ENGINEERING Information Technology & Electronic Engineering
https://www.academax.com/EITEE; https://www.engineering.org.cn
https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=11344065
E-mail: eng.itee@zju.edu.cn

Engineering Journals
Available on IEEE Xplore®

Review
https://doi.org/10.1631/ENG.ITEE.2026.0124

Intelligent interaction system for art image creation: challenges, methodologies, and future prospects

Huanghuang DENG[§1], Aoran SUN[§2], Zhan MENG[§3], Yongwei WANG[4✉], Luyang CHEN[4], Dunhan LI[2], Zhiyu LI[4], Yue HAN[2], Zejian LI[3], Kun KUANG[2], Zemin LIU[2], Ning ZOU[5], Zhou ZHAO[2], Kejun ZHANG[2], Lingyun SUN[2], Yu ZHAO[6,4✉], Fei WU[2✉], Yunhe PAN[5]

[1]*State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou 310027, China*
[2]*College of Artificial Intelligence, Zhejiang University, Hangzhou 310058, China*
[3]*School of Software Technology, Zhejiang University, Hangzhou 310058, China*
[4]*College of Media and International Culture, Zhejiang University, Hangzhou 310058, China*
[5]*College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*
[6]*Institute of Intelligent Media and Film Arts, Zhejiang University, Hangzhou 310058, China*

Abstract: Graphic visual art is undergoing a profound transformation driven by generative artificial intelligence (AI). This transformation extends beyond technological innovation, reshaping the paradigms of artistic image creation. This paper examines human-centered interactive systems for artistic image creation, aiming to bridge the gap between technology-oriented and art-oriented research. We identify three core challenges in interactive generative art creation: controllability, integrity, and creativity, and analyze in depth the underlying trade-offs among them. Building on this analysis, we systematically review the evolution of generative models, approaches to integrating artistic knowledge, and the emergence of collaborative agentic AI systems. Looking ahead, we emphasize the need to develop human–AI collaborative frameworks that offer greater controllability, preserve human creative agency, and ensure regulatory compliance, while exploring a wide range of potential applications. This study proposes a unified analytical framework that connects AI technologies, human–AI interaction processes, and the integration and innovation of artistic image creation. It further outlines a task-oriented roadmap to guide future research and the development of collaborative systems for artistic image creation.

Key words: Cross-media intelligence; Human–AI collaboration; Visual art creation; Interactive systems; Controllability; Integrity; Creativity

1 Introduction

1.1 Research background

Graphic visual art constitutes a central medium for human culture and emotional expression, playing an essential role in fostering creativity, transmitting social narratives, and driving economic and social innovation through industries such as education, media, culture, and design. Throughout human history, visual art has recorded the spirit of different eras, the diversity of lived experience, and the evolution of human emotions and ideas, serving as a cultural bridge connecting the past, present, and the future (Avakri and Avakri, 2025; Li ZJ et al., 2025; Oppenlaender et al., 2025).

Recent advances in generative artificial intelligence

§ Equal contribution
✉ Yongwei WANG, yongwei.wang@zju.edu.cn
✉ Yu ZHAO, yuzhao@zju.edu.cn
✉ Fei WU, wufei@zju.edu.cn
Yongwei WANG, https://orcid.org/0000-0001-9712-8964
Fei WU, https://orcid.org/0000-0003-2139-8807
CLC number: TP391.41

Received: Apr. 28, 2026; Revision accepted: Aug. 31, 2026;
Crosschecked: Sept. 22, 2026

© The Authors 2026. Published by Zhejiang University Press Co., Ltd. This is an open access article distributed under the terms of the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/)

(AI) have transformed the ways in which visual content can be created and refined. Artists can now use multimodal information such as text, sketches, and reference images to explore visual ideas and perform artistic practice. For example, the 2022 Colorado State Fair award received by Jason Allen's *Théâtre D'opéra Spatial* (https://www.jasonmallen.com/space-opera-theater) drew attention to AI-assisted art making and to the role of human selection and post-editing in the creative process. In this setting, generative AI is not only a tool for producing images, but also supports conceptual ideation and iterative refinement.

Existing surveys generally organize generative AI by model architectures or general image-generation tasks (Shahriar, 2022; Cao et al., 2024; Sengar et al., 2025). Within visual art, work (Wang BY et al., 2025) reviews diffusion model-based methods and identifies a shift towards more interactive and user-oriented systems. Although these existing perspectives are valuable for understanding technological progress, they provide limited discussion of how generative AI participates in artistic workflows across different model families. This review therefore examines interactive artistic image creation through the broader issues of intention alignment, multimodal interaction, creative agency, and responsible dissemination.

1.2 Conceptual clarifications

1. Image vs. artistic image. An image is a visual representation that may present, record, and communicate information. This broad category includes photographs, charts, illustrations, and other visual forms. By contrast, the artistic images examined in this study refer specifically to artworks in painting and digital paintings. They not only share the pixel structure of general digital images (e.g., color pixel arrangement and resolution composition), but also embed complete narrative structures and systematic artistic visual knowledge (including intra-image and extra-image factors). Created through distinctive artistic techniques and aesthetic expression, artistic images integrate personal emotion, conceptual thinking, and narrative intent to evoke deep aesthetic experiences and emotional resonance in viewers.

For example, Vincent van Gogh's *The Starry Night* constructs a tri-layer narrative of natural turbulence, human tranquility, and vital tension through swirling skies, quiet villages, and vertical cypresses. Its artistic visual knowledge includes intra-image factors such as color contrast (deep blue vs. bright yellow), rhythmic brushwork (vortex strokes evoking celestial motion), and compositional balance. Extra-image factors include Van Gogh's mental state and broader post-impressionist artistic movements, which enrich the interpretive depth of the work. This sharply contrasts with an ordinary photograph of a night sky that lacks such narrative and artistic knowledge. Artistic images prioritize aesthetic value and conceptual depth, often employing symbolic, metaphorical, or stylized techniques that integrate pixel structures, narrative structures, and artistic visual knowledge to enable higher-level emotional and intellectual expression.

2. Generation vs. creation. AI generation is primarily driven by data and knowledge. By learning from large-scale datasets, AI models acquire patterns that enable them to generate new content. In image generation, for instance, the model analyzes numerous visual features and then produces corresponding images based on user input. However, the model lacks the emotional depth, intentionality, and subjectivity inherent to human creation; it merely outputs results following predefined algorithms and learned patterns.

Creation, by contrast, is a human practice rooted in subjective intention, emotional experience, and cultural context. It is characterized by the dynamic integration of pictorial techniques (e.g., color theory, linework, and composition) and non-pictorial factors (e.g., personal experiences, emotional fluctuations, creative intention, and prevailing aesthetic ideologies). The unity of these elements allows artists to transform abstract emotions and ideas into perceptible artistic forms. For example, Picasso's *Guernica* combines cubist deconstruction and oppressive tonalities with anti-war sentiment and historical trauma to achieve profound expressive power.

3. Prompting vs. intelligent interaction. A dialogue prompt is a one-way natural language command mechanism. It follows a linear logic in which the user issues instructions and the system executes them based on predefined rules. This mechanism lacks multimodal fusion and bidirectional feedback, limiting its capacity to interpret abstract intentions, regulate generative uncertainty, or support collaborative creativity. It is thus suitable only for low-complexity tasks and cannot meet the demands of artistic creation, where intentions are nuanced and multi-layered.

Intelligent interaction—the core operational pathway of the "human–AI collaborative agent" paradigm—aims to achieve intent alignment between creators and generative models. As a multimodal and deeply collaborative process, intelligent interaction fulfills three key functions: first, translating abstract intentions, namely, converting emotional, aesthetic, and cultural information into structured, model-interpretable conditions; second, regulating generative uncertainty, which allows creators to iteratively refine and steer outputs toward their envisioned direction; third, fostering co-creative emergence that assists in simulating new creative ideas, enabling outcomes unattainable by either party alone. In artistic scenarios, creators may express intentions through speech, gesture, sketching, or textual cues. The system continuously interprets multimodal signals and optimizes outputs to achieve progressive alignment with human intent.

4. AI-generated content. AI-assisted creation (AIAC) represents an emerging paradigm of human–AI collaboration in creative practices, where AI functions as an intelligent assistant rather than an autonomous creator. Within this paradigm, AI contributes computational power, knowledge support, and generative capabilities, while human creators retain responsibility for creative intent, aesthetic judgment, and decision-making. By reducing the effort required for technical execution and content generation, AIAC enables creators to focus more on conceptual development, artistic exploration, and innovation. However, the effectiveness of AIAC depends not only on the capabilities of generative models but also on the extent to which users can control, guide, and refine the creative process. Therefore, enhancing controllability and supporting meaningful human–AI interaction have become fundamental challenges in the design of AI-assisted creative systems.

1.3 Literature review and current status

Current research on generative AI in visual art follows two largely isolated trajectories, i.e., technology-oriented research and art- and philosophy-oriented research:

Technology-oriented research focuses on models such as StyleGAN3, DALL·E, Imagen, and Muse. While technically insightful, such studies often overlook human–AI interaction, authentic artistic workflows, and associated societal and ethical issues.

Art- and philosophy-oriented research addresses the nature of art, creativity, and human experience, questioning whether AI can possess emotional depth or interpret cultural context. However, these discussions often lack grounding in underlying AI technologies and rarely offer actionable design guidance.

This review aims to bridge these isolated paths by establishing an integrated analytical framework spanning technology, interaction, art, and ethics, thereby offering a comprehensive and practice-oriented perspective for this interdisciplinary field.

1.4 Research scope and methodology

This study focuses on interactive generative systems for artistic image creation, excluding general-purpose image generation and purely model-centric studies. The literature is drawn primarily from major academic databases and representative high-impact conferences, journals, and preprint repositories. Included works must address at least two of the following aspects: artistic image creation, human–AI interaction, and iterative or multimodal workflows. The paper is structured around a dual framework consisting of a four-stage task workflow and three core challenges.

Specifically, we examine interactive generative art creation in visual image contexts across three dimensions: core tasks, key challenges, and future directions. Core tasks include multimodal inspiration generation, compositional previsualization, aesthetic optimization, and mediated dissemination. Key challenges involve controllability, integrity, and creativity, including improving compositional precision, establishing objective and quantifiable aesthetic evaluation metrics, and addressing copyright, bias, and ethical compliance. The study also outlines future development directions and potential applications in domains such as education, media, and culture.

