey.

–Q1 *What are the most attended topics in diffusion-based visual art creation?* This is the basic step to identify hot issues and construct a consistent framework. The question also concerns contrasts between diffusion-based and non-diffusion-based methods, and the temporal features and evolution of this field.

–Q2 *What are current research problems/needs/requirements in diffusion-based visual art creation?* This question ranges from an artistic/user perspective to a technical/designer perspective. In the following sections, we further break it down to artistic requirements and technical problems, featured by application scenarios, data modalities, and generative tasks, and attempt to establish connections between them.

–Q3 *What are the methods applied in diffusion-based visual art creation?* For each technical problem, we focus on diffusion-based method design according to its modalities and tasks. As DDPM, DDIM, and their extensions follow similar model structures, we can further categorize and organize the methods based on the unified structure of an extended diffusion model.

–Q4 *What are the frontiers, trends, and future works?* We are interested in the following questions: Are there any further problems to solve? How may we leverage the development of a diffusion model and its application in relevant fields to cope with the problems?

5 Method

In this section, we provide the method of our data collection and filtering process. We first searched for papers within the past 20 years with a focus on diffusion-based visual art creation. As shown

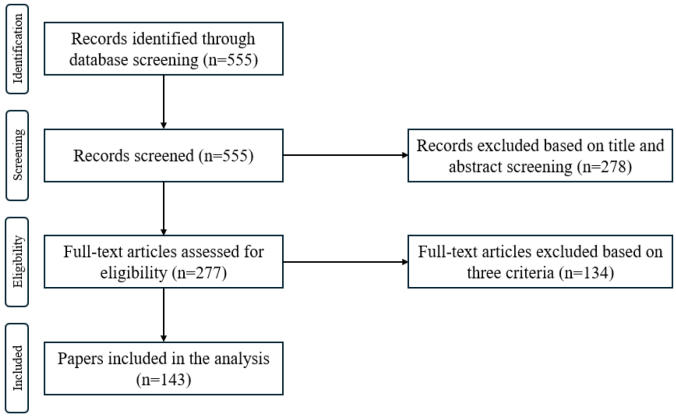

Fig. 3. The PRISMA [108] procedure of our literature review. We conducted a four-phase filtering process including identification, screening, eligibility, and final dataset coding, following the standards described in Section 5. The "n"denotes the number of papers included.

in Figure 3, we selected the papers following the PRISMA [108] guideline and its four-phase procedure. To meet our research goal, we set up three criteria to filter the literature:

- –C1 *Visual art creation*: The paper needs to include visual art as (one of) its application scenarios or demonstrate capacities in generating highly artistic results (according to figures displayed in the paper).
- –C2 *Diffusion-based model*: The paper needs to design its generative model based on DDPM [61], DDIM [138], or their extensions (e.g., ControlNet [174], T2I-Adapter [109]).
- –C3 *Beyond the context*: We also include some relevant works that establish datasets, provide understanding, or design systems that can facilitate, guide, or inspire diffusion-based visual art creation.

5.1 Phase 1: Identification

We aimed to identify (potentially) high-impact papers on diffusion-based visual art creation. Using Google Scholar metrics, we identified 10 venues as primary sources of data: Special Interest Group on Computer GRAPHics and Interactive Techniques Conference (SIGGRAPH), Conference on Computer Vision and Pattern Recognition (CVPR), International Conference on Computer Vision (ICCV), European Conference on Computer Vision (ECCV), IEEE Visualization Conference (IEEE VIS), ACM Conference on Human Factors in Computing Systems (ACM CHI), *IEEE Transactions on Graphics* (IEEE TOG), *Computer Graphics Forum* (CGF), *IEEE Transactions on Visualization and Computer Graphics* (IEEE TVCG), and ACM Multimedia Conference (ACM MM).

To focus our work according to the two dimensions in Section 4.1, we established a pool of index terms and phrases for the query. The common terms include "artist," "artistic," "visual art," "painting" (artistic perspective), and "diffusion model," "generation," "diffusion" (technical perspective). An example query of this on the ACM Digital Library is as follows:

Title: ((visual art OR painting OR artis*) AND (diffusion)) OR Abstract: ((visual art OR painting OR artis*) AND (diffusion)),

where the asterisk (*) denotes any number of unknown characters (wild cards). This way, we included words such as "artist," "artists," and "artistic." We included both full papers and posters (several artistic creations are featured on posters) in English from the past 22 years (2003–2024).

We used a period of more than two decades to cover publications with most of the modern technologies for visual content creation. Besides papers officially published in journals and conferences, we also included arXiv preprint papers (especially those within a review cycle) to provide the most updated research progress.

5.2 Phase 2: Screening

We screened the titles and abstracts of the 555 papers collected in Phase 1 using the aforementioned selection criteria (C1–C3). For example, papers that did not have any artistic results were excluded. Out of the 555 papers, we included 277 papers for Phase 3, thereby excluding 278 papers. We invited an external author and the two authors individually rated the same set of 20 randomly chosen papers for inclusion. The Cohen's kappa is 0.85.

5.3 Phase 3: Eligibility

Based on the results from Phase 2, we screened the full-text articles for eligibility using three criteria. The reasons for exclusion in this phase were either (a) that the paper met neither C1 nor C2 despite the abstract screening, or (b) that the paper met either C1 or C2 but not C3 (i.e., it had little relevance to diffusion-based visual art creation). In this phase, we further excluded 134 publications, ultimately leaving 143 papers for the final phase.

5.4 Phase 4: Final Dataset and Coding Process

We included the remaining 143 papers in the review. Among them, 85 papers meet both C1 and C2 and the remaining 58 papers meet C3 and either C1 or C2. We coded each paper based on the Application Domain (if applicable, same below), Artistic Category, Data Modality, Result Feature, Generative Task, Generative Method, and Method Category. In this way, we derived a whole table containing 143 entries with seven attributes and a subtable containing 85 entries.

6 Findings

In this section, we aim to fulfill G1 by answering questions Q1 through Q3.

6.1 Structural Analysis and Framework Construction

6.1.1 Data Classification. We focus on the first question: What are the currently most attended topics in diffusion-based visual art creation? (Q1) We first summarized different paper codes proposed in Section 5.4 along with the index terms of each selected paper. Among them, a major part is closely related to method design, whereas others concern data, modalities, artistic genres, and application scenarios. We found that these terms basically form three categories and thus applied a Venn chart to characterize different works, as shown in Figure 4. The three categories are as follows:

- *Application*: Different application scenarios (e.g., different art genres [51, 120], visualization [159, 167]).
- *Understanding*: Different data forms and corresponding modalities (e.g., image series [12, 140], 3D scenes [40, 56, 169]). From an artistic perspective, the first two categories characterize different art forms/genres with corresponding features.
- *Generation*: Different generative tasks (e.g., style control [87, 157], style transfer [180, 181], image editing [11, 58]) and different generative methods (e.g., ControlNet [99, 161], Textual Inversion [2, 178], LoRA [24, 132]).

With such a categorization method, we can approach the dataset from different perspectives and identify corresponding hot topics. As shown in Figure 4, most of the selected works lie in four subsets of the seven areas, including the following:

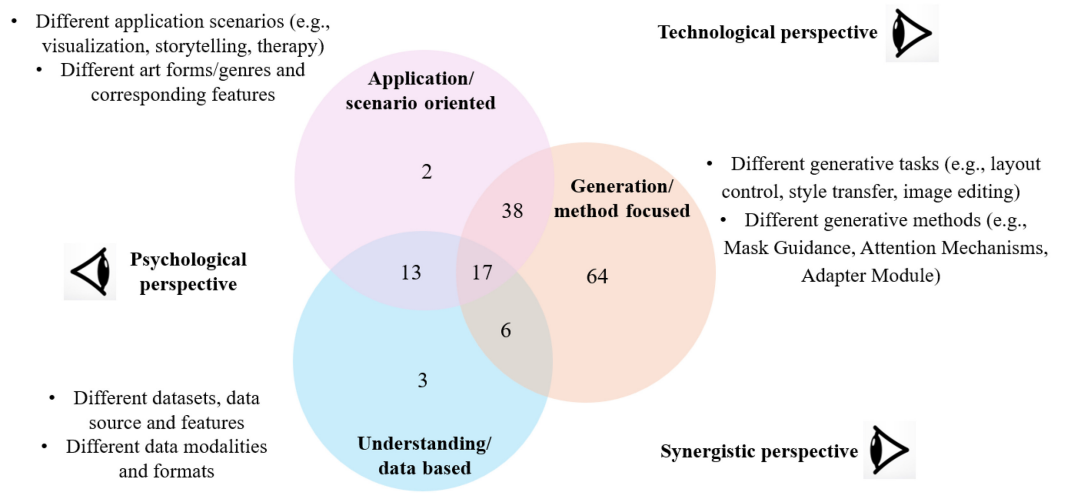

Fig. 4. Venn chart for topics in visual art creation. The chart is summarized from data distribution and annotations in our dataset. This framework is used to categorize and distinguish the blueprints of relevant research (Section 6.1) and to analyze the development and current state of this field (Section 6.2). In Section 7, we further provide technological, synergistic, and application perspectives as extensions of these three categories for development trends and future work.

— *Generation (125)*: Generation and editing with controllable style, subject, and layout (e.g., personalization [111, 156], stylization [141, 153], layout control [33, 162])
— *Generation* ∩ *Application (55)*: Application-oriented generation (e.g., art therapy [190], visual art education [37], computational arts metaverse [81])
— *Application* ∩ *Understanding (30)*: Mimicking specific historical context/ artistic genre (e.g., Chinese calligraphy [145], Indian Art [76], St Paul painting [36])
— *Generation* ∩ *Understanding (23)*: Analysis and understanding by generation (e.g., reflection on the essence of art and creativity [165], exploring new concepts and possibilities [139])

6.1.2 Framework Construction. Since most of the work is concentrated at the pole of generation, we dived into the generative part and paid more attention to the intersection areas. We found the following. First, research in diffusion-based visual art creation is typically characterized by different artistic scenarios and technical methods. Second, the artistic requirements and technical problems are basically connected by specifying data modality and extracting generative tasks. As a result, we summarized a new framework that can better characterize the current research paradigm (Figure 5).

Based on the framework, we can further break down Q1 into a series of consequential questions and approach a generative problem from different perspectives:

— *Scenario*: What are the common features and requirements of different artistic scenarios?
— *Modality*: What are the data modalities applied, including training dataset, input, conditions, and output?
— *Task*: What are popular research problems in generating visual arts, including their technical statements and classification?
— *Method*: What are the methods used to augment and adapt diffusion models?

In the following sections, we refer to this structure to analyze the relationships represented by each red dashed line.