Methodologically, we employ:

(1) literature review, drawing from domestic and international research to identify trends and current knowledge gaps, and

(2) case analysis, examining representative interactive generative art systems to highlight technical applications, interaction modes, and practical challenges.

1.5 Research innovations

This study contributes to the interdisciplinary field of AI-assisted artistic creation through three core innovations:

First, we propose an integrated analytical framework that bridges technological and artistic research by systematically combining generative AI technologies, human–AI interaction processes, and artistic image fusion innovations. Second, we construct a structured roadmap for research and development that decomposes human–AI collaborative creation into sequential tasks—multimodal inspiration, collaborative generation, and mediated dissemination—and summarizes advances in intent modeling, cultural and aesthetic knowledge, and creative fusion.

Third, we identify core tensions and pathways forward, where we analyze the triangular trade-offs among controllability, integrity, and creativity, and identify knowledge fusion, model collaboration, and agentic AI as foundational technologies for future breakthroughs in dynamic and co-creative visual art systems.

As shown in the structural framework of the paper in Fig. 1, the remainder of this article is organized as follows: Section 2 provides an in-depth discussion of interactive artistic image generation tasks, including multimodal inspiration generation, compositional pre-visualization, aesthetic optimization, and media-oriented dissemination. Section 3 examines the key challenges related to controllability, integrity, and creativity. Section 4 reviews the methodological landscape, covering visual generative models, representations of artistic knowledge, intelligent creation systems, and agentic AI. Section 5 looks ahead to the future development of intelligent interactive art systems and their potential applications in education, media, culture, design, and other domains. Finally, Section 6 concludes the paper.

2 Interactive art image generation task

Image-based visual art creation, as a form of artistic expression, embodies both creativity and aesthetic value. Unlike function-oriented image generation in industrial applications, artistic creation places greater emphasis on symbolic meaning, stylistic aesthetics, and the communication of subjective intent. As illustrated in Fig. 2, by integrating perspectives from visual studies, art theory, and media studies with the paradigm of generative AI, we divide the workflow of interactive artistic image creation into four interrelated stages: multimodal ideation, progressive generation, aesthetic refinement, and mediated dissemination. This framework not only reflects the cognitive progression inherent in artistic creation but also offers a task-centered analytical lens for embedding interactive creative AI into artistic practice (Han and Fussell, 2025).

1. Multimodal ideation. The creative process begins with intention formation and associative exploration. In this interactive stage, artists integrate narrative themes, symbolic schemas, and stylistic priors—sourced from diverse multimodal inputs—to construct an initial semantic concept. Such multimodal inputs may include text, reference images, and sample materials. Through a collaborative feedback loop with an AI system, the artist iteratively refines the initial idea and transforms subjective inspiration into structured prompts. These prompts serve as machine-interpretable constraints or conditioning signals that guide generative models.

After an initial creative intention has been formed, the ideation process enters a critical stage of semantic concretization. At this stage, artists translate abstract themes, symbolic concepts, and fragmented stylistic impressions into concrete visual elements and structural relationships that can be

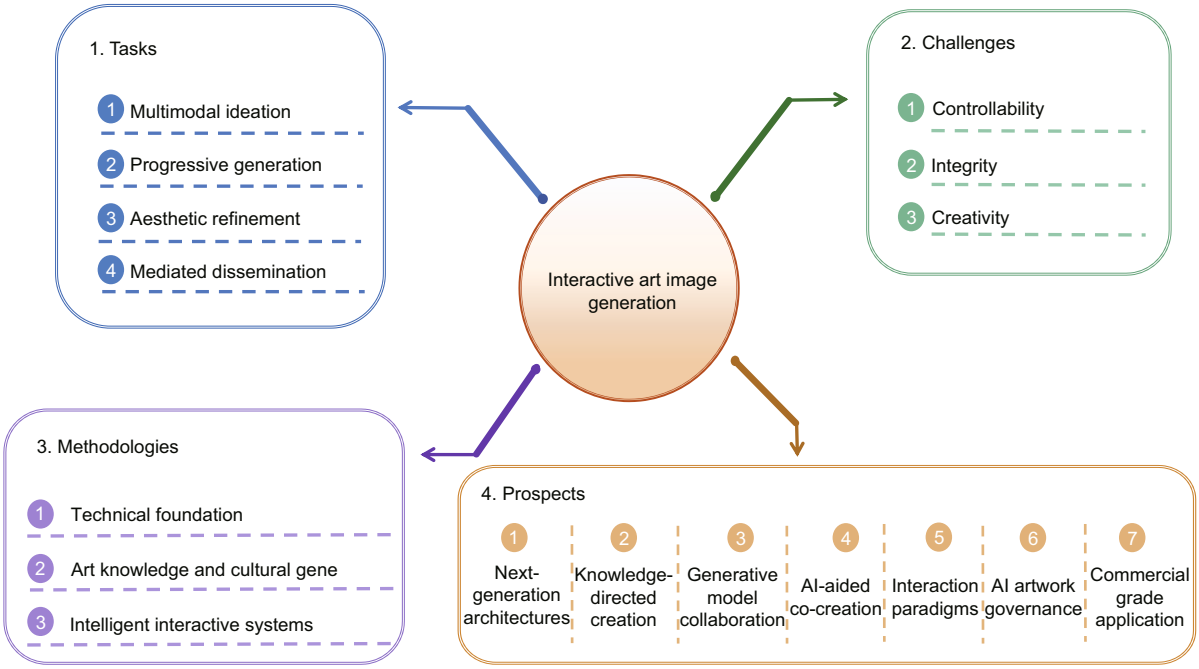


Fig. 1 Outline of this paper, which consists of four main components: art image generation tasks, key challenges, prevalent methodologies, and prospects

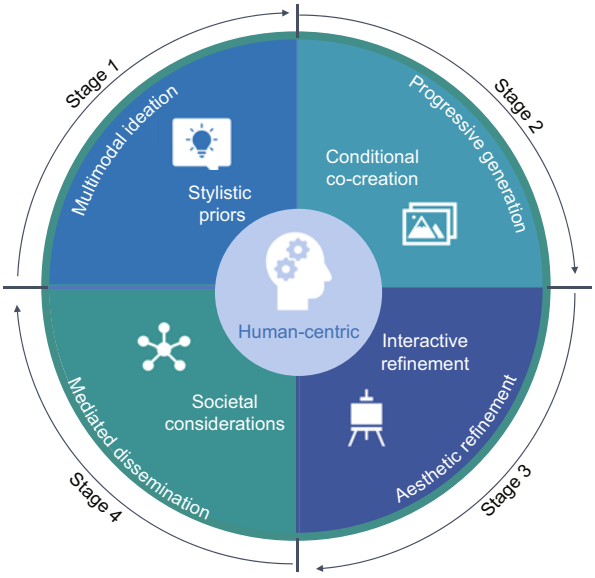


Fig. 2 Illustration of interactive art image generation task that involves four sequential stages: multimodal ideation, progressive generation, aesthetic refinement, and mediated dissemination

interpreted and manipulated by generative models. This process typically involves analysis, decomposition, and recombination of multimodal inputs. For instance, key concepts and affective tones may be distilled from textual descriptions, compositional patterns, color palettes, or textures may be extracted from reference images, and motion cues or material qualities may be inferred from example materials. Importantly, this process is not a simple aggregation or collage of heterogeneous inputs. Instead, guided by the artist's aesthetic judgment, a coherent internal logic is gradually established among multiple candidate elements. Through this process, dispersed sources of inspiration are organized into a preliminary visual blueprint with internal consistency. Subsequently, this emerging visual blueprint is encoded and articulated in the form of textual prompts. During this stage, artists often iteratively refine and validate these prompt descriptions by observing whether the generated outputs faithfully translate the intended creative concepts. Such an exploratory feedback loop facilitates semantic alignment between the artist's internal vision and the system's visual outputs, thereby establishing a clear and iterative foundation for the subsequent stage of fine-grained control and generation (Cai et al., 2023; Tholander and Jonsson, 2023; Adamkiewicz et al., 2025; Rafner et al., 2025; Satkunarajan et al., 2025; Wang P et al., 2026).

A core challenge in this stage lies in the potential semantic mismatch between multimodal prompts and visual outputs. Continuous interaction is therefore required to achieve cross-modal semantic alignment. The multimodal ideation phase sets the foundational semantic boundaries for later decisions on controllable generation and stylistic direction.

2. Progressive generation. A key turning point in the workflow lies in translating creative intent into concrete visual structures. This interactive process aims to convert abstract ideas and inspirations into verifiable spatial compositions, such as layout, form and posture, perspective and space, lighting, color relationships, brushwork and texture, and edge relations (Matsumoto et al., 2025). Based on artistic theory and principles from Gestalt psychology, artists interact with the AI system through structured inputs such as pose skeletons, segmentation masks, depth maps, and edge sketches.

To achieve the aforementioned level of controllability, artists can employ several types of structured conditioning signals, including the following:

(1) Pose and skeleton maps. Human keypoints or skeletal representations enable precise control over the posture, movement, and interaction relationships of characters or creatures (Shi HY et al., 2022; Ju et al., 2023). Such signals provide a structural foundation for narrative composition and dynamic expression.

(2) Edge maps and line drawings. Extracted edges or manually drawn sketches can strictly constrain object contours, shapes, and major structural lines within the image, ensuring that the generated content closely aligns with the intended composition (Zhang TY et al., 2024; Shi WY et al., 2025).

(3) Depth maps. Depth information provides explicit guidance on the spatial organization of the scene, enabling clear control over foreground and background layers. This facilitates the representation of perspective relationships and enhances the perceived three-dimensionality of the generated imagery.

(4) Semantic segmentation maps. Semantic labels are assigned to different regions of the image (e.g., sky, human, building, and tree), allowing precise control over the spatial layout, positioning, and relative scale of different object categories within the scene (Tan ZT et al., 2021).

(5) Reference-based style or structural transfer. The stylistic characteristics or compositional structures of a reference image can be extracted and transferred to newly generated content, enabling the imitation or inheritance of stylistic elements, color palettes, or layout patterns.

These control signals can be applied individually or flexibly combined to form multi-level and multidimensional constraints. Through such mechanisms, artists are able to explore the generative space within a more predictable and guided framework, where serendipitous outcomes are increasingly situated within rational planning. This capability allows artists to more effectively address challenges such as multi-object layout, detail consistency, and overall semantic coherence in complex scenes, gradually converging toward their original creative intentions.

Conditional generative models (e.g., ControlNet (Zhang L et al., 2023) and T2I-Adapter (Mou et al., 2024)) provide feasible pathways to incorporate such structured signals. To provide a more intuitive visual illustration, we present several generation examples in Fig. 3 to demonstrate some of the image-conditioned generation capabilities supported by ControlNet. However, a central challenge remains: how to handle multi-object layout, consistency, and semantic coherence in a smooth and user-friendly way.

This stage directly relates to the core controllability challenge of the system: how can multimodal control signals enable generative models to produce more reliable and interpretable outputs? The issue is particularly critical in complex and narrative-rich artistic scenarios.

3. Aesthetic refinement. The final value of an artwork is often expressed through the subtlety of its stylistic qualities and the coherence of its formal presentation. Once the composition is established, this stage focuses on refining color schemes, stylistic consistency, and the realism of local structures. Artists leverage AI systems to perform color transfer, style transfer, image inpainting or outpainting, detail enhancement, and edge relation optimization. These refinements are realized through a series of interactive editing operations, allowing artists to adjust and perfect the artwork iteratively (Huang YK et al., 2025) with the following approaches:

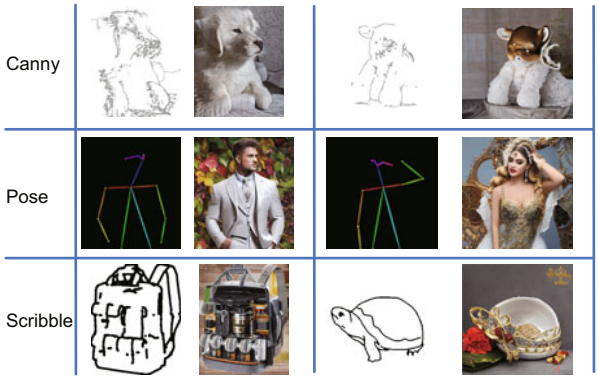


Fig. 3 Illustration of two sets of generated results for each of the different image conditions (Canny edges, human pose, and hand-drawn sketches from top to bottom) from ControlNet. On the left side of each example is the input control image, and on the right is the image generated based on the corresponding condition. It can be seen that ControlNet supports control with various structural inputs while still producing high-quality images

(1) Local redrawing and inpainting. Specific regions of an already generated image (e.g., facial details, object textures, or background elements) can be selectively masked and regenerated. This allows artists to correct artifacts, enrich local details, or replace undesired content while maintaining visual coherence with the surrounding context (Ju et al., 2024; Wang RC et al., 2024; Xu ZY et al., 2025).

(2) Global outpainting. Based on the existing content and stylistic characteristics of an image, the system can extend the canvas outward and generate additional visual content. This enables seamless scene completion or aspect-ratio adjustments, providing greater flexibility for compositional refinement (Zhang SF et al., 2024).

(3) Global stylization. Deep learning-based style transfer filters can be applied to impose specific artistic styles (e.g., oil painting, watercolor, or cartoon aesthetics) or emulate the characteristic brushwork of particular artists. Such transformations can be applied globally or locally to reinforce stylistic consistency across the image (Wang HF et al., 2024).

(4) Resolution enhancement and detail amplification. Super-resolution models and detail synthesis methods can increase the physical resolution of an image while intelligently introducing or strengthening plausible fine-grained textures, such as skin microstructures, fabric weaves, or leaf veins. This ensures that the artwork maintains clarity and realism even when enlarged.

Through iterative application of these techniques, artists can progressively transform an initially "approximately correct" generated draft into a refined final piece. This process allows the artwork to gradually meet the creator's expectations in terms of stylistic coherence, color expressiveness, textural realism, and overall completion quality, ultimately elevating it

to a state suitable for presentation and dissemination.

As artistic aesthetics are inherently subjective and diverse, there is currently no unified standard for quantifying the "aesthetic quality" of generative digital visual art. This gap has prompted the convergence of human preference modeling and automated aesthetic prediction metrics—together offering more robust and reproducible evaluation methods to guide interactive refinement. The challenges in this stage are directly tied to aesthetics, as the objective is not merely to achieve visual realism but to realize the expressive integrity of an artistic form.

4. Mediated dissemination. The lifecycle of an artwork does not end at its initial creation but extends into the complex dynamics of mediated dissemination. Within the "text–image–platform" ecosystem, artworks undergo reinterpretation, reframing, and recombination, with their meaning progressively established through textual anchoring (titles and captions) and narrative transmission (sequencing and commentary).

For works produced via human–AI collaboration, this stage raises profound socio-technical questions, such as ambiguous authorship in copyright determination and the need for provenance tracking to document the history of generation and editing. Addressing these issues requires AI systems to incorporate watermarking, blockchain technologies, and content authenticity verification into the creative workflow, thereby extending the interactive feedback loop into ethical and infrastructural dimensions.

Thus, this final stage closely links the artistic lifecycle with regulatory and ethical considerations, establishing the foundation for the responsible application of generative AI in cultural production.

In summary, in this section, we have introduced four stages of the interactive artistic image creation workflow. As shown in Table 1, the pipeline systematically progresses from multimodal ideation, through progressive generation and aesthetic refinement, to mediated dissemination, forming a coherent end-to-end creation paradigm. Each stage is characterized by distinct user input modalities, task objectives, and technical focuses, ranging from concept formation and structure construction to aesthetic enhancement and final content dissemination.

3 Challenges

As AI continues to penetrate the domain of artistic creation, understanding model capabilities alone is no longer sufficient. To move beyond the limitations of basic functionality, it is essential to simultaneously account for cutting-edge research progress and the practical needs of creative industries. Against this backdrop (as shown in Fig. 4), we identify three key challenges at the intersection of AI and the arts: controllability (the precise transmission and dynamic adaptation of multidimensional, cross-modal human intent), integrity (the scientific understanding, systematic mining, and compliant application of visual art heritage across cultures), and creativity (the integrative and original generation of image-based artworks that align with human aesthetic consensus and emotional resonance). We further examine the inherent tensions among these three challenges and explore how they shape the future trajectory of AI-driven artistic creation. Table 2 summarizes our analysis and key findings regarding controllability, integrity, and creativity.

3.1 Controllability: accurate conveyance and dynamic adaptation of human multidimensional cross-modal intentions

In generative visual systems, controllability refers to the ability of a model to generate or edit visual content according to user-provided conditions. Unlike general image quality or text–image alignment, controllability concerns whether users can explicitly influence specific aspects of the generation process and obtain predictable modifications. This capability becomes increasingly important for interactive creation, where users iteratively refine composition, semantics, structure, and local details rather than relying solely on text prompts (Hertz et al., 2022; Pan XG et al., 2023; Zhang L et al., 2023).

Existing research identifies three fundamental aspects of controllability. First, condition fidelity refers to whether

Table 1 An overview of four stages of the interactive artistic image creation workflow

Stage	Multimodal ideation	Progressive generation	Aesthetic refinement	Mediated dissemination
Sub-task	Inspiration elicitation; intent externalization	Layout control; pose control; structure-constrained rendering	Inpainting; outpainting; style unification	Provenance; attribution; release compliance
User input modality	Text; reference image; sketch	Pose; edge; depth; mask		
Goal	Concept formation	Structure creation	Aesthetic quality	Artwork dissemination; cultural production
Key challenge	Semantic alignment	Reliability; consistency	Subjective evaluation	Authorship
Representative method	Prompt engineering	ControlNet	Interactive editing operations	Watermarking
Evaluation cue	Semantic alignment; user satisfaction	Control fidelity; consistency	Edit locality; style coherence; preference	Watermark detectability; traceability; policy compliance

Table 2 Compact comparison of controllability, integrity, and creativity

Challenge	Metric	Case	Strength	Limitation & trade-off
Controllability	CLIPScore (alignment); IoU/mIoU (structure); locality; multi-turn consistency	ControlNet; Prompt-to-Prompt; DragGAN; GANPaint	Structured control; layout/pose editing; predictable steering	Reduced exploration; creativity trade-off; over-constrained edits; prior-dependent
Integrity	Provenance (watermark, metadata); fairness/bias; cultural audit; style similarity	C2PA; bias audits; Glaze; cultural analysis	Traceability; compliance; fairness awareness; artist protection	Non-unified metrics; usability trade-off; reduced flexibility; limits reuse
Creativity	FID/IS/KID; human preference; editing benchmarks; user studies	SketchRNN; GANCollage; LatentCreative	Exploration; variation; ideation support; co-creation	Hard to quantify; output bias; control trade-off; integrity constraints

IoU: intersection over union; mIoU: mean intersection over union; IS: inception score; KID: kernel inception distance

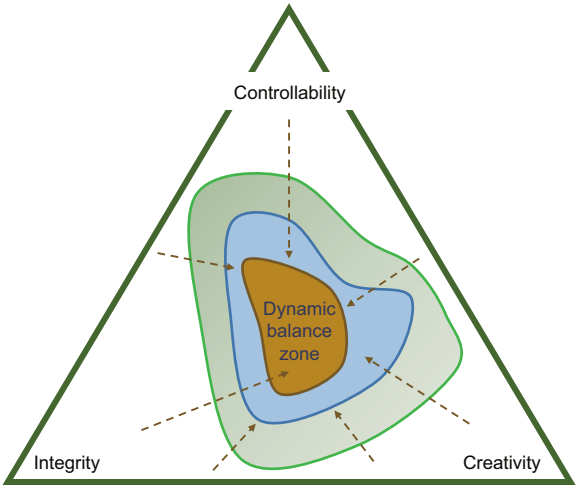


Fig. 4 Illustration of three core challenges, i.e., controllability, aesthetic quality (creativity aspect), and integrity, and their intricate triangular check-and-balance relationships in an intelligent interactive art creation workflow

generated outputs accurately follow external guidance, including textual descriptions, sketches, poses, depth information, and semantic layouts (Zhang L et al., 2023; Zhao SH et al., 2023). Second, targeted modification and preservation describe the ability to introduce desired changes to specific attributes, regions, or concepts while preserving unrelated visual content, such as identity, structure, and style (Bau et al., 2019; Hertz et al., 2022; Pan XG et al., 2023). Third, iterative consistency concerns whether user intentions and previously established visual elements can be maintained during successive generation and editing steps. This property is particularly important in multi-turn generation and editing scenarios, where models must adapt to new instructions while avoiding semantic drift and unintended changes across interactions (Cheng et al., 2024; Wang ZY et al., 2024).