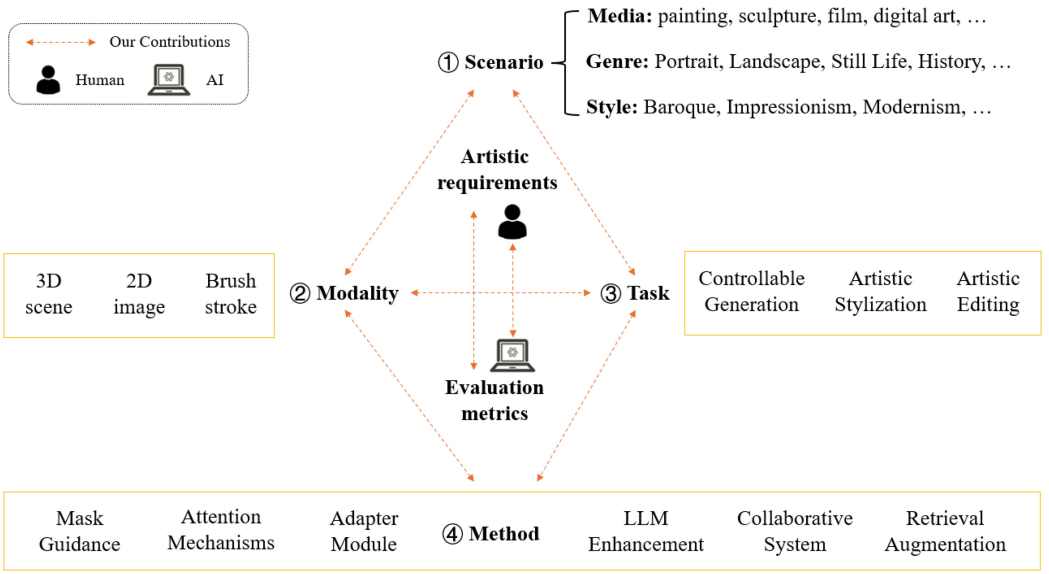

Fig. 5. An overall framework for diffusion-based visual art creation. The main contributions of this work lie in establishing the connections among scenario, modality, task, and method, as well as outlining a general roadmap from artistic requirements (human perspective) to technical problems (AI perspective). This framework is then used to analyze each individual paper in our dataset (Sections 6.3 and 6.4).

6.2 Temporal Analysis and Trend Detection

We investigate how the number of publications, categories, and keywords in different dimensions evolve over time in our dataset in the supplementary material. In this section, we specifically focus on the difference between pre-diffusion and post-diffusion eras.

6.2.1 Qualitative Comparison. From a microscopic perspective, we are interested in how the development of diffusion-based models introduces new methods for solving traditional problems. We thus selected five artistic genres or scenarios, including robotic painting, Chinese landscape painting, ink painting, story visualization, and artistic font synthesis, to compare different approaches for similar problems and tasks.

Figure 6 displays an example to compare different methods for similar tasks (portrait stylization) before and after the diffusion era. According to the similarities between the two workflows, we derived a unified, model-agnostic structure for solving the problem. As shown in Figure 6, in solving the task of face stylization (stylized portrait generation), both methods generate the new image based on the previous two images. Meanwhile, they both refer to the input image (content reference) for details and local information, and the template image (style reference) for background and global information in the generative process. When traditional tasks meet with new methods, such frameworks provide an interesting perspective to capture the embedded human expertise. However, the two approaches display multiple differences in result quality, model complexity, computational cost, and so on. Based on multiple pairs of selected work in our dataset, we summarize the following trends:

— *Input format*: From images only to images and masks as conditional input (improved controllability)
— *Dataset*: From fixed database to arbitrary image (enhanced generalizability)

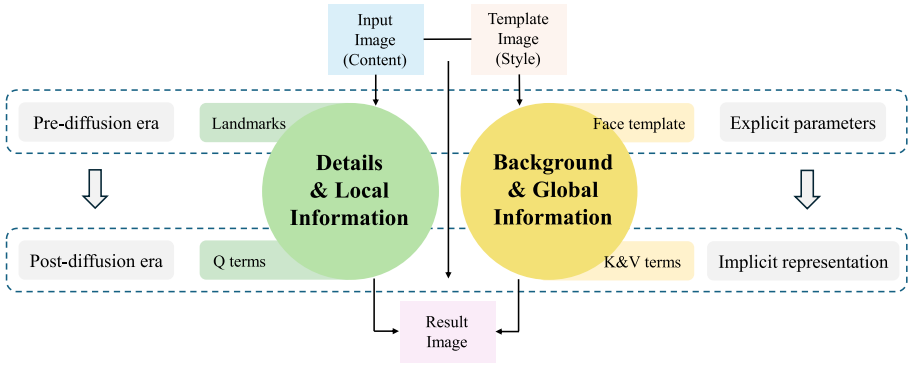

Fig. 6. Comparison between different methods on similar tasks before and after the diffusion era. We select the task of portrait stylization as an example. We compare traditional methods based on image blending [26] and the model structure of the newly proposed Portrait Diffusion [91]. Based on them, we abstract a unified paradigm for portrait stylization.

Table 1. Top Growing Keywords in Method Features and User Requirements by Year

Year	Method Features	User Requirements
2021	basic model, dataset, metric, evaluation	(no specific requirements)
2022	framework, adaptive, sampling, fine-tuning	photorealistic, multilayer, artistic, creative, coherent
2023	tuning-free, training-free, system, prompt	controllable, subject, disentanglement, interaction, painterly
2024	inversion, dilation, layer-aware, step-aware	personalization, composition, visualization, concept, context

— *Generative process*: From explicit/pixel manipulation to implicit/latent manipulation (higher complexity)
— *Method category*: From traditional rule-based image processing to diffusion-based image stylization (more computational cost and time consumption)

6.2.2 A Brief Summary and Outlook. In the previous section, we compared methods before and after the diffusion era, to borrow frameworks and ideas from the pre-diffusion era and inspire new method design. Here we are further interested in identifying research gaps from temporal trends and task-method relationships in diffusion-based visual art creation.

In Table 1, we present the top growing keywords in method features and user requirements for the post-diffusion era. We first calculate keyword frequencies for each year from 2020 to 2024 (up until May 2024). We then calculate the difference in each word's frequency compared to the previous year. The words with the highest frequency growth are identified as "Top Growing Keywords." We identify some major trends in diffusion-based visual art creation:

— *Technically*, the research type developed from a basic model to a generative framework to an interactive system. Researchers' design focus also shifted from developing benchmarks (dataset, metric, evaluation) to introducing generative methods (sampling, inversion, dilation), with a general trend to simplify the generative process (tuning-free and training-free).
— *Artistically*, user requirements are diverging from higher quality (photorealistic, artistic, coherent) to multiple diversified needs (controllable, composition, visualization), and research focus has shifted from the generated visual content (multilayer, coherent) to creative subject (personalization, concept, context). The most notable requirement is *creative*, which emerged 2 years ago but has not been well resolved until now [148].
— *Interdisciplinarily*, the keywords also manifested more collaborations between human creators and AI models. On the one hand, experts introduced principles in the diffusion process

(e.g., step-aware) and concepts from artistic areas (e.g., layer-aware), to boost controllability and performance. On the other hand, researchers ventured into understanding implicit latent (e.g., disentanglement), adapting the system to user inputs (e.g., prompt), and catering the diffusion model to the human thinking process (e.g., interaction).

In Section 7, we will go into detail to discuss the trends and future outlook from multiple perspectives.

6.3 From Artistic Requirements to Technical Problems

In this section, we focus on the upper half of the rhombus framework (see Figure 5), to summarize current research problems/needs/requirements in diffusion-based visual art creation (Q2). Specifically, we start from application scenarios and artistic genres, analyze their corresponding data modality and generative task, and dive into their key requirements/goals and computational statements. By doing this, we aim to fill in the three connections and bridge the gap between artistic requirements and technical problems.

6.3.1 Application Domain and Artistic Category. Among our 143 selected papers, 70 are coded as application/scenario oriented. Within this subset, 55 papers focus on specific artistic categories (e.g., traditional paintings, human portraits, and specific art genres), and 17 focus on relevant domains (e.g., story visualization, replication prevention, human-AI collaboration). We summarize representative work in different application scenarios, focusing on how they formulate and tackle the domain issues.

The first series of works view visual art (or digital art, fine art) as a general category. Abrahamsen and Yao [1] introduce innovative methods to invent art styles using models trained exclusively on natural images, thereby circumventing the issue of plagiarism in human art styles. Their approach leverages the inductive bias of artistic media for creative expression, harnessing abstraction through reconstruction loss and inspiration from additional natural images to forge new styles. This holds the promise of ethical generative AI use in art without infringing upon human creators' originality. In a similar vein, Zhang et al. [181] address the limitations of existing artistic style transfer methods, which either fail to produce highly realistic images or struggle with content preservation, by proposing ArtBank. This novel framework, underpinned by a pre-trained diffusion model and an ISPB (Implicit Style Prompt Bank), adeptly generates lifelike stylized images while maintaining the content's integrity. The added SSAM (Spatial-Statistical-based self-Attention Module) further refines training efficiency, with their method surpassing contemporary artistic style transfer techniques in both qualitative and quantitative evaluations. Meanwhile, Qiao et al. [119] explore the use of image prompts in conjunction with text prompts to enhance subject representation in multimodal AI-generated art. Their annotation experiment reveals that initial images significantly improve subject depiction, particularly for concrete singular subjects, with icons and photos fostering high-quality, aesthetically varied generations. They provide valuable design guidelines for leveraging initial images in AI art creation. Furthermore, Huang et al. [66] present the MGAD (Multimodal Guided Artwork Diffusion) model, a novel approach to digital art synthesis that leverages multimodal prompts to direct a classifier-free diffusion model, thereby achieving greater expressiveness and result diversity. The integration of the CLIP model unifies text and image modalities, with substantial experimental evidence endorsing the efficacy of the diffusion model coupled with multimodal guidance. Last, Liao et al. [88] contribute to the field by introducing ArtBench-10, a class-balanced, high-grade dataset for benchmarking artwork generation. It stands out with its clean annotations, high-quality images, and standardized dataset creation process, addressing the skewed class distributions prevalent in prior artwork datasets. Available in multiple resolutions and formatted for seamless

integration with prevalent machine learning frameworks, ArtBench-10 facilitates comprehensive benchmarking experiments and in-depth analyses to propel generative model research forward. Collectively, these works illustrate the dynamic intersection of AI and art, where innovative methodologies and datasets are expanding the frontiers of artistic creation, opening avenues for novel styles, ethical considerations, and enhanced representation in the digital art sphere.

The second series of works focus on specific artistic genres or historical contexts, among which traditional Chinese painting is most frequently visited. Wang et al. [154] introduce CCLAP, a pioneering method for controllable Chinese landscape painting generation. By leveraging a latent diffusion model, CCLAP consists of a content generator and style aggregator that together produce paintings with specified content and style, evidenced by both qualitative and quantitative results that showcase the model's artful composition capabilities. A dedicated dataset, CLAP, has been developed to evaluate the model comprehensively, and the code has been made accessible for broader use. Addressing the issue of low-resolution images in the digital preservation of Chinese landscape paintings, Lyu et al. [101] propose the diffusion probabilistic model CLDiff. It employs iterative refinement steps akin to the Langevin dynamic process to transform Gaussian noise into high-quality, ink-textured super-resolution images, while a novel attention module enhances the U-Net architecture's generative power. Fu et al. [48] tackle the challenge of generating traditional Chinese flower paintings in various styles, such as line drawing, meticulous painting, and ink painting, using a deep learning approach. Their Flower-GAN framework, bolstered by attention-guided generators and discriminators, facilitates style transfer and overcomes common artifacts and blurs. A new loss function, Multi-Scale Structural Similarity, is introduced to enforce structural preservation, resulting in higher-quality multistyle Chinese art paintings. From the perspective of generative teaching aids, Wang et al. [150] present "Intelligent-paint,"a method for generating the painting process of Chinese artworks. Using a ViT-based generator and an adversarial learning approach, this method emphasizes the unique characteristics of Chinese painting, such as void and brush strokes, employing loss constraints to align with traditional techniques. The coherence of the generated painting sequences with real painting processes is further validated by expert evaluations, making it a valuable tool for beginners learning Chinese painting. Finally, Li et al. [84] introduce the novel task of artistically visualizing classical Chinese poems. For this, they construct the Paint4Poem dataset, comprising high-quality poem-painting pairs and a larger collection to assist in training poem-to-painting generation models. Despite the models' capabilities in capturing pictorial quality and style, reflecting poem semantics remains a challenge. Paint4Poem opens many research avenues, such as transfer learning and text-to-image generation for low-resource data, enriching the intersection of literature and visual art. These works collectively highlight the potential of diffusion-based techniques in enriching the field of traditional Chinese painting, offering advanced tools for both creation and restoration and enhancing the educational process for aspiring artists.