Accordingly, controllability in interactive visual art creation can be understood as the capability of generative systems to translate human intentions into controllable generation and editing operations while maintaining semantic and visual constraints (Hertz et al., 2022; Pan XG et al., 2023; Zhang L et al., 2023; Cheng et al., 2024).

3.1.1 Precise conditional control

A fundamental challenge of controllable visual generation is enabling models to accurately follow diverse user-provided conditions (Zhang L et al., 2023; Zhao SH et al., 2023). Early text-to-image systems primarily relied on textual prompts as high-level semantic guidance; however, artistic creation often requires more explicit control over spatial structures, object arrangements, poses, and visual attributes. Consequently, recent research has explored additional conditioning signals, including sketches, edge maps, depth information, human poses, and semantic segmentation layouts, to provide more precise guidance beyond natural language descriptions (Li YH et al., 2023; Mou et al., 2024).

One major challenge in precise conditional control is the alignment between heterogeneous conditions and the generated visual content. Different modalities represent user intentions at different abstraction levels: text typically conveys semantic concepts, while sketches, layouts, and structural maps provide spatial constraints (Huang LH et al., 2023; Zhang L et al., 2023). When multiple conditions are jointly applied, conflicts may occur because different modalities can provide inconsistent or competing guidance. Existing approaches address this issue through mechanisms such as modality-specific attention alignment, adaptive condition weighting, and multi-stage conditioning strategies, which aim to balance semantic information and structural constraints during generation (Hertz et al., 2022; Rombach et al., 2022; Hu et al., 2023).

Despite substantial progress, precise conditional control remains challenging because stronger adherence to external constraints does not always guarantee visually coherent or aesthetically desirable results (Rezwana and Maher, 2023; Zhang L et al., 2023; Zhao SH et al., 2023). For example, enforcing strict structural conditions may limit the flexibility of generation, while weak constraints may lead to semantic drift or insufficient control. Therefore, effective controllable generation requires not only following individual conditions, but also balancing different forms of guidance according to the requirements of the creative task.

3.1.2 Human–AI interaction and workflow adaptation

Beyond the ability to follow conditions, controllable visual systems must support effective interaction between human creators and generative models. In artistic workflows, users rarely

provide a complete specification at the beginning; instead, they progressively explore ideas, evaluate intermediate results, and refine outputs through repeated interactions. Therefore, controllability depends not only on generation accuracy, but also on whether users can understand, adjust, and revise model behaviors during the creative process (Frich et al., 2019; Rezwana and Maher, 2023).

Recent interactive generation systems have moved beyond prompt-only interfaces by incorporating multimodal interaction mechanisms, including sketches, masks, point-based manipulation, and structured editing operations (Avrahami et al., 2022; Pan XG et al., 2023). These interfaces allow users to directly express spatial or semantic intentions that are difficult to describe through language alone. Studies on human–AI co-creative systems suggest that effective interaction requires users to maintain a sense of agency, receive interpretable feedback, and iteratively refine outputs rather than repeatedly regenerate images through trial and error (Shneiderman, 2022).

The integration of generative models into existing creative workflows further introduces challenges in maintaining consistency between user actions and system responses. Practical tools increasingly adopt interactive editing paradigms, where generation and modification are combined into a continuous process rather than isolated operations. For example, instruction-based editing and direct manipulation approaches demonstrate that controllable systems can better support creative workflows when users can selectively modify generated content while preserving previously accepted results (Hertz et al., 2022; Pan XG et al., 2023). Thus, human–AI interaction in visual creation requires not only controllable generation techniques, but also interfaces and workflows that enable efficient communication of intent and incremental refinement.

3.1.3 Generation consistency: spatial, temporal, and multi-turn

Maintaining consistency during generation and editing is essential for interactive visual creation, where users often expect modifications to preserve previously established content. At the spatial level, consistency requires that different regions within an image remain coherent in terms of identity, structure, and visual style, while local edits should avoid unintended changes to unrelated areas. Image editing studies demonstrate that preserving non-target regions during modification remains a key challenge for controllable generation systems (Hertz et al., 2022; Mokady et al., 2023).

Beyond individual images, temporal consistency becomes critical in applications such as visual storytelling, animation, and video generation. Generated sequences must maintain stable identities, environments, and stylistic characteristics across frames. Early research on story-based image generation highlighted the importance of maintaining both local image coherence and global narrative consistency, while subsequent video generation studies further investigated mechanisms for preserving object appearance and motion continuity across time (Li YT et al., 2019; Blattmann et al., 2023; Wu JZ et al., 2023).

Multi-turn interaction introduces an additional consistency requirement because users progressively refine their intentions through successive instructions. In such scenarios, the system must incorporate new modifications while preserving previously confirmed elements and avoiding accumulated semantic drift. Recent studies on interactive image generation show that models may fail to maintain subject attributes or contextual information across multiple turns, motivating approaches that incorporate dialogue history and contextual representations for more consistent generation (Cheng et al., 2024; Wang ZY et al., 2024; Yang ZY et al., 2024).

Therefore, generation consistency extends controllability from single-step condition following to long-term interaction. A controllable visual system should not only respond to current user inputs, but also maintain the continuity of visual content and user intentions throughout iterative generation and editing processes.

3.2 Integrity: cultural fidelity, rights protection, and responsible application

Integrity is a prerequisite for the sustainable development of interactive AI art. It concerns two overarching goals: preserving the cultural and representational foundations of human visual art and ensuring that generative systems are used in a traceable, fair, and rights-compliant manner. Operationally, these goals involve four coupled dimensions. Cultural integrity concerns whether artistic symbols, styles, and conventions are represented with sufficient contextual fidelity; representational integrity concerns whether generated outputs avoid systematic demographic distortion and harmful stereotypes; provenance integrity concerns whether the origin and editing history of an artwork can be reliably traced; rights-related integrity concerns unauthorized reproduction, training-sample memorization, authorship, and artist-specific style mimicry. These dimensions are particularly important in interactive creation, where images are repeatedly generated, edited, recombined, and disseminated across different platforms (Coalition for Content Provenance and Authenticity, 2024; AlDahoul et al., 2025; U.S. Copyright Office, 2025a; Vice et al., 2026).

3.2.1 Cultural and representational integrity

Improving cultural integrity requires coordinated intervention at the data, representation, knowledge, and governance levels. At the data level, Western-art-centered collections should be supplemented with geographically, historically, and linguistically diverse images, art-historical documents, creative contexts, and community interpretations, since current text-to-image models exhibit substantial disparities in representing culturally specific architecture, clothing, food, and implicit cultural expectations (Bayramli et al., 2025; Nayak et al., 2025). At the representation and knowledge levels, multimodal models, art-historical ontologies, and knowledge graphs can connect visible properties—including materials, composition, iconography, and brushwork—with historical conditions, aesthetic principles, and symbolic meanings. Because such structures may reproduce exclusions and classificatory assumptions of their sources, they should preserve attribution, uncertainty, alternative interpretations, and community-specific terminology rather than presenting culture as a fixed canonical taxonomy.

These technical measures should be complemented by participatory practices in which artists, curators, historians, anthropologists, tradition bearers, and represented communities

contribute to data selection, annotation, concept definition, and output review. This approach is consistent with UNESCO frameworks requiring the participation of communities that create and transmit living heritage and recommending multicultural and inclusive assessment throughout the AI life cycle (UNESCO, 2003, 2021). Cultural integrity therefore cannot be established solely through visual similarity or benchmark optimization, particularly because automatic evaluation metrics remain weak proxies for human judgments of explicit and implicit cultural expectations (Nayak et al., 2025).

Representational integrity extends this concern to demographic and social portrayals. Text-to-image systems can reproduce associations involving gender, race, age, authority, and occupation, as demonstrated by open-set and occupation-oriented audits of contemporary generators (D'Incà et al., 2024; Klassert et al., 2026). However, the magnitude and direction of measured bias depend on the model, prompt language, sampling procedure, and demographic classifier. Claims of amplification require particular caution, because apparent gender–occupation amplification can decrease substantially after distributional differences between training captions and evaluation prompts are controlled (Seshadri et al., 2024). Quantitative measures of representation, diversity, stereotypes, and group-conditioned reliability should therefore be combined with curator review, community audit, and contextual-appropriateness assessment.

3.2.2 Provenance, authorship, and rights-related integrity

Rights-related integrity concerns the provenance of training data, the reproduction of protected expression, and the conditions under which artist-specific materials are used. Because web-scale corpora often obscure individual sources, licenses, and rights reservations, the legality of training depends on jurisdiction, applicable exceptions, licensing arrangements, and case-specific facts rather than on data collection alone (U.S. Copyright Office, 2025b; Royal Courts of Justice, 2025; United States District Court for the Northern District of California, 2026). Nevertheless, experiments extracting training examples from diffusion models demonstrate that near-duplicate reproduction and prompt-triggered memorization are measurable risks rather than merely hypothetical concerns (Carlini et al., 2023; Chen C et al., 2024; Ma et al., 2025). Rights-related evaluation should consequently distinguish general stylistic or semantic learning from dataset duplication, memorized reproduction, and substantial similarity to particular sources.

Artist-specific style imitation is related to, but distinct from, the reproduction of protected works. Although stylistic similarity alone does not automatically establish copyright infringement, artist-specific fine-tuning can affect consent, attribution, professional identity, and market substitution, motivating defensive mechanisms such as style cloaking (Shan et al., 2023). Accordingly, style mimicry should be assessed through artist-centered and market-contextual evidence without treating every instance of stylistic influence, reference, or homage as legally equivalent.

Authorship likewise requires a contribution-specific analysis. U.S. law requires human authorship and excludes works identified as autonomously generated by a machine, as affirmed in Thaler v. Perlmutter (United States Court of Appeals for the District of Columbia Circuit, 2025). AI-assisted works may nevertheless receive protection when humans determine sufficiently expressive elements through selection, arrangement, modification, compositing, or other creative contributions, whereas prompts alone will not ordinarily establish authorship of the resulting expression (U.S. Copyright Office, 2025a). Workflow records can make these human and machine contributions more identifiable, but they provide evidence for, rather than automatically determine, authorship and ownership.