With the development of diffusion-based generative methods, the application scenario has expanded to cover a wide range of artistic categories, including human images, portraits, fonts, and more. Ju et al. [71] have crafted the Human-Art dataset to bridge the gap between natural and artificial human representations. Spanning natural and artificial scenes, this dataset is comprehensive, covering 2D and 3D instances, and is poised to enable advancements in various computer vision tasks such as human detection, pose estimation, image generation, and motion transfer. Liu et al. [91] present Portrait Diffusion, a training-free face stylization framework that utilizes text-to-image diffusion models for detailed style transformation. This novel framework integrates content and style images into latent codes, which are then delicately blended using Style Attention Control, yielding precise face stylization. The innovative Chain-of-Painting method allows for gradual redrawing of images from coarse to fine details. In the realm of secondary painting for artistic productions like comics and animation, Ai and Sheng [3] introduce

Stable Diffusion Reference Only, a method that accelerates the process with a dual-conditioning approach using image prompts and blueprint images for precise control. This self-supervised model integrates seamlessly with the original UNet architecture, enhancing efficiency and controllability without the need for complex training methods. Wang et al. [144] tackle the challenge of creating nontypical aspect-ratio images with MagicScroll, a diffusion-based image generation framework. It addresses issues of content repetition and style inconsistency by allowing fine-grained control of the creative process across object, scene, and background levels. This model is benchmarked against mediums like paintings, comics, and cinema, demonstrating its potential in visual storytelling. Last, Tanveer et al. [141] introduce DS-Fusion, a method for generating artistic typography that balances stylization with legibility. Utilizing large language models and an unsupervised generative model with a diffusion model backbone, it creates typographies that visually convey semantics while remaining coherent. DS-Fusion is validated through user studies and stands out against prominent baselines and artist-crafted typographies. Together, these advancements signify a major leap in the application of diffusion-based methods to myriad artistic categories. By encompassing human-centric datasets, training-free frameworks, speed-enhancing models for artists, tools for visual storytelling, and typography generation techniques, the scope of AI in art creation is being pushed to new, previously unimagined heights.

6.3.2 Representing Scenarios as Modalities and Tasks. Next, we attempt to structure different application scenarios by their corresponding data modalities and generative tasks. In this way, we aim to approach the embedded technical problems and establish alignment between artistic requirements and technical problems.

Following the common practice in AIGC, we first categorize artistic scenarios by different data modalities:

— *Thread/brushstroke*: The first series of work focus on brush stroke generation. The problem has been long studied and technically attended since around 2000 and can be well solved by traditional rendering and rule-based methods, with little involvement of diffusion-based models [9, 45, 82, 110].
— *2D pixels/image*: Among all modalities, 2D images are the most common representation in visual art, and thus a great bunch of work adopts it as (one of) the target representations [4, 32, 187, 189].
— *Image series/video*: Typically considered as a temporal extension or duplication of a single image, image series and videos are common in certain scenarios such as storytelling and animation [12, 52, 140, 146].
— *3D model/scene*: Some art forms are based on spatial expression, and the field of 3D generation is also extending its artistic perspective, and thus 3D visual art creation is growing rapidly [40, 56, 120, 172].
— *Others*: Other artistic genres are commonly believed to possess certain modality features. For example, sketches share both raster and vector representations, thus inspiring researchers to explore different generative approaches [30, 149, 151, 152].

Next, we summarize typical tasks in diffusion-based visual art creation:

— *Quality enhancement*: As the baseline task in content generation and the basic requirement in visual art creation, the generated content should possess higher resolution and better quality. This is commonly realized by aesthetic training data, advanced model structure, more parameters, and result optimization designs. In the post-diffusion era, these methods are integrated into training foundation models [22, 57, 113, 124, 180].
— *Controllable generation*: The requirement emerges from artists' need to precisely control each perspective of their generated results, including context, subject, content, and style.

Table 2. Correspondence between Artistic Goals and Evaluation Metrics

Artistic Goal	Example Evaluation Metric
Controllability	CLIP Score [121], CLIP Directional Similarity [117]
Visual Quality	User studies, LAION-AI Aesthetics [131]
Fidelity	Fréchet Inception Distance [59], Inception Score [128]
Interpretability	Disentanglement metrics, feature attribution

Researchers adapt diversified ways, including additional information encoding, the cross-attention mechanism, and retrieval augmentation, to support different forms of conditions [67, 86, 174, 183].

— *Content editing and stylization*: This task is seen in various scenarios such as iterative generation, collaborative creation, and image inpainting. Following the understanding of high-level concept and low-level style in deep latent structure, experts are also working on decoupling the two aspects, to improve the performance of diffusion-based models on style transfer, style control, style inversion, and so forth [34, 58, 73, 98].

— *Specialized tasks*: According to different visual art scenarios and inspired by human concepts, experts summarized and proposed new tasks including compositional generation (e.g., concept, layout, layer) and latent manipulation. Still, more research is application oriented, designed, and optimized for specific data types (e.g., human portrait) or specific scenarios (e.g., multiview art) [21, 94, 144, 157, 186].

6.3.3 From Artistic Goals to Evaluation Metrics. In diffusion-based visual art creation, artistic goals drive the development of generative tasks, and the success of these tasks is measured using specific evaluation metrics. In Table 2, we summarize common artistic goals and list several evaluation metrics as an example:

- *Controllability*: Achieving precise control over generated outcomes is measured by metrics that evaluate the adherence to user-specified prompts and directions.
  - *CLIP Score*: Assesses alignment between text prompts and generated images using CLIP embeddings [121].
  - *CLIP Directional Similarity*: Measures the semantic similarity between changes in text prompts and corresponding changes in generated images [117].
- *Visual quality*: The quality of generated art is quantified by subjective and objective metrics that reflect the aesthetic and technical excellence of the artwork.
  - *User studies*: Subjective evaluations where users rate the visual appeal and aesthetic qualities of generated content [147].
  - *LAION-AI Aesthetics*: A metric that uses a dataset from LAION-AI to objectively evaluate the aesthetic aspects of generated images, such as harmony, balance, and composition [131].
- *Fidelity*: The fidelity of the generated content to the target data distribution is gauged using metrics that compare the statistical properties of generated and real artwork.
  - *FID* [59]: Quantifies the distance between feature distributions of generated and real images to assess the realism and diversity of the content [95].
  - *IS (Inception Score)* [128]: Measures the clarity and variety of generated images based on the Inception network's confidence in classifying the content and the diversity across the dataset [39].
- *Interpretability*: For the goal of interpretability, metrics assess how well we can understand and manipulate the generative model's inner workings.

- *Disentanglement metrics*: Utilize methods such as the β-VAE metric [60] to quantify the independence of different factors in the latent variables [69].
- *Feature attribution*: Employ techniques such as SHAP (SHapley Additive exPlanations) [100] to determine which features or latent variables have the greatest impact on the characteristics of the generated content [77].

6.4 Design and Application of Diffusion-Based Methods

In the previous discussion, we gradually shifted from an artistic/user perspective to a technical/designer perspective. In this part, we focus on the lower half of the Rhombus framework (see Figure 5), to summarize specific methods applied in diffusion-based visual art creation (Q3).

6.4.1 From Generative Tasks to Method Design. Based on the previously summarized generative tasks, we first categorize representative diffusion-based methods applied to solve each problem. We specifically focus on controllable generation, content editing, and stylization, which together make up more than 80% of research in generative/method-based research.

Controllable Generation. In the realm of controllable generation, various studies have presented innovative approaches to guide diffusion models effectively. The work by Choi et al. [28] introduces ILVR (Iterative Latent Variable Refinement, which conditions DDPMs using a reference image. The ILVR method directs a single DDPM to generate images with various attributes informed by the reference, enhancing the controllability and quality of generated images across multiple tasks like multidomain image translation and image editing. Gal et al. [49] propose a method that personalizes text-to-image generation by learning new "words"to represent user-provided concepts. This approach, named *Textual Inversion*, adapts a frozen text-to-image model to generate images of unique concepts. By embedding these unique "words"into natural language sentences, users have the creative freedom to guide the AI in generating personalized images. In another breakthrough, Zhang et al. [174] present ControlNet, an architecture that adds spatial conditioning controls to pretrained text-to-image diffusion models. ControlNet takes advantage of "zero convolutions" and existing deep encoding layers from large models, allowing the fine-tuning of conditional controls like edges and segmentation with robust training across different dataset sizes. Building further on control mechanisms, Zhao et al. [183] introduce Uni-ControlNet, a unified framework that enables the simultaneous use of multiple control modes, both local and global, without the need for extensive training from scratch. The framework's unique adapter design ensures cost-effective and composable control, enhancing both controllability and generation quality. Finally, Ruiz et al. [125] present DreamBooth, a fine-tuning approach that personalizes text-to-image diffusion models to generate novel renditions of subjects in varying contexts using a small reference set. This method, empowered by a class-specific prior preservation loss, maintains the subject's defining features across different scenes, opening the door to new applications like subject recontextualization and artistic rendering. These studies collectively illustrate the evolving landscape of design and application within diffusion-based methods. They highlight the progress from generative tasks to refined method design and the ongoing pursuit of enhanced controllability in image generation.

Content Editing. The design and application of diffusion-based methods have paved the way for breakthroughs in content editing, offering enhanced photorealism and greater control in the text-guided synthesis and manipulation of images. Nichol et al. [113] delve into text-conditional image generation using diffusion models, contrasting CLIP guidance with classifier-free guidance. The latter is favored for producing realistic images that closely align with human expectations. Their 3.5 billion parameter model outperforms DALL-E in human evaluations, and further demonstrates its flexibility in image inpainting, facilitating text-driven editing capabilities. Hertz

et al. [58] introduce an intuitive image editing framework, where modifications are steered solely by textual prompts, bypassing the need for spatial masks. Their analysis highlights the crucial role of cross-attention layers in mapping text to image layout, enabling precise control over local and global edits while preserving fidelity to the original content. Kumari et al. [78] propose an efficient approach for incorporating user-defined concepts into text-to-image diffusion models, Custom Diffusion. By optimizing a subset of parameters, the method allows for rapid adaptation to new concepts and the combination of multiple concepts, yielding high-quality images that outperform existing methods in both efficiency and effectiveness. Brooks et al. [13] present InstructPix2Pix, a conditional diffusion model trained on a dataset generated by combining the expertise of GPT-3 and Stable Diffusion. This model can interpret human-written instructions to edit images accurately, operating swiftly without needing per-example fine-tuning, showcasing its proficiency across a wide array of editing tasks. Last, Parmar et al. [116] tackle the challenge of content preservation in image-to-image translation with pix2pix-zero. Through the discovery of editing directions in text embedding space and cross-attention guidance, their method ensures the input image's content remains intact. They further streamline the process with a distilled conditional GAN, achieving superior performance in both real and synthetic image editing without necessitating additional training. Collectively, these advancements in diffusion-based methods signify a transformative period in content editing, where the synthesis of images is becoming increasingly controllable, customizable, and responsive to textual nuance, greatly expanding the potential for creative expression and practical applications.