Provenance technologies make parts of this workflow inspectable. Coalition for Content Provenance and Authenticity (C2PA) content credentials can cryptographically bind assertions about origin, ingredients, and modification history to an asset, while fingerprints and watermarks can support recovery after metadata loss (Coalition for Content Provenance and Authenticity, 2026b). Their evidentiary scope remains limited: they verify the association and integrity of recorded assertions, not their completeness, factual truth, legality, fairness, or cultural appropriateness (Coalition for Content Provenance and Authenticity, 2026a). Provenance robustness should therefore be assessed through manifest validity, edit-chain completeness, signer trust, metadata survival, and recovery capability, in combination with licensing records, memorization analysis, and artist-centered review.

3.2.3 Governance and multidimensional evaluation

Regulatory frameworks increasingly translate selected integrity concerns into life-cycle obligations. The European Union (EU) requires general-purpose AI providers to maintain a copyright-compliance policy and publish a summary of training content, while China requires explicit and implicit labels in specified generation and dissemination contexts (European Union, 2024; Cyberspace Administration of China, 2025; European Commission, 2025). U.S. policy and case law similarly distinguish copyrightability, human authorship, and the legality of generative-AI training (U.S. Copyright Office, 2025a, 2025b; United States Court of Appeals for the District of Columbia Circuit, 2025). These frameworks collectively support greater transparency and traceability, but their different legal bases, scopes, and enforcement mechanisms should not be presented as interchangeable.

Integrity should consequently be evaluated as a structured profile rather than a single scalar score. Provenance metrics assess origin and edit-history claims; fairness audits assess demographic distributions, stereotypes, and group-conditioned reliability; cultural review assesses historically situated meaning; rights-related analysis assesses licensing, reproduction, memorization, and mimicry. These dimensions are not directly commensurable: a system may preserve valid provenance while producing stereotypes, or improve demographic diversity while introducing historical inaccuracies (Wan et al., 2024a). Evaluation should therefore report at least provenance robustness, demographic fairness and reliability, cultural-context appropriateness, and rights-related reproduction and mimicry risks, together with their uncertainty and responsible reviewers.

This profile should include evidence concerning the final image as well as the data, inputs, human edits, model changes, review, and dissemination. Integrity safeguards may

strengthen accountability but can also impose documentation burdens, visual perturbations, latency, or creative restrictions. Their protective effects and costs to controllability, accessibility, and exploratory creation should therefore be evaluated explicitly. Where living heritage is involved, the acceptable balance should be determined with the participation of represented communities rather than exclusively through developer-defined or aggregate benchmark criteria (UNESCO, 2003, 2021).

Perceived creativity depends partly on whether audiences attribute meaningful intention and creative agency to the process that produced the work. Experimental studies show that identical or comparable works tend to receive lower evaluations when attributed to AI, particularly in perceived creativity, sensitivity, awe, and aesthetic value (Chamberlain et al., 2018; Agudo et al., 2022; Millet et al., 2023). These effects are not determined by visual form alone: they are moderated by viewers' beliefs about whether creativity is a uniquely human capacity and by the degree of intentional agency they attribute to the creator (Jucker et al., 2014; Millet et al., 2023).

Another important artistic challenge concerns emotional resonance. Artistic interpretation is often shaped not only by visible form but also by contextual information concerning the artist, production process, intended meaning, and historical setting (Leder et al., 2004; Jucker et al., 2014; Darda and Chatterjee, 2023). Contemporary image generators, by contrast, produce outputs by sampling from learned statistical representations conditioned on prompts or other inputs rather than by drawing on autobiographical experience in the human sense (Ho et al., 2020; Rombach et al., 2022). This difference contributes to the perception that AI-generated art lacks emotional depth or expressive intention, especially when viewers are explicitly informed of AI authorship (Agudo et al., 2022; Millet et al., 2023).

Emotional responses to AI art should nevertheless be evaluated empirically rather than assumed to be impossible. In a 2×2 experiment, Demmer et al. (2023) found that viewers reported emotions and attributed intentionality to both computer- and human-created artworks, although the actually human-created works elicited stronger reported emotions under the tested conditions. Taken together, these findings support a relational account in which emotional resonance emerges from the interaction among the perceptual properties of an artwork, viewers' expectations and prior beliefs, and contextual information about authorship and intention (Leder et al., 2004; Demmer et al., 2023; Millet et al., 2023).

3.3 Creativity: fusion-based innovation, human perception, and evaluation

Creativity is a central objective of interactive AI art systems, but it should not be equated with visual fidelity, prompt compliance, or stylistic novelty alone. A widely adopted definition characterizes creativity as the production of outcomes that are both original and effective, where effectiveness may involve usefulness, appropriateness, or value within a particular context (Runco and Jaeger, 2012). This definition is especially relevant to AI-assisted art: an unexpected image is not necessarily creative if its novelty results from incoherence or error, while a technically polished imitation may remain insufficiently original. Creativity in interactive generation should therefore be understood as a situated property involving the generated artifact, human creative process, and social interpretation of the result (Rhodes, 1961; Amabile, 1982).

3.3.1 Technological perspective: quantifying artistic value and enabling cultural fusion

A central technological challenge is that commonly used automatic metrics capture only partial aspects of artistic creativity. Fréchet inception distance (FID) estimates distributional similarity between generated and reference images (Heusel et al., 2017); CLIPScore measures compatibility between an image and a textual description (Hessel et al., 2021); preference datasets such as the Pick-a-Pic model comparative user choices (Kirstain et al., 2023). These measures can provide evidence about realism, distributional coverage, text–image alignment, or aggregate preference. However, none directly determines whether an artwork is simultaneously original, coherent, meaningful, and contextually valuable. Comparative studies further show that rankings produced by generative-model evaluation metrics can diverge substantially from human judgments (Stein et al., 2023). Automatic metrics should therefore be treated as complementary diagnostic instruments rather than stand-alone measures of artistic creativity.

Human evaluation remains necessary because judgments of creativity depend on disciplinary conventions, audience knowledge, and the purpose of the creative task. The consensual assessment technique evaluates creativity through independent judgments by domain-relevant experts instead of relying on a universal formal definition (Amabile, 1982). The creative product semantic scale (CPSS) similarly separates product creativity into novelty, resolution, and elaboration–synthesis, showing that originality and technical or conceptual appropriateness are distinguishable dimensions (O'Quin and Besemer, 1989). These methods provide stronger conceptual grounding than evaluations based only on attractiveness or realism, although their dependence on human judges makes repeated and large-scale assessment comparatively demanding.

Beyond quantification, a second technological challenge concerns the support of meaningful conceptual and cultural fusion. Boden (2004) described combinational creativity as the production of new ideas through unfamiliar combinations of existing concepts. In culturally situated generation, however, meaningful combination requires more than placing recognizable motifs from different traditions in the same image. Cultural coherence may depend on relationships among symbols, materials, colors, compositional conventions, ritual functions, and historical contexts.

Recent benchmarks provide empirical evidence for this limitation. CULTDIFF finds that text-to-image models frequently fail to reproduce culturally specific artifacts in categories such as architecture, clothing, and food, with more pronounced limitations for underrepresented regions (Bayramli et al., 2025). CulturalFrames similarly reports substantial failures in satisfying explicit and implicit cultural expectations and finds that existing automatic metrics correlate poorly with human judgments of cultural alignment (Nayak et al., 2025). Consequently, an image that combines Dunhuang imagery with

Cyberpunk elements should not be considered creative merely because both visual vocabularies are recognizable. Its creativity also depends on whether the relationship between those vocabularies is visually intelligible, conceptually motivated, and culturally defensible.

Structured cultural knowledge offers a possible, although not yet conclusive, response to this problem. Art-specific knowledge graphs have been used to encode relationships among artists, artistic schools, historical periods, and other attributes, improving classification and cross-modal retrieval in automatic art analysis (Garcia et al., 2019). Retrieval-based grounding can likewise improve the representation of underrepresented cultural concepts. Culture-TRIP, for example, retrieves contextual and visual information about culturally specific concepts and iteratively refines prompts; its user study reports improved alignment between generated images and underrepresented culture nouns (Jeong et al., 2025).

These findings support the narrower claim that explicit relational knowledge can improve contextual grounding. They do not yet demonstrate that knowledge graphs, retrieval systems, or prompt-refinement methods can autonomously produce culturally meaningful artistic innovation. Whether such representations can support novel relationships between cultural elements without simplifying, decontextualizing, or stereotyping their source traditions therefore remains an open research problem.

3.3.2 Artistic perspective: audience perception and emotional resonance

From an artistic perspective, creativity is not determined exclusively by the formal properties of the generated image. Judgments of artistic value are also influenced by what audiences believe about the creator, the creative process, and the intentions behind the work. Experiments in art perception show that contextual information about an artist's intentions can change how viewers interpret and evaluate an artwork (Jucker et al., 2014). The perceived presence of intention therefore contributes to artistic reception even when the visible properties of the artifact remain unchanged.

Studies comparing human- and AI-attributed art provide further evidence for this effect. AI authorship labels can reduce evaluations of creativity, aesthetic value, or emotional significance even when the visual stimulus is held constant or closely controlled (Chamberlain et al., 2018; Millet et al., 2023). These effects are not uniform, however. They vary according to viewers' prior beliefs about machine creativity, their attitudes toward human uniqueness, and the formal qualities of the artwork. It is therefore more accurate to conclude that the reception of AI-generated art is partly mediated by perceived human intention, effort, and agency than to claim that audiences categorically reject AI authorship.

The emotional value of AI-generated art requires similar qualification. Contemporary generative models learn statistical regularities from training data and do not possess autobiographical experience or artistic intention in the ordinary human sense. Nevertheless, the absence of human-like subjective experience does not imply that their outputs cannot produce emotional responses. Controlled studies report emotional and intentionality-related responses to both human- and computer-generated artworks, while also identifying differences in the strength of those responses and in the intentions attributed to their creators (Demmer et al., 2023).

Emotional resonance should therefore be understood relationally rather than as an intrinsic property guaranteed by either human or machine authorship. It emerges through the interaction among the formal properties of the artifact, the viewer's interpretation, contextual information about its production, and the role attributed to human collaborators. For interactive AI art systems, this means that preserving visible human choices, revisions, and intentions may affect not only the user's sense of authorship but also the way in which audiences interpret the resulting work.

3.3.3 Evaluation perspective: creative products and human creative processes

A defensible evaluation framework should distinguish the creativity of the resulting product from the system's support for the human creative process. At the product level, evaluation can combine the general criteria of originality and effectiveness (Runco and Jaeger, 2012) with the CPSS dimensions of novelty, resolution, and elaboration–synthesis (O'Quin and Besemer, 1989). Novelty concerns the extent to which the work departs from familiar solutions; resolution concerns whether the work functions coherently in relation to its purpose; elaboration–synthesis concerns how successfully its elements are integrated into a complete and intelligible artifact.