Stylization. Recent advancements in diffusion-based methods have significantly enhanced the stylization capabilities in the domain of generative AI, enabling more intuitive and precise artistic expression. Zhang et al. [178] propose an inversion-based style transfer technique that captures the artistic style directly from a single painting, circumventing the need for complex textual descriptions. This method, named *InST*, efficiently captures the essence of a painting's style through a learnable textual description and applies it to guide the synthesis process, thus achieving high-quality style transfer across diverse artistic works. Huang et al. [67] present DiffStyler, a novel architecture that leverages dual diffusion processes to control the balance between content and style during text-driven image stylization. By integrating cross-modal style information as guidance and proposing a content image-based learnable noise, DiffStyler ensures that the structural integrity of the content image is maintained while achieving a compelling style transformation. In the realm of artistic image synthesis, Ahn et al. [2] propose DreamStyler, a framework that optimizes multistage textual embedding with context-aware text prompts. DreamStyler excels at both text-to-image synthesis and style transfer, providing the flexibility to adapt to various style references and producing images that exhibit high-quality and unique artistic traits. Sohn et al. [136] develop StyleDrop, a method designed to synthesize images that adhere closely to a specific style using a text-to-image model. StyleDrop stands out for its ability to capture intricate style nuances with minimal parameter fine-tuning. It demonstrates impressive results even when provided with a single image, effectively synthesizing styles across different patterns, textures, and materials. Together, these methodologies exemplify the ongoing innovation in the field of image stylization through diffusion-based methods. They afford users an unprecedented level of control and flexibility in generating and editing images, breaking new ground in the creation of stylized artistic content. These tools not only facilitate the expression of visual art but also promise to expand the possibilities for personalized and creative digital media.

Quality Enhancement. The exploration of diffusion-based methods has led to significant enhancements in the quality of text-to-image synthesis, pushing the boundaries of resolution, fidelity, and customization. Balaji et al. [6] propose eDiff-I, an ensemble of text-to-image diffusion

models that specialize in different stages of the image synthesis process. This approach results in images that better align with the input text while maintaining visual quality. The models use various embeddings for conditioning and introduce a "paint-with-words"feature, which allows users to control the output by applying words to specific areas of an image canvas, providing a more intuitive way to craft images. Chang et al. [20] introduce Muse, a Transformer model that surpasses diffusion and autoregressive models in efficiency. Muse achieves state-of-the-art performance with a masked modeling task on discrete tokens, informed by text embedding from a large pretrained language model. This method allows for fine-grained language understanding and diverse image editing applications without additional fine-tuning, such as inpainting and mask-free editing. In the realm of cost-effective and environmentally conscious training, Chen et al. [22] present PIXART-α, a Transformer-based diffusion model that significantly reduces training time and costs while maintaining competitive image quality. Through a decomposed training strategy, efficient text-to-image Transformer design, and higher informative data, PIXART-α demonstrates superior speed, saving resources and minimizing CO2 emissions. It provides a template for startups and the AI community to build high-quality, low-cost generative models. Last, He et al. [57] delve into higher-resolution visual generation with ScaleCrafter, an approach that addresses the challenges of object repetition and structure in images created at resolutions beyond those of the training datasets. By re-dilating convolutional perception fields and implementing dispersed convolution and noise-damped classifier-free guidance, ScaleCrafter enables the generation of ultra-high-resolution images without additional training or optimization, setting a new standard for texture detail and resolution in synthesized images. Collectively, these advancements represent a paradigm shift in the quality enhancement of diffusion-based generative models, offering innovative solutions to meet the ever-growing demands for high-quality, customizable, and efficient image generation and editing in the AI-powered creative landscape.

Specialized Tasks. The design and application of diffusion-based methods have extended into specialized tasks, revealing both the potential and the challenges associated with these powerful generative tools. Somepalli et al. [137] raise concerns about the originality of the content produced by diffusion models, particularly questioning whether these models generate unique art or merely replicate existing training data. Through image retrieval frameworks, they analyze content replication rates in models like Stable Diffusion and stress the significance of diverse and extensive training sets to mitigate direct copying. Zhang et al. [177] tackle the limitation of personalizing specific visual attributes in generative models. Introducing ProSpect, they utilize the stepwise generation process of diffusion models to represent images with inverted textual token embeddings, corresponding to different stages of image synthesis. This method enhances disentanglement and controllability, enabling attribute-aware personalization in image generation without the need for fine-tuning the diffusion models. In the realm of vector graphics, Jain et al. [68] demonstrate that text-conditioned diffusion models trained on pixel representations can be adapted to produce SVG-format vector graphics. Through Score Distillation Sampling loss and a differentiable vector graphics rasterizer, VectorFusion abstracts semantic knowledge from pretrained diffusion models, yielding coherent vector graphics suitable for scalable design applications. Zhang and Agrawala [173] introduce LayerDiffusion, an innovative approach that equips large-scale pretrained latent diffusion models with the capability to generate transparent images and image layers. By incorporating "latent transparency" into the model's latent space, LayerDiffusion maintains the quality of the original diffusion model while enabling transparency, facilitating applications like layer generation and structural content control. These specialized applications of diffusion-based methods highlight the versatility of generative AI, addressing the need for authenticity in digital art, personalization of visual attributes, scalability in design formats, and transparency in image layers. As these technologies advance, they promise to

Table 3. Diffusion-Based Method Categorization by Different Tasks

Task	Example Methods
Controllable Generation	ILVR [28], Textual Inversion [49], ControlNet [174], Uni-ControlNet [183], DreamBooth [125], RPG framework [164], PHDiffusion [98], etc.
Content Editing	GLIDE [113], Prompt-to-Prompt [58], Custom Diffusion [78], InstructPix2Pix [13], pix2pix-zero [116], DiT4Edit [46], etc.
Stylization	InST [178], DiffStyler [67], DreamStyler [2], StyleDrop [136], DEADiff [118], etc.
Quality Enhancement	eDiff-I [6], Muse [20], PIXART-α [22], ScaleCrafter [57], etc.
Other Specialized Tasks	ProSpect [177], VectorFusion [68], LayerDiffusion [173], Diffusion Model Originality [137], DS-Fusion [141], Dynamic Typography [97], MagicColor [179], etc.

reshape the landscape of digital content creation, offering tools that can adapt to an array of specialized tasks while preserving the integrity and quality of the generated materials.

In Table 3, we summarize the discussed research and provide more examples, to establish correspondence between different generative tasks and methods.

6.4.2 Method Classification by Diffusion Model Structure. Based on Section 4.2.1 and Figure 2, we classify different methods to design or refine diffusion-based models by a unified model structure and summarize representative methods to optimize each module.

Encoder-Decoder. Lu et al. [98] innovate with a dual encoder setup in their PHDiffusion model for painterly image harmonization, which features a lightweight adaptive encoder and a DEF (Dual Encoder Fusion) module, allowing for a more nuanced manipulation of foreground features to blend photographic objects into paintings seamlessly. Yang et al. [164] push the boundaries of text-to-image diffusion models by introducing the RPG framework, which leverages the complex reasoning capabilities of multimodal large language models. This model employs a global planner that decomposes the image generation task into subtasks, enhancing the model's ability to handle prompts with multiple objects and intricate relationships.

Denoiser-Noise Predictor. Liu et al. [94] propose a compositional visual generation technique that interprets diffusion models as energy-based models. This allows for the combination of multiple diffusion processes, each representing different components of an image, enabling the generation of scenes with a level of complexity not encountered during training. Bar-Tal et al. [7] present MultiDiffusion, a framework that fuses multiple diffusion paths for controlled image generation. The key innovation lies in its optimization task that allows for high-quality, diverse image output without requiring retraining or fine-tuning. Chefer et al. [21] introduce an attention-based semantic guidance system, Attend-and-Excite, for text-to-image diffusion models. This method refines the cross-attention units during inference time, ensuring that generated images more faithfully represent the text prompt's content. Cao et al. [15] develop a tuning-free image synthesis and editing approach, MasaCtrl, which transforms self-attention in diffusion models into mutual self-attention. This allows for consistent generation and editing by querying correlated local content and texture from source images.

Additional Modules. Hu et al. [64] innovate in the adaptation of large language models through LoRA, which introduces low-rank matrices into the Transformer architecture, significantly reducing the number of trainable parameters required for downstream tasks. Mou et al. [109] create T2I-Adapters, specialized modules that enhance the controllability of text-to-image models. These adapters tap into the models' implicit knowledge for more nuanced control over the generation outputs, emphasizing color and structure without retraining the entire model.

Table 4. Diffusion-Based Method Categorization by Different Modules

Module	Example Methods
Encoder-Decoder	Textual Inversion [49], DreamBooth [125], GLIDE [113], InstructPix2Pix [13], Muse [20], RPG framework [164], PHDiffusion [98], etc.
Denoiser-Noise Predictor	ILVR [28], Compositional Generation [94], MultiDiffusion [7], Custom Diffusion [78], Attend-and-Excite [21], MasaCtrl [15], PIXART-α [22] eDiff-I [6], etc.
Additional Modules	LoRA [64], T2I-Adapters [109], ControlNet [174], Uni-ControlNet [183], ProSpect [177], VectorFusion [68], LayerDiffusion [173], etc.

Table 5. Applications of Different Modules in Generative Tasks

Module	Conditional Generation	Content Editing	Stylization	Quality and Others
Encoder-Decoder	Textual Inversion [49], DreamBooth [125]	GLIDE [113], InstructPix2Pix [13]	InST [178], InstaStyle [34]	ScaleCrafter [57], Muse [20]
Denoiser-Noise Predictor	ILVR [28], Attend-and-Excite [21], Custom Diffusion [78]	Prompt-to-Prompt [58], MasaCtrl [15], DiT4Edit [46]	StyleDrop [136], DiffStyler [67]	ScaleCrafter [57], PIXART-α [22], eDiff-I [6]
Additional Modules	LoRA [64], ControlNet [174]	pix2pix-zero [116], T2I-Adapters [109]	DEADiff [118]	VectorFusion [68], LayerDiffusion [173]

Table 4 illustrates method categorization by model structure. Each of these innovations contributes significantly to the design and application of diffusion models, enhancing their capacity for a range of generative tasks with improved efficiency, control, and output quality.

6.4.3 Summary and Trend Identification. Following the previous illustration of visual art generative tasks, methods, and diffusion-based model structures, we form them into Table 5 and discuss how they manifest features and trends in diffusion-based method design.

This table has multiple implications. In designing a method for specific generative tasks, we may start from a column and select different corresponding modules to test their performance. We may also combine modules from different columns, which may help us accomplish multiple tasks simultaneously. However, we summarize the following trends in adapting diffusion modules and designing methods for visual art creation:

— *Integration of attention mechanisms*: The trend towards incorporating sophisticated attention mechanisms [143, 170] within generative models is evident, allowing for more detailed and contextually relevant image generation [15, 21].
— *Enhanced personalization and fine-tuning*: Techniques for fine-tuning pretrained models to adapt to specific styles, subjects, or user preferences with minimal computation are gaining traction [49, 125].
— *Control and precision in content generation*: Methods have been developed for precise control over layout, style, and content, indicating an increased focus on user-guided generation [28, 174].
— *Quality enhancement through advanced training and loss functions*: Innovations in training strategies and loss function designs aim to produce high-fidelity outputs [6, 57, 152].
— *Modularity and composability in model design*: The design of modular and composable components reflects a trend toward more adaptable generative systems [64, 109, 183].

— *Multitask and multimodal generative models*: The development of models capable of handling multiple tasks or modalities points to a trend toward versatile models [13, 78, 177].
— *Efficiency and scalability*: Innovations in model architecture aim to enhance the generation process while ensuring computational efficiency [22, 68].

7 Discussion

In this section, we focus on the frontiers, trends, and future work of diffusion-based visual art creation (Q4). Specifically, we adopt a technical and synergistic perspective to better characterize the multidimensional essence of this interdisciplinary field. In this way, we aim to shed light on emerging topics and possible future developments to provide inspiration and guidance for scientific researchers, artistic practitioners, and the whole community (G2).

7.1 Breaking the Fourth Wall: A Technical Perspective

The first trend is facilitating the creation of a more artistic, controllable, and realistic environment through the transcendence of dimensions. Researchers combine higher-dimension visual content and more diverse modalities with advanced computational power to create an immersive experience.

7.1.1 Higher Dimension. The first series of works revolve around the innovative integration of AI with 3D artistic expression and scene generation. Among them, ARF [172] presents a method to transfer artistic features from a 2D style image to a 3D scene, using a radiance field representation that overcomes geometric reconstruction errors found in previous techniques. It introduces a nearest neighbor-based loss for capturing style details and a deferred back-propagation method, optimizing memory-intensive radiance fields and improving the visual quality of stylized scenes. CoARF [169] builds on this by introducing a novel algorithm for controllable 3D scene stylization. It offers fine-grained control over the style transfer process using segmentation masks with label-dependent loss functions, and a semantic-aware nearest neighbor matching algorithm, achieving superior style transfer quality. Instruct-NeRF2NeRF [56] proposes a method for editing NeRF scenes with text instructions, using an image-conditioned diffusion model to achieve realistic targeted edits. The technique allows large-scale, real-world scene edits, expanding the possibilities for user-driven 3D content creation. DreamWire [120] presents an AI system for crafting multiview wire art, using a combination of 3D Bézier curves, Prim's algorithm, and knowledge distillation from diffusion models. This system democratizes the creation of multiview wire art, making it accessible to nonexperts while ensuring visual aesthetics. Last, RealmDreamer [135] introduces a technique for text-driven 3D scene generation using 3D Gaussian Splatting and image-conditional diffusion models. It uniquely generates high-quality 3D scenes in diverse styles without the need for video or multiview data, showcasing the potential for 3D synthesis from single images. Together, these papers advance the fusion of generative AI and 3D art, enabling new levels of creativity and control in artistic content creation.

7.1.2 Diverse Modalities. The second series of works showcase innovations that span across visual and auditory domains, aligning technologies with the nuanced dynamics of human perception and artistic creation. The Human-Art dataset [71] addresses a void in computer vision by collating 50k images from both natural and artificial human depictions across 20 different scenarios, marking a leap forward in human pose estimation and image generation tasks. This versatile dataset, with more than 123k annotated person instances in both 2D and 3D, stands to offer new insights and research directions as it bridges the gap between natural and artificial scenes. SonicDiffusion [11] introduces an audio-driven approach to image generation and editing, leveraging the multimodal aspect of human perception. By translating audio features into

tokens compatible with diffusion models, and incorporating audio-image cross-attention layers, SonicDiffusion demonstrates superior performance in creating and editing images conditioned on auditory inputs. Sprite-from-sprite [175] unravels the complexity of cartoon animations by decomposing them into basic "sprites" using a pioneering self-supervised framework that leverages Pixel MLPs. This method cleverly simplifies the decomposition of intricate animated content by first resolving simpler sprites, thus easing the overall process and enhancing the quality of cartoon animation analysis. WonderJourney [166] transforms scene generation by introducing a modularized framework designed to create a perpetual sequence of diverse and interconnected 3D scenes from any starting point, be it a textual description or an image. This approach yields imaginative and visually diverse scene sequences, showcasing the framework's robust versatility. Last, Intelligent-paint [150] propels the generation of Chinese painting processes forward. Utilizing a ViT-based generator, adversarial learning, and loss constraints that adhere to the characteristics of Chinese painting, this method vastly improves the plausibility and clarity of intermediate painting steps. The approach not only successfully bridges the gap between generated sequences and real painting processes but also serves as a valuable learning tool for novices in the art of Chinese painting. Collectively, these contributions present a multifaceted view of the convergence between AI technologies and the arts, pushing the boundaries of what can be achieved in terms of human-centric data analysis, multimodal synthesis, and artistic process generation.

7.2 "1 + 1 > 2": A Synergistic Perspective

The second trend is to promote human and AI's understanding and collaboration with each other, and finally to unleash human potential and stimulate creativity in diffusion-based visual art creation. Research under this topic is mostly understanding oriented and application driven, including creative system design, multiple intuitive interactions, content reception, and modality alignment. We summarize different approaches and solutions for the problems and tasks.

7.2.1 Interactive Systems. Emerging research showcases interactive technologies that amalgamate human intuition with AI's capabilities to enhance the process of creation across various artistic domains. PromptPaint [29] revolutionizes text-to-image models by allowing users to intuitively guide image generation through paint-medium-like interactions, akin to mixing colors on a palette. This system enables the iterative application of prompts to canvas areas, enhancing the user's ability to shape outputs in ways that language alone could not facilitate. Collaborative Neural Painting [35] introduces the task of joint art creation between humans and AI. Through a novel Transformer-based architecture that models the input and completion strokes, users can iteratively shape the artwork, making the painting process both creative and collaborative. ArtVerse [54] proposes a human-machine collaborative creation paradigm in the metaverse, where AI participates in artistic exploration and evolution, shaping a decentralized ecosystem for art creation, dissemination, and transaction. ARtVista [63] empowers individuals to bridge the gap between conceptual ideas and their visual representation. By integrating AR and generative AI, ARtVista assists users in creating sketches from abstract thoughts and generating vibrant paintings in diverse styles. Its unique paint-by-number simulation further simplifies the artistic process, enabling anyone to produce stunning artwork without advanced drawing skills. The Interactive3D [40] framework elevates 3D object generation by granting users unparalleled control over the creation process. Utilizing Gaussian Splatting and InstantNGP representations, this framework allows for comprehensive interaction, including adding, removing, transforming, and detailed refinement of 3D components, pushing the boundaries of precision in generative 3D modeling. Finally, Neural Canvas [134] integrates generative AI into a 3D sketching interface, facilitating scenic design prototyping. It transcends the limitations of traditional tools by enabling

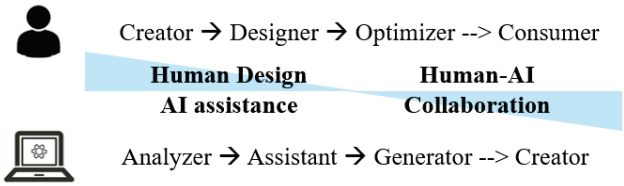

Fig. 7. New perspectives on the emerging frontiers and future work of diffusion-based visual art creation. From a synergistic perspective, we inspect a continuous paradigm shift in the roles of human and AI.

rapid iteration of visual ideas and atmospheres in 3D space, expediting the design process for both novices and professionals. These contributions collectively demonstrate a synergistic approach where the sum of human and machine efforts yields greater creative outcomes than either could achieve independently, marking a new era in interactive and generative art making.

7.2.2 Reception and Alignment. Another series of works focus on latent space disentanglement and multimodality alignment, combining the perspectives of content reception and generation for understanding, to enable human and AI to better understand each other. For example, a study on multisensory experience [27] emphasizes the potential of various sensory elements such as sound, touch, and smell to convey visual artwork elements to visually impaired individuals. By leveraging patterns, temperature, and other sensory cues, this research opens up new avenues for inclusive art appreciation and paves the way for further exploration in multisensory interfaces. In the realm of knowledgeable art description, a new framework [5] has been introduced for generating rich descriptions of paintings that cover artistic styles, content, and historical context. This multitopic approach, augmented with external knowledge, has been shown to successfully capture diverse aspects of artwork and its creation, enhancing the viewer's understanding and engagement with art. Initial Images [119] explores the use of image prompts alongside text to improve the subject representation in AI-generated art. This research demonstrates how image prompts can exert significant control over final compositions, leading to more accurate and user-aligned creations. CLIP-PAE [186] addresses the challenge of disentangled and interpretable text-guided image manipulation by introducing projection-augmentation embedding. This method refines the alignment between text and image features, enabling more precise and controllable manipulations, particularly demonstrated in the context of facial editing. Evaluating text-to-visual generation has been advanced with the introduction of VQAScore [90], a metric that utilizes a visual-question-answering model to assess image-text alignment. This approach offers a more nuanced evaluation of complex prompts and has led to the creation of GenAI-Bench, a benchmark for rigorously testing generative AI models against compositional text prompts. Collectively, these contributions signify a synergistic advancement where the combination of multiple approaches, senses, and technologies results in a more profound and aligned interaction between AI and human perception, pushing the boundaries of art creation, appreciation, and evaluation.

Figure 7 illustrates the paradigm shift in human and AI roles in content creation. With technological advancements, human roles in AIGC are shifting from creators to optimizers to consumers, whereas AI develops from analyzers to generators to creators. In all, the creative paradigm shifts from human design and AI assistance to human-AI collaboration, where the two counterparts learn from and inspire each other.

8 Conclusion

This survey has charted the course of diffusion-based generative methods within the rich terrain of visual art creation. We began by identifying the research scope, pinpointing diffusion models

and visual art as pivotal concepts, and outlining our dual research goals and the quartet of research questions. A robust dataset was assembled, encompassing relevant papers that underwent a rigorous four-phase filtering process, leading to their categorization across seven dimensions within the thematic angles of application, understanding, and generation. Through a blend of structural and temporal analysis, we discovered prevailing trends and constructed a comprehensive analytical framework. Our synthesis of findings crystallized into a paradigm, encompassing the quadrants of scenario, modality, task, and method, which collectively shape the nexus of diffusion-based visual art creation. As we gaze into the future, we propose a novel perspective that intertwines technological and synergistic aspects, characterizing the collaborative ventures between humans and AI in creating visual art. This perspective beckons a future where AI not only complements human artistry but also actively contributes to the creative process. However, amidst the rapid technical strides, we are compelled to ponder the implications of AI potentially surpassing human capacity in both understanding and task execution. In such a scenario, we are prompted to question the pursuits we should embrace. If human aspirations and desires continue to expand, how can AI evolve to meet these ever-growing needs? How can we ensure that the evolution of AI in visual art creation remains aligned with human values and creative aspirations? In conclusion, while we acknowledge the remarkable progress made thus far, this survey also serves as a clarion call to the research community. As the horizon of AI in visual art creation broadens, we must continue to explore, innovate, and critically reflect on the role of AI in this field. The future beckons with a promise of AI that not only mimics but enriches human creativity, forming an indelible part of our artistic and cultural expression.