For fusion-based AI art, these product dimensions should be supplemented by relational coherence. This criterion asks whether combined elements establish a perceptible conceptual, stylistic, or cultural relationship rather than merely co-occurring in the same composition. Under this framework, unusual visual noise would not qualify as creative simply because it is statistically novel, and a polished reproduction would not qualify merely because it is coherent. A persuasive creative product must balance difference with intelligibility and contextual appropriateness. Automatic measures of diversity, alignment, and realism may supplement this assessment, but they cannot replace context-sensitive human judgment (Stein et al., 2023; Nayak et al., 2025).

Product evaluation alone, however, cannot determine whether an interactive system supports human creativity. Rhodes's distinction among person, process, product, and press indicates that creativity involves not only an artifact but also the individual, the acts through which it is produced, and the surrounding social environment (Rhodes, 1961). Correspondingly, the evaluation of computational co-creative systems should specify who or what is being evaluated, when the evaluation takes place, which aspects of the interaction are considered, and whose judgment provides the evidence (Maher et al., 2018).

Relevant process-level evidence includes the range of alternatives explored, the diversity of generated directions, the extent and direction of revision, and the user's ability to reject, appropriate, or substantially transform system suggestions. It should also examine whether the users retain decision-making agency and can identify the resulting work as an expression of their own creative intentions. These dimensions distinguish a system that merely produces an impressive image from one

that supports exploration, reflection, revision, and sustained creative inquiry.

This distinction is particularly important because generative assistance can facilitate ideation while also constraining it. In a controlled visual-ideation study, exposure to generative-AI examples increased design fixation and was associated with fewer, less varied, and less original ideas (Wadinambiarachchi et al., 2024). This result should be interpreted within the specific experimental task and should not be generalized to every form of AI-assisted creation. Nevertheless, it demonstrates that polished outputs and rapid generation are insufficient evidence of creativity support. An interactive system should also be evaluated according to whether it expands or narrows the user's search space, encourages meaningful revision, and enables departure from the model's initial proposals.

Accordingly, creativity in interactive AI art generation is best evaluated across three connected levels. The technological level concerns whether the system can represent and manipulate meaningful conceptual and cultural relationships rather than optimize only fidelity or alignment. The artistic level concerns how human intention, authorship, and production context shape audience interpretation and emotional resonance. The evaluation level concerns whether the resulting artifact is original, effective, and relationally coherent and whether the interaction supports exploration, revision, agency, and human creative ownership. This integrated perspective avoids treating visual polish as equivalent to creativity and provides a stronger basis for comparing systems intended not only to generate images but also to participate in human creative practice.

To further illustrate the difference in creativity between an original artwork and AI-generated images, Fig. 5 compares Xu Beihong's *The Foolish Old Man Removes the Mountains* with images generated by the Kling and Jimeng platforms with the same text prompt. The original work in Fig. 5 features varied figures, expressive gestures, and a coherent spatial organization. The picture has distinct layers and harmonious spatial relationships, achieving an organic integration of realistic modeling and the brushwork language of traditional Chinese ink painting.

Kling's result in Fig. 5b retains the general theme elements of the original artwork, but shows a lower aesthetic quality, e.g., discontinuities between figures and the background, repetitive poses, and limited stylistic consistency with the original ink-painting language. Jimeng's result in Fig. 5c provides richer figure variation and stronger overall visual coherence. However, local distortions in faces and clothing remain, and some figure arrangements do not clearly support the narrative of collective labor.

These examples suggest that current models can reproduce broad visual elements from the prompt, but may still struggle to preserve compositional coherence, narrative relationships, and culturally specific artistic expression. The comparison is intended as a qualitative illustration rather than a quantitative evaluation.

3.4 Dynamic trade-offs among controllability, integrity, and creativity

The preceding sections have treated controllability, integrity, and creativity as distinct but complementary challenges. In interactive artistic creation, however, they are tightly coupled. Conditions that improve the predictability of generation may restrict exploratory variation; integrity safeguards may introduce constraints into user interaction and model outputs; greater generative autonomy may increase semantic drift and integrity-related risks. The objective is therefore not to maximize any single dimension or to identify a uni-

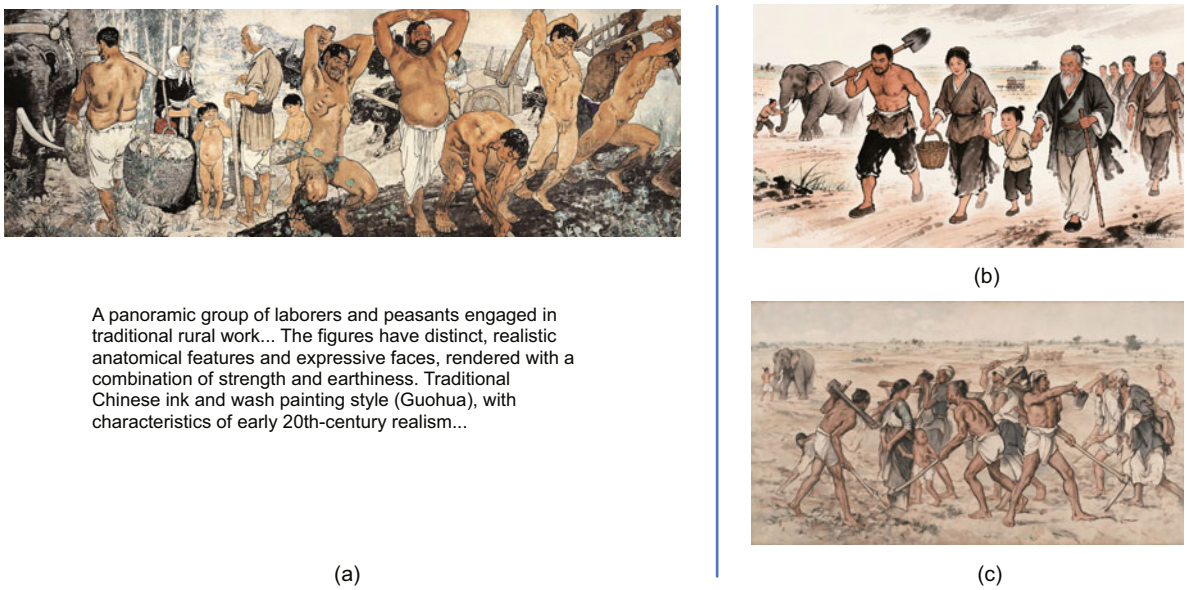


Fig. 5 Comparison between Xu Beihong's original artwork *The Foolish Old Man Removes the Mountains* and AI-generated images: (a) the original Chinese ink painting; (b) an image generated by Kling; (c) an image generated by Jimeng. Images in (b) and (c) were generated using the same text prompt derived from the original artwork

versal optimum, but to establish a dynamic equilibrium that varies with the creator's intention, the stage of the creative process, and the social and legal context of the task.

1. Controllability and creativity. Controllability enables creators to communicate semantic, structural, and stylistic intentions through prompts, sketches, layouts, masks, and other conditioning signals (Pan XG et al., 2023; Zhang L et al., 2023; Mou et al., 2024). Stronger conditions can improve fidelity and support targeted modification, but excessive specification may narrow the model's exploration space and reduce its contribution beyond the execution of predetermined instructions. This concern is consistent with evidence that exposure to highly developed AI-generated examples can, under some ideation conditions, increase fixation and reduce the variety of subsequent ideas (Wadinambiarachchi et al., 2024). Conversely, insufficient guidance can produce semantic drift, unintended changes, and repeated trial-and-error generation, weakening the creator's sense of agency (Shneiderman, 2022; Rezwana and Maher, 2023). Interactive systems should therefore represent control as a user-steerable continuum between precise realization and exploratory variation. Multimodal conditions should be available when fidelity is required but remain optional when users seek broader ideation, while selective editing and preservation mechanisms should allow creators to retain accepted elements as they explore alternatives (Hertz et al., 2022; Pan XG et al., 2023).

2. Integrity and controllability. Integrity safeguards constrain not only final outputs but also prompts, reference materials, editing operations, and model resources through which users exercise control. Such safeguards are necessary for addressing representational harm, unauthorized reproduction, provenance, and regulatory compliance (Carlini et al., 2023; D'Incà et al., 2024; U.S. Copyright Office, 2025b; Coalition for Content Provenance and Authenticity, 2026b). Nevertheless, context-insensitive restrictions can impose unnecessary friction when culturally specific, historical, or critical subject matter is mistaken for harmful intent. A purely restrictive response may force benign users to repeatedly reformulate inputs without clarifying which aspect of their request creates risk, thereby reducing interaction efficiency and meaningful user control. Integrity mechanisms should consequently incorporate contextual and multimodal evidence, distinguish the intended use of sensitive concepts, and communicate actionable reasons for intervention. Where possible, systems should support revision—for example, by identifying a risky region, unlicensed reference, or unsupported provenance claim—instead of rejecting the entire creative request. Provenance records, licensing information, and transparent safeguards can thereby function as control-supporting feedback, although such evidence cannot by itself determine legality, fairness, or cultural appropriateness (U.S. Copyright Office, 2025a; Coalition for Content Provenance and Authenticity, 2026a).

3. Integrity and creativity. Integrity also shapes the range of expressions that a generative system can explore. Cultural grounding can improve the contextual fidelity of artistic symbols, especially for traditions that are poorly represented in training data (Bayramli et al., 2025; Nayak et al., 2025). However, reducing cultural integrity to a fixed collection of approved motifs or canonical interpretations risks reinforcing stereotypes and overlooking alternative or community-specific meanings. Similar restrictions may arise from safety-oriented optimization and artist-specific rights protection: they can reduce harmful or unauthorized outputs, but they also alter the themes, styles, and combinations available for exploration (Shan et al., 2023; U.S. Copyright Office, 2025b). These effects should not be treated as evidence that weaker integrity necessarily produces greater creativity. Novelty based on cultural decontextualization, representational harm, or memorized reproduction does not satisfy the joint requirements of originality and contextual effectiveness (Runco and Jaeger, 2012; Carlini et al., 2023). Systems should instead distinguish tunable creative attributes—such as cultural explicitness, abstraction, stylistic distance, and the intensity of cross-cultural fusion—from non-negotiable requirements concerning rights, provenance, and harmful content. Cultural guidance may then support meaningful combinations without forcing every output toward a dominant or standardized aesthetic. Participatory review and the preservation of alternative interpretations are particularly important when living or marginalized cultural traditions are involved (UNESCO, 2003, 2021).

4. Toward a user-steerable and accountable equilibrium. The triangular relationship is consequently asymmetric rather than a simple zero-sum trade-off. High controllability does not guarantee integrity, because precise controls may also be used to produce biased or unauthorized content; integrity does not require maximal user control, because it also depends on data governance, provenance, rights management, and system-level safeguards; high creativity does not require the complete removal of constraints, because meaningful novelty depends on coherence and contextual appropriateness (O'Quin and Besemer, 1989; Runco and Jaeger, 2012). An effective co-creative system should therefore allow users to adjust the balance between fidelity and exploration while maintaining transparent integrity boundaries that cannot be delegated entirely to the user. Evaluation should correspondingly examine not only final-image quality but also condition fidelity, revision and exploration behavior, user agency, cultural and representational appropriateness, provenance robustness, and rights-related risks (Maher et al., 2018; Nayak et al., 2025; Coalition for Content Provenance and Authenticity, 2026a). The desired outcome is thus a user-steerable but accountable equilibrium in which human creators can move between precise reproduction and exploratory surprise without losing either meaningful agency or the protections required for sustainable artistic practice.

4 Comprehensive review of related methodological frameworks

4.1 Technical foundations for interactive visual generation

The development most relevant to interactive AI art is not the succession of individual model families, but the transition from autonomous image synthesis to generation that can be conditioned, revised, and incorporated into a human creative workflow. Early adversarial models demonstrated that neural generators could synthesize visually plausible images

from learned data distributions (Goodfellow et al., 2014), while style-based generators showed that intermediate latent representations could support attribute interpolation, style mixing, and coarse-to-fine manipulation (Karras et al., 2019). Discrete and continuous latent representations subsequently provided compact spaces in which visual content could be encoded, recombined, or modeled as tokens (van den Oord et al., 2017). Large-scale text-conditioned and latent diffusion systems extended these capabilities to open-domain image generation, combining pre-trained language representations with iterative visual synthesis (Ramesh et al., 2021; Saharia et al., 2022b; Rombach et al., 2022). From the perspective of artistic interaction, the important consequence of this development is that generative models increasingly expose intermediate representations and conditioning interfaces through which users can express and revise visual intentions.

Three technical capabilities are particularly important for interactive art creation. First, controllable representations make aspects of an image available for manipulation without requiring users to operate directly on pixels. StyleGAN's intermediate latent space, VQ-VAE's discrete visual codebook, and latent diffusion's compressed representation illustrate different ways of separating image generation from the full-resolution pixel space (van den Oord et al., 2017; Karras et al., 2019; Rombach et al., 2022). Such representations support interpolation, recombination, personalization, and reuse, but the meaning of their dimensions is not automatically transparent to artists: manipulation remains limited by the concepts and correlations learned from training data (Higgins et al., 2017; Karras et al., 2019). They should therefore be understood as enabling representations rather than complete user interfaces.

Second, multimodal and spatial conditioning enables creators to express intentions that are difficult to specify through text alone. Latent diffusion introduced cross-attention-based conditioning that can connect visual synthesis with text and other representations (Rombach et al., 2022), while ControlNet adds explicit spatial guidance from edges, depth maps, segmentation maps, and human poses to pre-trained diffusion models (Zhang L et al., 2023). These mechanisms allow an artist to combine semantic instructions with sketches, layouts, masks, poses, or reference structures. Systems such as Cocktail further demonstrate that several conditioning modalities can be fused and applied to designated image regions (Hu et al., 2023). This flexibility, however, creates a new interaction problem: simultaneous conditions may describe different abstraction levels or provide partially incompatible constraints, so successful control requires mechanisms for aligning, weighting, and spatially localizing their influence (Hu et al., 2023).

Third, local and iterative editing transforms generation from a one-shot request into a cycle of proposal, inspection, and revision. Prompt-to-Prompt shows that manipulating cross-attention maps can change selected textual concepts while retaining much of an existing spatial layout (Hertz et al., 2022). InstructPix2Pix instead learns to transform an input image according to a written editing instruction without requiring example-specific fine-tuning or inversion (Brooks et al., 2023). Token-based methods provide another route: MaskGIT predicts masked visual tokens in parallel and then refines them iteratively, supporting inpainting, extrapolation, and image manipulation while reducing the sequential decoding cost of earlier autoregressive generators (Chang et al., 2022). These methods establish the technical basis for generate–inspect–edit–regenerate workflows, but they do not by themselves provide a complete interaction design. PromptPaint, for example, was motivated by the difficulty of expressing visual concepts through language alone and demonstrated an interface in which prompts can be mixed and applied to different canvas regions and stages of generation (Chung and Adar, 2023). This illustrates how model control must be translated into interaction techniques that correspond to artists' iterative and spatial working practices.

Despite these advances, single-image controllability does not guarantee persistent control across an extended creative session. Text-guided editing research notes that even small prompt changes may produce unintended changes outside the requested concept, making preservation of accepted content a distinct technical objective (Hertz et al., 2022). Multi-turn generation introduces an additional problem: models must preserve character identity, layout, and contextual information while incorporating new instructions. TheaterGen identifies semantic and subject inconsistency as central limitations of multi-turn image generation and introduces explicit prompt and character management to mitigate them (Cheng et al., 2024). Thus, current systems still face interacting challenges of condition conflict, edit locality, cross-turn memory, identity preservation, and predictable control strength. These limitations arise across model families and are better analyzed as workflow-level problems than as properties of generative adversarial networks (GANs), variational autoencoders (VAEs), diffusion models, or autoregressive models in isolation.

The underlying architecture is therefore only one enabling layer of an interactive artistic system. Artistic usefulness depends on how representations, multimodal controls, editing mechanisms, feedback, and interaction history are combined. Likewise, fairness, provenance, cultural appropriateness, and content safety cannot be inferred from the generative architecture alone: they also depend on training data governance, alignment, filtering, interface design, and deployment policies. Rather than assigning universal ranking to model families, we select five representative generative models for experimental evaluation. Among them, the open-source models include Stable Diffusion v1.5, CogView4-6B, and FLUX.1-schnell, while the commercial models include Jimeng and Kling. We evaluate the generative capabilities of each model from three dimensions: controllability, aesthetic quality, and novelty. Specifically, we design 15 prompts covering different artistic styles and depiction contents, and generate 5 samples for each prompt. The average values of the evaluation results for each indicator are calculated. For the contrastive language-image pre-training (CLIP) similarity indicator, we use the CLIP-ViT-B/32 model to extract the semantic embeddings of the generated images and the corresponding prompts, and calculate the similarity between these two to measure the semantic consistency between the generated results and the text prompts. For controllability, aesthetic quality, and novelty, we use the Qwen-VL-Plus multimodal large model to score the generated images. The results are shown in Table 3. Stable Diffusion v1.5 achieves the highest score in the CLIP similarity indicator, indicating

Table 3 Quantitative comparison

Model	CLIP similarity	Controllability	Aesthetic quality	Novelty
Stable Diffusion v1.5	**0.3329**	0.7847	0.8040	0.6387
CogView4-6B	0.3011	0.7793	0.8193	0.6507
FLUX.1-schnell	0.3248	<u>0.8207</u>	<u>0.8367</u>	<u>0.6627</u>
Jimeng	<u>0.3328</u>	**0.8842**	**0.8642**	**0.6973**
Kling	0.3139	0.7833	0.8033	0.6267

A higher score indicates a better performance. The best results are highlighted in bold, while the second-best results are underlined

that its generated results have a higher semantic consistency with the input prompts; that is, it follows the prompts more closely. Also, Jimeng performs the best in the comprehensive evaluation based on large models, achieving the highest scores in controllability, aesthetic quality, and novelty, demonstrating the most comprehensive image generation capability.

As generative models become increasingly integrated into artistic and visual tasks, injecting structured knowledge has become pivotal in enhancing their interpretability, controllability, and expressive capacity (Lv et al., 2025; Song ZR et al., 2025). This subsection investigates how visual knowledge representations—spanning low-level perceptual cues, mid-level structural attributes, and high-level semantic or affective concepts as illustrated in Table 4—support visual recognition and generation. We first examine how models acquire and operationalize such knowledge, with a focus on mechanisms that learn to capture artistic style, formal structure, and salient visual attributes from curated datasets and domain-specific priors. We then discuss methods for modeling affective information in visual content, including affective knowledge graphs, emotion representation learning, and other techniques that enable generative models to align outputs with fine-grained emotional intent. By integrating insights from style analysis, art interpretation, and affective modeling, these approaches provide a foundation for more robust interactions between recognition and generation processes in visual AI systems, ultimately enabling models that better capture human-aligned structure and intent.

Table 4 Knowledge tiers→controllable generation mapping

Tier	Extractor	Control representation
Low-level	Estimator: edge, pose, seg, depth; palette extractors	Maps: edge, pose, seg, depth; palette vectors
Mid-level: style	Style classifiers; reference-image encoders	Style tags/embeddings; reference features
High-level: emotion	Emotion classifiers; affect lexicons graphs	Emotion vector or discrete affect label
Meta-knowledge	Art ontologies; critique models; rule engines	Scene graphs; rule sets; critique scores

Seg: segmentation

4.1.1 Low-level feature recognition

In computational art analysis, low-level visual features provide pixel-space regularities that summarize an artwork's measurable appearance: palettes (global and local color statistics), local textures (micro/macro textons), edges and contours (intensity gradients), and geometry proxies such as surface normals, monocular depth, and human pose keypoints. Classical descriptors—e.g., local binary patterns for texture, Canny-style gradients for edges, and histogram of oriented gradients (HOG)-like contour orientation summaries—remain useful, while modern monocular depth/normal predictors extend these cues to a coarse 3D structure. Early convolutional neural network (CNN) layers are known to behave as color/edge/texture detectors, so deep activations complement hand-crafted features when building style-sensitive embeddings for paintings (Canny, 1986; Ojala et al., 2002; Dalal and Triggs, 2005; Zeiler and Fergus, 2014; Ranftl et al., 2021).