Acknowledgments

The authors extend gratitude to the anonymous reviewers for their expert critique. Special acknowledgment is accorded to You Zhou (Computational Media and Arts, HKUST(GZ)) for graphic refinement.

References

[1] Nilin Abrahamsen and Jiahao Yao. 2023. Inventing art styles with no artistic training data. *arXiv preprint arXiv:2305.12015* (2023). https://api.semanticscholar.org/CorpusID:258833473

[2] Namhyuk Ahn, Junsoo Lee, Chunggi Lee, Kunhee Kim, Daesik Kim, Seung-Hun Nam, and Kibeom Hong. 2024. *DreamStyler*: Paint by style inversion with text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 674–681.

[3] Hao Ai and Lu Sheng. 2023. Stable diffusion reference only: Image prompt and blueprint jointly guided multi-condition diffusion model for secondary painting. *arXiv preprint arXiv:2311.02343* (2023).

[4] Qingyan Bai, Yinghao Xu, Zifan Shi, Hao Ouyang, Qiuyu Wang, Ceyuan Yang, Xuan Wang, Gordon Wetzstein, Yujun Shen, and Qifeng Chen. 2024. Real-time 3D-aware portrait editing from a single image. *arXiv preprint arXiv:2402.14000* (2024).

[5] Zechen Bai, Yuta Nakashima, and Noa Garcia. 2021. Explain me the painting: Multi-topic knowledgeable art description generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5422–5432.

[6] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. 2022. EDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324* (2022).

[7] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. 2023. MultiDiffusion: Fusing diffusion paths for controlled image generation. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*. 1737–1752.

[8] Siwar Bengamra, Olfa Mzoughi, André Bigand, and Ezzeddine Zagrouba. 2024. A comprehensive survey on object detection in visual art: Taxonomy and challenge. *Multimedia Tools and Applications* 83, 5 (2024), 14637–14670.

[9] Ardavan Bidgoli, Manuel Ladron De Guevara, Cinnie Hsiung, Jean Oh, and Eunsu Kang. 2020. Artistic style in robotic painting: A machine learning approach to learning brushstroke from human artists. In *Proceedings of the 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN'20)*. IEEE, 412–418.

[10] Fengxiang Bie, Yibo Yang, Zhongzhu Zhou, Adam Ghanem, Minjia Zhang, Zhewei Yao, Xiaoxia Wu, Connor Holmes, Pareesa Golnari, David A. Clifton, et al. 2023. RenAIssance: A survey into AI text-to-image generation in the era of large model. *arXiv preprint arXiv:2309.00810* (2023).

[11] Burak Can Biner, Farrin Marouf Sofian, Umur Berkay Karakaş, Duygu Ceylan, Erkut Erdem, and Aykut Erdem. 2024. SonicDiffusion: Audio-driven image generation and editing with pretrained diffusion models. *arXiv preprint arXiv:2405.00878* (2024).

[12] Tom Braude, Idan Schwartz, Alex Schwing, and Ariel Shamir. 2022. Ordered attention for coherent visual storytelling. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3310–3318.

[13] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18392–18402.

[14] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712* (2023).

[15] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. 2023. MASACTRL: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22560–22570.

[16] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun. 2023. A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT. *arXiv preprint arXiv:2303.04226* (2023).

[17] Giovanna Castellano and Gennaro Vessio. 2021. A brief overview of deep learning approaches to pattern extraction and recognition in paintings and drawings. In *Proceedings of the International Conference on Pattern Recognition*. 487–501.

[18] Giovanna Castellano and Gennaro Vessio. 2021. Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview. *Neural Computing and Applications* 33, 19 (2021), 12263–12282.

[19] Eva Cetinic and James She. 2022. Understanding and creating art with AI: Review and outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications* 18, 2 (2022), 1–22.

[20] Huiwen Chang, Han Zhang, Jarred Barber, A. J. Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T. Freeman, Michael Rubinstein, et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704* (2023).

[21] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-Excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics* 42, 4 (2023), 1–10.

[22] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. 2023. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426* (2023).

[23] Minshuo Chen, Song Mei, Jianqing Fan, and Mengdi Wang. 2024. An overview of diffusion models: Applications, guided generation, statistical rates and optimization. *arXiv preprint arXiv:2404.07771* (2024).

[24] Weifeng Chen, Jiacheng Zhang, Jie Wu, Hefeng Wu, Xuefeng Xiao, and Liang Lin. 2024. ID-Aligner: Enhancing identity-preserving text-to-image generation with reward feedback learning. *arXiv preprint arXiv:2404.15449* (2024).

[25] Xiaojiao Chen, Zhebing Hu, and Chengliang Wang. 2024. Empowering education development through AIGC: A systematic literature review. *Education and Information Technologies* 29 (2024), 17485–17537.

[26] Pei-Ying Chiang, Chun-Von Lin, and Cheng-Hua Tseng. 2018. Generation of Chinese ink portraits by blending face photographs with Chinese ink paintings. *Journal of Visual Communication and Image Representation* 52 (2018), 33–44.

[27] Jun Dong Cho. 2021. A study of multi-sensory experience and color recognition in visual arts appreciation of people with visual impairment. *Electronics* 10, 4 (2021), 470.

[28] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. 2021. ILVR: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938* (2021).

[29] John Joon Young Chung and Eytan Adar. 2023. PromptPaint: Steering text-to-image generation through paint medium-like interactions. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–17.

[30] Shen Ciao, Zhongyue Guan, Qianxi Liu, Li-Yi Wei, and Zeyu Wang. 2024. Ciallo: GPU-accelerated rendering of vector brush strokes. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Papers'24 (SIGGRAPH Conference Papers'24)*. ACM, New York, NY, USA, 1–11. https://doi.org/10.1145/3641519.3657418

[31] Dario Cioni, Lorenzo Berlincioni, Federico Becattini, and Alberto Del Bimbo. 2023. Diffusion based augmentation for captioning and retrieval in cultural heritage. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1707–1716.

[32] Xiaoyan Cong, Yue Wu, Qifeng Chen, and Chenyang Lei. 2024. Automatic controllable colorization via imagination. *arXiv preprint arXiv:2404.05661* (2024).

[33] Guillaume Couairon, Marlène Careil, Matthieu Cord, Stéphane Lathuilière, and Jakob Verbeek. 2023. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2174–2183.
[34] Xing Cui, Zekun Li, Peipei Li, Huaibo Huang, Xuannan Liu, and Zhaofeng He. 2024. INSTASTYLE: Inversion noise of a stylized image is secretly a style adviser. In *Proceedings of the European Conference on Computer Vision*. 455–472.
[35] Nicola Dall'Asen, Willi Menapace, Elia Peruzzo, Enver Sangineto, Yiming Wang, and Elisa Ricci. 2023. Collaborative neural painting. *arXiv preprint arXiv:2312.01800* (2023).
[36] Sebastiano D'Amico, Valentina Venuti, Emanuele Colica, Vincenza Crupi, Giuseppe Paladini, Sante Guido, Giuseppe Mantella, and Domenico Majolino. 2021. A combined 3D surveying, XRF and Raman in-situ investigation of *The Conversion of St Paul* painting (Mdina, Malta) by Mattia Preti. *Acta IMEKO* 10, 1 (2021), 173–179.
[37] Nassim Dehouche and Kullathida Dehouche. 2023. What's in a text-to-image prompt? The potential of stable diffusion in visual arts education. *Heliyon* 9, 6 (2023), e16757.
[38] Gwen DePolo, Marc Walton, Katrien Keune, and Kenneth R. Shull. 2021. After the paint has dried: A review of testing techniques for studying the mechanical properties of artists' paint. *Heritage Science* 9 (2021), 1–24.
[39] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 8780–8794.
[40] Shaocong Dong, Lihe Ding, Zhanpeng Huang, Zibin Wang, Tianfan Xue, and Dan Xu. 2024. Interactive3D: Create what you want by interactive 3D generation. *arXiv preprint arXiv:2404.16510* (2024).
[41] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
[42] Denis Dutton. 2009. *The Art Instinct: Beauty, Pleasure, & Human Evolution*. Oxford University Press, USA.
[43] Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. 2017. CAN: Creative adversarial networks, generating "Art" by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068* (2017).
[44] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12873–12883.
[45] Xiao-Nan Fang, Bin Liu, and Ariel Shamir. 2018. Automatic thread painting generation. *arXiv preprint arXiv:1802.04706* (2018).
[46] Kunyu Feng, Yue Ma, Bingyuan Wang, Chenyang Qi, Haozhe Chen, Qifeng Chen, and Zeyu Wang. 2024. Dit4Edit: Diffusion transformer for image editing. *arXiv preprint arXiv:2411.03286* (2024).
[47] Lin Geng Foo, Hossein Rahmani, and Jun Liu. 2023. AI-generated content (AIGC) for various data modalities: A survey. *arXiv preprint arXiv:2308.14177* 2 (2023).
[48] Feifei Fu, Jiancheng Lv, Chenwei Tang, and Mao Li. 2021. Multi-style Chinese art painting generation of flowers. *IET Image Processing* 15, 3 (2021), 746–762.
[49] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618* (2022).
[50] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2414–2423.
[51] Daniel Geng, Inbum Park, and Andrew Owens. 2023. Visual anagrams: Generating multi-view optical illusions with diffusion models. *arXiv preprint arXiv:2311.17919* (2023).
[52] Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Yingqing He, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, et al. 2023. Interactive story visualization with multiple characters. In *SIGGRAPH Asia 2023 Conference Papers*. ACM, New York, NY, USA, 1–10.
[53] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems* 27 (2014), 1–9.
[54] Chao Guo, Yong Dou, Tianxiang Bai, Xingyuan Dai, Chunfa Wang, and Yi Wen. 2023. ArtVerse: A paradigm for parallel human–machine collaborative painting creation in metaverses. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 53, 4 (2023), 2200–2208.
[55] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R. Martin, Ming-Ming Cheng, and Shi-Min Hu. 2022. Attention mechanisms in computer vision: A survey. *Computational Visual Media* 8, 3 (2022), 331–368.
[56] Ayaan Haque, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. 2023. Instruct-NeRF2NeRF: Editing 3D scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 19740–19750.

[57] Yingqing He, Shaoshu Yang, Haoxin Chen, Xiaodong Cun, Menghan Xia, Yong Zhang, Xintao Wang, Ran He, Qifeng Chen, and Ying Shan. 2023. ScaleCrafter: Tuning-free higher-resolution visual generation with diffusion models. In *Proceedings of the 12th International Conference on Learning Representations*.

[58] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022).

[59] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems* 30 (2017), 6629–6640.

[60] Irina Higgins, Loic Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the the 5th International Conference on Learning Representations (ICLR'17): Poster*.

[61] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

[62] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

[63] Trong-Vu Hoang, Quang-Binh Nguyen, Duy-Nam Ly, Khanh-Duy Le, Tam V. Nguyen, Minh-Triet Tran, and Trung-Nghia Le. 2024. ARtVista: Gateway to empower anyone into artist. *arXiv preprint arXiv:2403.08876* (2024).

[64] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[65] Jingxuan Hu, Jinhuan Zhang, Liyu Hu, Haibo Yu, and Jinping Xu. 2021. Art therapy: A complementary treatment for mental disorders. *Frontiers in Psychology* 12 (2021), 686005.

[66] Nisha Huang, Fan Tang, Weiming Dong, and Changsheng Xu. 2022. Draw your art dream: Diverse digital art synthesis with multimodal guided diffusion. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1085–1094.

[67] Nisha Huang, Yuxin Zhang, Fan Tang, Chongyang Ma, Haibin Huang, Weiming Dong, and Changsheng Xu. 2025. DiffStyler: Controllable dual diffusion for text-driven image stylization. *IEEE Transactions on Neural Networks and Learning Systems* 36, 2 (2025), 3370–3383.

[68] Ajay Jain, Amber Xie, and Pieter Abbeel. 2023. VectorFusion: Text-to-SVG by abstracting pixel-based diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1911–1920.

[69] Xin Jin, Bohan Li, Baao Xie, Wenyao Zhang, Jinming Liu, Ziqiang Li, Tao Yang, and Wenjun Zeng. 2024. Closed-loop unsupervised representation disentanglement with β-VAE distillation and diffusion probabilistic feedback. *arXiv preprint arXiv:2402.02346* (2024).

[70] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. 2019. Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics* 26, 11 (2019), 3365–3385.

[71] Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. 2023. Human-ART: A versatile human-centric dataset bridging natural and artificial scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 618–629.

[72] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8110–8119.

[73] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6007–6017.

[74] Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114* (2013).

[75] Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. 2023. Large-scale text-to-image generation models for visual artists' creative works. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 919–933.

[76] Saptarshi Kolay. 2016. Cultural heritage preservation of traditional Indian art through virtual new-media. *Procedia: Social and Behavioral Sciences* 225 (2016), 309–320.

[77] Patrycja Kowalek, Hanna Loch-Olszewska, Łukasz Łaszczuk, Jarosław Opała, and Janusz Szwabiński. 2022. Boosting the performance of anomalous diffusion classifiers with the proper choice of features. *Journal of Physics A: Mathematical and Theoretical* 55, 24 (2022), 244005.

[78] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1931–1941.

[79] Jan Eric Kyprianidis, John Collomosse, Tinghuai Wang, and Tobias Isenberg. 2012. State of the "Art": A taxonomy of artistic stylization techniques for images and video. *IEEE Transactions on Visualization and Computer Graphics* 19, 5 (2012), 866–885.

[80] Margaret Lazzari and Dona Schlesier. 2008. Exploring art: A global, thematic approach. Cengage Learning.
[81] Lik-Hang Lee, Zijun Lin, Rui Hu, Zhengya Gong, Abhishek Kumar, Tangyao Li, Sijia Li, and Pan Hui. 2021. When creators meet the metaverse: A survey on computational arts. *arXiv preprint arXiv:2111.13486* (2021).
[82] Tong-Yee Lee, Shaur-Uei Yan, Yong-Nien Chen, and Ming-Te Chi. 2005. Real-time 3D artistic rendering system. In *Knowledge-Based Intelligent Information and Engineering Systems*. Lecture Notes in Computer Science, Vol. 3683. Springer, 456–462.
[83] Chenghao Li, Chaoning Zhang, Atish Waghwase, Lik-Hang Lee, Francois Rameau, Yang Yang, Sung-Ho Bae, and Choong Seon Hong. 2023. Generative AI meets 3D: A survey on text-to-3D in AIGC era. *arXiv preprint arXiv:2305.06131* (2023).
[84] Dan Li, Shuai Wang, Jie Zou, Chang Tian, Elisha Nieuwburg, Fengyuan Sun, and Evangelos Kanoulas. 2021. Paint4Poem: A dataset for artistic visualization of classical Chinese poems. *arXiv preprint arXiv:2109.11682* (2021).
[85] Hao Li, Zhongyue Guan, and Zeyu Wang. 2024. An inverse procedural modeling pipeline for stylized brush stroke rendering. In *Proceedings of the 45th Annual Conference of the European Association for Computer Graphics (EUROGRAPHICS'24)*.
[86] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. GLIGEN: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22511–22521.
[87] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. 2023. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. *arXiv preprint arXiv:2302.04578* (2023).
[88] Peiyuan Liao, Xiuyu Li, Xihui Liu, and Kurt Keutzer. 2022. The ArtBench dataset: Benchmarking generative models with artworks. *arXiv preprint arXiv:2206.11404* (2022).
[89] Troy TianYu Lin, James She, Yu-Ao Wang, and Kang Zhang. 2025. Future ink: The collision of AI and Chinese calligraphy. *ACM Journal on Computing and Cultural Heritage* 18, 1 (2025), Article 15, 17 pages.
[90] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291* (2024).
[91] Jin Liu, Huaibo Huang, Chao Jin, and Ran He. 2023. Portrait Diffusion: Training-free face stylization with chain-of-painting. *arXiv preprint arXiv:2312.02212* (2023).
[92] Lingxi Liu, Tsveta Miteva, Giovanni Delnevo, Silvia Mirri, Philippe Walter, Laurence de Viguerie, and Emeline Pouyet. 2023. Neural networks for hyperspectral imaging of historical paintings: A practical review. *Sensors* 23, 5 (2023), 2419.
[93] Mingyang Liu. 2023. Overview of artificial intelligence painting development and some related model application. In *SHS Web of Conferences*, Vol. 167. EDP Sciences, 01004.
[94] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. 2022. Compositional visual generation with composable diffusion models. In *Proceedings of the European Conference on Computer Vision*. 423–439.
[95] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. 2023. InstaFlow: One step is enough for high-quality diffusion-based text-to-image generation. In *Proceedings of the 12th International Conference on Learning Representations*.
[96] Zhengliang Liu, Yiwei Li, Qian Cao, Junwen Chen, Tianze Yang, Zihao Wu, John Hale, John Gibbs, Khaled Rasheed, Ninghao Liu, et al. 2023. Transformation vs tradition: Artificial general intelligence (AGI) for arts and humanities. *arXiv preprint arXiv:2310.19626* (2023).
[97] Zichen Liu, Yihao Meng, Hao Ouyang, Yue Yu, Bolin Zhao, Daniel Cohen-Or, and Huamin Qu. 2024. Dynamic typography: Bringing text to life via video diffusion prior. *arXiv e-prints arXiv:2404.11614* (2024).
[98] Lingxiao Lu, Jiangtong Li, Junyan Cao, Li Niu, and Liqing Zhang. 2023. Painterly image harmonization using diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*. 233–241.
[99] Denis Lukovnikov and Asja Fischer. 2024. Layout-to-image generation with localized descriptions using ControlNet with cross-attention control. *arXiv preprint arXiv:2402.13404* (2024).
[100] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 30 (2017), 4768–4777.
[101] Qiongshuai Lyu, Na Zhao, Yu Yang, Yuehong Gong, and Jingli Gao. 2024. A diffusion probabilistic model for traditional Chinese landscape painting super-resolution. *Heritage Science* 12, 1 (2024), 4.
[102] Anne-Sofie Maerten and Derya Soydaner. 2023. From paintbrush to pixel: A review of deep neural networks in AI-generated art. *arXiv preprint arXiv:2302.10913* (2023).
[103] Weijian Mai, Jian Zhang, Pengfei Fang, and Zhijun Zhang. 2023. Brain-conditional multimodal synthesis: A survey and taxonomy. *arXiv preprint arXiv:2401.00430* (2023).
[104] Marian Mazzone and Ahmed Elgammal. 2019. Art, creativity, and the potential of artificial intelligence. In *Arts*. Vol. 8. MDPI, 26.

[105] Jon McCormack and Mark D'Inverno. 2014. On the future of computers and creativity. In *Proceedings of the AISB 2014 Symposium on Computational Creativity*.

[106] Jon McCormack, Toby Gifford, and Patrick Hutchings. 2019. Autonomy, authenticity, authorship and intention in computer generated art. In *Proceedings of the International Conference on Computational Intelligence in Music, Sound, Art. and Design (Part of EvoStar)*. 35–50.

[107] Arthur I. Miller. 2019. *The Artist in the Machine: The World of AI-Powered Creativity*. MIT Press.

[108] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. AltmanPrisma Group. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine* 151, 4 (2009), 264–269.

[109] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 2024. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4296–4304.

[110] Reiichiro Nakano. 2019. Neural painters: A learned differentiable constraint for generating brushstroke paintings. *arXiv preprint arXiv:1904.08410* (2019).

[111] Jisu Nam, Heesu Kim, DongJae Lee, Siyoon Jin, Seungryong Kim, and Seunggyu Chang. 2024. DreamMatcher: Appearance matching self-attention for semantically-consistent text-to-image personalization. *arXiv preprint arXiv:2402.09812* (2024).

[112] Seonghyeon Nam, Chongyang Ma, Menglei Chai, William Brendel, Ning Xu, and Seon Joo Kim. 2019. End-to-end time-lapse video synthesis from a single outdoor image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1409–1418.

[113] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

[114] OpenAI. 2023. Model Index for Researchers. Retrieved November 2, 2023 from https://platform.openai.com/docs/model-index-for-researchers

[115] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2337–2346.

[116] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. 2023. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

[117] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. StyleCLIP: Text-driven manipulation of StyleGAN imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2085–2094.

[118] Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. 2024. DEADiff: An efficient stylization diffusion model with disentangled representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8693–8702.

[119] Han Qiao, Vivian Liu, and Lydia Chilton. 2022. Initial images: Using image prompts to improve subject representation in multimodal AI generated art. In *Proceedings of the 14th Conference on Creativity and Cognition*. 15–28.

[120] Zhiyu Qu, Lan Yang, Honggang Zhang, Tao Xiang, Kaiyue Pang, and Yi-Zhe Song. 2023. Wired perspectives: Multi-view wire art embraces generative AI. *arXiv preprint arXiv:2311.15421* (2023).

[121] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*. 8748–8763.

[122] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125* (2022).

[123] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning*. 8821–8831.

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

[125] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510.

[126] Irene Russo. 2022. Creative text-to-image generation: Suggestions for a benchmark. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*. 145–154.

[127] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.

[128] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training GANs. *Advances in Neural Information Processing Systems* 29 (2016), 2234–2242.
[129] Andreza Sartori. 2014. Affective analysis of abstract paintings using statistical analysis and art theory. In *Proceedings of the 16th International Conference on Multimodal Interaction*. 384–388.
[130] Thorsten-Walther Schmidt, Fabio Pellacini, Derek Nowrouzezahrai, Wojciech Jarosz, and Carsten Dachsbacher. 2016. State of the art in artistic editing of appearance, lighting and material. *Computer Graphics Forum* 35 (2016), 216–233.
[131] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114* (2021).
[132] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. 2023. ZipLoRA: Any subject in any style by effectively merging LoRAs. *arXiv preprint arXiv:2311.13600* (2023).
[133] Sakib Shahriar. 2022. GAN computers generate arts? A survey on visual arts, music, and literary text generation using generative adversarial network. *Displays* 73 (2022), 102237.
[134] Yulin Shen, Yifei Shen, Jiawen Cheng, Chutian Jiang, Mingming Fan, and Zeyu Wang. 2024. Neural Canvas: Supporting scenic design prototyping by integrating 3D sketching and generative AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI'24)*. ACM, New York, NY, USA, 1–17. https://doi.org/10.1145/3613904.3642096
[135] Jaidev Shriram, Alex Trevithick, Lingjie Liu, and Ravi Ramamoorthi. 2024. RealmDreamer: Text-driven 3D scene generation with inpainting and depth diffusion. *arXiv preprint arXiv:2404.07199* (2024).
[136] Kihyuk Sohn, Lu Jiang, Jarred Barber, Kimin Lee, Nataniel Ruiz, Dilip Krishnan, Huiwen Chang, Yuanzhen Li, Irfan Essa, Michael Rubinstein, et al. 2024. StyleDrop: Text-to-image synthesis in any style. *Advances in Neural Information Processing Systems* 36 (2024), 66860–66889.
[137] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Diffusion art or digital forgery? Investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6048–6058.
[138] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).
[139] Junrong Song, Bingyuan Wang, Zeyu Wang, and David Kei-Man Yip. 2023. From expanded cinema to extended reality: How AI can expand and extend cinematic experiences. In *Proceedings of the 16th International Symposium on Visual Information Communication and Interaction*. 1–5.
[140] Yun-Zhu Song, Zhi Rui Tam, Hung-Jen Chen, Huiao-Han Lu, and Hong-Han Shuai. 2020. Character-preserving coherent story visualization. In *Proceedings of the European Conference on Computer Vision*. 18–33.
[141] Maham Tanveer, Yizhi Wang, Ali Mahdavi-Amiri, and Hao Zhang. 2023. DS-Fusion: Artistic typography via discriminated and stylized diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 374–384.
[142] Jaakko Turkka, Outi Haatainen, and Maija Aksela. 2017. Integrating art into science education: A survey of science teachers' practices. *International Journal of Science Education* 39, 10 (2017), 1403–1419.
[143] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017), 1–11.
[144] Bingyuan Wang, Hengyu Meng, Rui Cao, Zeyu Cai, Lanjiong Li, Yue Ma, Qifeng Chen, and Zeyu Wang. 2025. MagicScroll: Enhancing immersive storytelling with controllable scroll image generation. In *Proceedings of the 2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR'25)*. IEEE, 431–441.
[145] Bingyuan Wang, Kang Zhang, and Zeyu Wang. 2023. Naturality: A natural reflection of Chinese calligraphy. In *Proceedings of the 16th International Symposium on Visual Information Communication and Interaction*. 1–8.
[146] Bingyuan Wang, Pinxi Zhu, Hao Li, David Kei-Man Yip, and Zeyu Wang. 2023. Simonstown: An AI-facilitated interactive story of love, life, and pandemic. In *Proceedings of the 16th International Symposium on Visual Information Communication and Interaction*. 1–7.
[147] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, et al. 2023. Imagen editor and EditBench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18359–18369.
[148] Wenhao Wang, Yifan Sun, Zongxin Yang, Zhengdong Hu, Zhentao Tan, and Yi Yang. 2024. Replication in visual diffusion models: A survey and outlook. *arXiv preprint arXiv:2408.00001* (2024).
[149] Zeyu Wang. 2022. *Enhancing the Creative Process in Digital Prototyping*. Ph. D. Dissertation. Yale University.
[150] Zunfu Wang, Fang Liu, Zhixiong Liu, Changjuan Ran, and Mohan Zhang. 2024. Intelligent-Paint: A Chinese painting process generation method based on Vision Transformer. *Multimedia Systems* 30, 2 (2024), 1–17.
[151] Zeyu Wang, Sherry Qiu, Nicole Feng, Holly Rushmeier, Leonard McMillan, and Julie Dorsey. 2021. Tracing versus freehand for evaluating computer-generated drawings. *ACM Transactions on Graphics* 40, 4 (2021), 1–12.

[152] Zeyu Wang, T. Wang, and Julie Dorsey. 2022. Learning a style space for interactive line drawing synthesis from animated 3D models. In *PG2022 Short Papers, Posters, and Work-in-Progress Papers*. 1–12.
[153] Zhouxia Wang, Xintao Wang, Liangbin Xie, Zhongang Qi, Ying Shan, Wenping Wang, and Ping Luo. 2023. StyleAdapter: A single-pass LoRA-free model for stylized image generation. *arXiv preprint arXiv:2309.01770* (2023).
[154] Zhongqi Wang, Jie Zhang, Zhilong Ji, Jinfeng Bai, and Shiguang Shan. 2023. CCLAP: Controllable Chinese landscape painting generation via latent diffusion model. In *Proceedings of the 2023 IEEE International Conference on Multimedia and Expo (ICME'23)*. IEEE, 2117–2122.
[155] Xianchao Wu. 2022. Creative painting with latent diffusion models. In *Proceedings of the 2nd Workshop on When Creative AI Meets Conversational AI*. 59–80.
[156] You Wu, Kean Liu, Xiaoyue Mi, Fan Tang, Juan Cao, and Jintao Li. 2024. U-VAP: User-specified visual appearance personalization via decoupled self augmentation. *arXiv preprint arXiv:2403.20231* (2024).
[157] Yankun Wu, Yuta Nakashima, and Noa Garcia. 2023. Not only generative art: Stable diffusion for content-style disentanglement in art analysis. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*. 199–208.
[158] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. 2022. GAN inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2022), 3121–3138.
[159] Shishi Xiao, Liangwei Wang, Xiaojuan Ma, and Wei Zeng. 2024. TypeDance: Creating semantic typographic logos from image through personalized generation. *arXiv preprint arXiv:2401.11094* (2024).
[160] Minrui Xu, Hongyang Du, Dusit Niyato, Jiawen Kang, Zehui Xiong, Shiwen Mao, Zhu Han, Abbas Jamalipour, Dong In Kim, Xuemin Shen, et al. 2024. Unleashing the power of edge-cloud generative AI in mobile networks: A survey of AIGC services. *IEEE Communications Surveys & Tutorials* 26, 2 (2024), 1127–1170.
[161] Youze Xue, Binghui Chen, Yifeng Geng, Xuansong Xie, Jiansheng Chen, and Hongbing Ma. 2024. Strictly-ID-preserved and controllable accessory advertising image generation. *arXiv preprint arXiv:2404.04828* (2024).
[162] Moyuru Yamada. 2024. GLoD: Composing global contexts and local details in image generation. *arXiv preprint arXiv:2404.15447* (2024).
[163] Jingyuan Yang, Jiawei Feng, and Hui Huang. 2024. EmoGen: Emotional image content generation with text-to-image diffusion models. *arXiv preprint arXiv:2401.04608* (2024).
[164] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. 2024. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal LLMs. *arXiv preprint arXiv:2401.11708* (2024).
[165] Yilin Ye, Rong Huang, Kang Zhang, and Wei Zeng. 2023. Everyone can be Picasso? A computational framework into the myth of human versus AI painting. *arXiv preprint arXiv:2304.07999* (2023).
[166] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T. Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. 2023. WonderJourney: Going from anywhere to everywhere. *arXiv preprint arXiv:2312.03884* (2023).
[167] Xingchen Zeng, Ziyao Gao, Yilin Ye, and Wei Zeng. 2024. IntentTuner: An interactive framework for integrating human intents in fine-tuning text-to-image generative models. *arXiv preprint arXiv:2401.15559* (2024).
[168] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. 2023. Text-to-image diffusion model in generative AI: A survey. *arXiv preprint arXiv:2303.07909* (2023).
[169] Deheng Zhang, Clara Fernandez-Labrador, and Christopher Schroers. 2024. CoARF: Controllable 3D artistic style transfer for radiance fields. *arXiv preprint arXiv:2404.14967* (2024).
[170] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2019. Self-attention generative adversarial networks. In *Proceedings of the International Conference on Machine Learning*. 7354–7363.
[171] Jiajing Zhang, Yongwei Miao, and Jinhui Yu. 2021. A comprehensive survey on computational aesthetic evaluation of visual art images: Metrics and challenges. *IEEE Access* 9 (2021), 77164–77187.
[172] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. 2022. ARF: Artistic radiance fields. In *Proceedings of the European Conference on Computer Vision*. 717–733.
[173] Lvmin Zhang and Maneesh Agrawala. 2024. Transparent image layer diffusion using latent transparency. *arXiv preprint arXiv:2402.17113* (2024).
[174] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.
[175] Lvmin Zhang, Tien-Tsin Wong, and Yuxin Liu. 2022. Sprite-from-sprite: Cartoon animation decomposition with self-supervised sprite estimation. *ACM Transactions on Graphics* 41, 6 (2022), 1–12.
[176] Tianyi Zhang, Zheng Wang, Jing Huang, Mohiuddin Muhammad Tasnim, and Wei Shi. 2023. A survey of diffusion-based image generation models: Issues and their solutions. *arXiv preprint arXiv:2308.13142* (2023).
[177] Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. 2023. Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. *ACM Transactions on Graphics* 42, 6 (2023), 1–14.

[178] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2023. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10146–10156.

[179] Yinhan Zhang, Yue Ma, Bingyuan Wang, Qifeng Chen, and Zeyu Wang. 2025. MagicColor: Multi-instance sketch colorization. *arXiv preprint arXiv:2503.16948* (2025).

[180] Zhanjie Zhang, Quanwei Zhang, Huaizhong Lin, Wei Xing, Juncheng Mo, Shuaicheng Huang, Jinheng Xie, Guangyuan Li, Junsheng Luan, Lei Zhao, et al. 2024. Towards highly realistic artistic style transfer via stable diffusion with step-aware and layer-aware prompt. *arXiv preprint arXiv:2404.11474* (2024).

[181] Zhanjie Zhang, Quanwei Zhang, Wei Xing, Guangyuan Li, Lei Zhao, Jiakai Sun, Zehua Lan, Junsheng Luan, Yiling Huang, and Huaizhong Lin. 2024. ArtBank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 7396–7404.

[182] Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. 2024. Retrieval-augmented generation for AI-generated content: A survey. *arXiv preprint arXiv:2402.19473* (2024).

[183] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong. 2024. Uni-ControlNet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems* 36 (2024), 11127–11150.

[184] Yixiao Zheng, Kaiyue Pang, Ayan Das, Dongliang Chang, Yi-Zhe Song, and Zhanyu Ma. 2024. CreativeSeg: Semantic segmentation of creative sketches. *IEEE Transactions on Image Processing* 33 (2024), 2266–2278.

[185] Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 2023. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. *arXiv preprint arXiv:2312.02439* (2023).

[186] Chenliang Zhou, Fangcheng Zhong, and Cengiz Öztireli. 2023. CLIP-PAE: Projection-augmentation embedding to extract relevant features for a disentangled, interpretable and controllable text-guided face manipulation. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–9.

[187] Chenyang Zhu, Kai Li, Yue Ma, Chunming He, and Li Xiu. 2024. MultiBooth: Towards generating all your concepts in an image from text. *arXiv preprint arXiv:2404.14239* (2024).

[188] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 2223–2232.

[189] Yuchao Zhuo, Bingyuan Wang, and Zeyu Wang. 2024. Ink Harmony: An AI-and VR-enhanced system for calligraphy education. In *Proceedings of the ACM SIGGRAPH 2024 Immersive Pavilion*. 1–2.

[190] Ania Zubala, Nicola Kennell, and Simon Hackett. 2021. Art therapy in the digital world: An integrative review of current practice and future directions. *Frontiers in Psychology* 12 (2021), 600070.

Received 17 August 2024; revised 13 February 2025; accepted 1 April 2025