These low-level cues can be distilled into structured controls suited for downstream models: edge/line maps (e.g., Canny/line segment detector (LSD)) to capture linework and contours; human pose keypoints (e.g., OpenPose) for figure layout; semantic segmentation masks for region-wise style/texture gating; depth (and, where available, normals) to anchor composition in coarse 3D. Additionally, palette vectors and compact texture codes summarize color harmony and brushstroke microstructure for appearance control (Canny, 1986; von Gioi et al., 2010; O'Donovan et al., 2011; Long et al., 2015; Chen LC et al., 2018; Ranftl et al., 2021).

In controllable art-generation pipelines, these signals are injected as structural side-branches or adapters that modulate a frozen text-to-image backbone: ControlNet adds spatial conditioning paths (e.g., edges, depth, segmentation, and pose), while T2I-Adapter supplies lightweight adapters that align external maps with internal features. Appearance can be steered by palette-aware prompting or direct palette-conditioning via palette vectors or palette-to-image modules (Bahng et al., 2018; Zhang L et al., 2023; Mou et al., 2024).

For computational art analysis, precise control over spatial layout (contours, segmentation, pose, and depth) and palette/texture statistics enables reproducible, editable, and deterministic workflows: analysts can isolate brushstroke geometry from color design, enforce palette constraints for comparative studies, and generate counterfactuals that test hypotheses about style, composition, or authorship—turning low-level evidence into rigorous, controllable experiments.

4.1.2 High-level feature analysis

Early works in affective image analysis demonstrated that psychological features such as color harmony, brightness, and object themes correlate with the emotions evoked by images. One seminal study introduced the ArtPhoto dataset, consisting of 806 artistic photographs labeled with discrete emotions, and showed that combining art-inspired features (e.g., warm/cool

colors and composition geometry) with content descriptors could predict viewer emotions to some extent (Machajdik and Hanbury, 2010).

With the advent of deep learning, CNNs and vision Transformers have become dominant tools for visual style recognition. Large-scale art datasets such as WikiArt have enabled models to classify artworks by style or artist with high accuracy (Tan WR et al., 2019). CNN-based classifiers have achieved strong performance in recognizing fine-grained artistic styles and genres from WikiArt images. For example, a VGG16-based approach reached state-of-the-art results in style recognition by learning discriminative brushstroke and color patterns for movements such as Impressionism versus Cubism. More recently, studies have explored Transformers and multimodal pre-training to enhance art understanding. The CLIP model—pre-trained on image–text pairs—has been leveraged to perform zero-shot classification of artworks by style, artist, and medium (Radford et al., 2021). Such multimodal models can match paintings to textual style descriptions (e.g., "Baroque" and "Surrealist") without explicit fine-tuning, indicating that language supervision can effectively encode art-related knowledge.

Beyond predicting style labels, research has increasingly shifted toward explaining why an image exhibits a certain style or evokes a specific effect—a step toward automated art criticism. Explainable models integrate human-understandable attributes and expert art knowledge. For example, Costa et al. (2023) developed the explainable style classifier ANYXI, which uses logical rules derived from art history to justify its predictions. This model encodes art specialists' knowledge of styles, correlates it with low-level image features, and produces rationales for style classifications. Another approach represents each style through a set of signature visual attributes—such as color palette, stroke texture, and subject matter—allowing semantic comparisons between styles. These representations can power content-based art recommendation and automated critique generation.

A notable example is Achlioptas et al. (2021)'s ArtEmis dataset, which explores the interplay between an artwork's visual content, its perceived emotional impact, and viewers' free-form explanations. ArtEmis contains 80 000 WikiArt paintings annotated with a dominant emotion label and a natural-language explanation describing why that emotion was felt. These grounded explanations often reference subtle visual aspects or abstract interpretations—for example, "the gloomy sky and muted colors give a feeling of sadness." Training captioning models on ArtEmis enables them to generate critiques that reflect both emotional and semantic content, going beyond literal object descriptions. This illustrates how combining affective and contextual knowledge with visual recognition allows AI systems to engage in rudimentary "art criticism."

Recent models have further emphasized fusing stylistic features with high-level content to improve visual understanding. Jing et al. (2023) proposed a style-guided network for emotion distribution prediction in images, called StyleEDL, which extracts hierarchical style representations and builds a stylistic graph convolutional network. Their results demonstrated that incorporating style knowledge significantly enhances the ability to understand an artwork's affective impact.

In addition to style analysis, visual sentiment analysis has made significant progress. Yang JY et al. (2023) compiled EmoSet, a dataset of 3.3 million images annotated with emotion categories. Each image was also labeled with interpretable attributes, grounded in psychological research linking visual characteristics to mood. Integrating these attributes into CNN models improved emotion prediction while providing interpretable rationales—for instance, "image evokes awe because it depicts a grand natural scene with bright colors."

To capture subtler emotional cues, Man et al. (2025) introduced K-EVER2, a knowledge-enhanced visual emotion reasoning framework. This approach integrates an affective knowledge graph derived from SenticNet and aligns visual and textual cues with graph nodes during inference, enabling more robust emotion classification.

The field's frontier now extends to emotion-guided image generation. Yang JY et al. (2024) proposed EmoGen, which defines an emotion space aligned with the CLIP semantic space. By mapping emotional vectors to text prompts, EmoGen conditions diffusion models to generate images that match intended emotional effects. Evaluations using metrics such as emotion accuracy and semantic clarity show substantial improvements over standard diffusion-based approaches.

4.1.3 Definition and significance of cultural genes

Cultural genes in art refer to the artistic and cultural factors accumulated through long-term historical and artistic practices. They encode regional and ethnic value orientations, aesthetic preferences, stylistic conventions, and conceptual ideologies. At the level of concrete artistic production, cultural genes manifest as artistic styles, representational forms, symbolic motifs, and related attributes. As cultural-level hereditary factors, they exhibit stability, transmissibility, and evolvability, shaping human artistic creation across generations.

In the digital-intelligence era, cultural genes can be digitized, stored, transmitted, and regenerated, enabling new computational workflows for generative art. Within generative AI systems, cultural genes play three primary roles:

1. Data-level encoding of artistic cultural genes. Cultural genes—embodied in historical artworks and their interpretations—constitute a rich reservoir of human artistic knowledge. By training on image corpora containing diverse stylistic and cultural signatures, along with accompanying art-historical and interpretive text, generative models can construct a digitized, multicultural, and inclusive cultural-gene repository that serves as the data foundation for controllable generation.

2. Feature-level style understanding and construction. Because cultural genes are relatively stable, visual features in historical artworks (e.g., color palettes, compositions, and brushwork) often correspond consistently to particular artistic styles. Through multimodal feature fusion, generative models can learn these associations between visual attributes and culture-specific semantics, enabling controllable image generation that inherits and evolves human artistic cultural genes at the feature level.

3. Semantic-level modeling of cultural logic. Beyond style features, cultural genes encode deeper artistic knowledge—value systems, conceptual thinking, aesthetic principles, and

experiential insights. These factors form a semantic network that has guided artistic creation for centuries. Incorporating art-historical knowledge graphs allows generative models to capture relations among cultural genes, stylistic lineages, and cultural contexts, enabling models to generate outputs with coherent cultural logic rather than surface-level stylistic imitation.

4. Applications in generative art. Initial work has begun integrating cultural-gene modeling into generative systems. For instance, in AI-generated content (AIGC) design for intangible cultural heritage (ICH), models decompose visual features of ICH motifs and assign semantic labels to establish cultural constraint rules, ensuring cultural fidelity while supporting creative diversity (Zheng HZ, 2025). Similarly, research on the cultural-gene phylogeny of Changsha kiln plant patterns employs a dual-track framework of explicit cultural genes (pattern, color, and decorative technique) and implicit cultural genes (social or religious semantics) to build a structured motif lineage model that supports effective digital-era inheritance of ICH (Cui et al., 2025).

4.1.4 Fairness and bias

1. Data and algorithmic biases. Bias in generative models is not limited to cultural symbols. Data collection and model design can amplify societal inequalities. Wan et al. (2024b)'s survey notes that by 2025, 90% of Internet content may be AI-generated, and that current models exhibit severe biases. These biases manifest along gender, race, and occupation: Stable Diffusion portrays white male chief executive officers (CEOs) leading corporations, while dark-skinned individuals are depicted as criminals or service workers. The survey warns that such stereotypes can lead to both representational harms—misrepresenting individuals—and allocational harms, such as discriminatory outcomes when AI is used in hiring or law enforcement (Wan et al., 2024b). In a more general context, Buolamwini and Gebru (2018)'s gender shades audit of commercial gender-classification systems found that error rates for classifying gender were below 0.8% for light-skinned men but exceeded 34% for dark-skinned women. These disparities show that unbalanced training data lead to inaccurate predictions for marginalized groups, a problem that can carry over into generative models that rely on the same corpora.

The representation of gender identity reveals another dimension of bias. Ungless et al. (2023) compared outputs of three image-generation models for prompts containing cisgender versus non-cisgender identity terms. They found that non-cisgender identities were consistently misrepresented as less human, more stereotyped, and more sexualized. Their survey of non-cisgender individuals indicated that participants were concerned about misrepresentation and harmful stereotypes and called for community involvement, curated training data, and customization to mitigate these harms (Ungless et al., 2023).

2. Fairness, diversity, and reliability. While fairness is often conflated with accuracy, Jordan Vice and colleagues argue that it is distinct from reliability. They proposed an evaluation framework that uses global and local perturbations in the embedding space to probe model behavior. Their method measures generative diversity, defined as the breadth of visual representations for learned concepts, and generative fairness, which examines how removing concepts from prompts affects control over outputs under a low-guidance setup (Vice et al., 2026). The framework can detect unreliable or bias-injected models and trace the provenance of embedded biases. Importantly, the authors noted that fairness and reliability are interrelated but distinct: a model may be reliable in producing consistent outputs yet systematically biased. They also highlighted the threat of backdoor or intentionally biased models, which may require advanced detection strategies (Vice et al., 2026).

By integrating metrics for fairness, diversity, and reliability, researchers can compare models across demographic groups, occupations, and cultures. However, Wan et al. (2024b) cautioned that evaluation datasets and metrics are "scattered" and that no unified framework exists. They called for human-centric approaches to define, evaluate, and mitigate bias (Wan et al., 2024b). Such approaches include co-designing evaluation metrics with affected communities, using participatory audits and employing intersectional analysis (considering gender, race, class, and culture together) instead of treating each dimension independently.

3. Fairness mitigation strategies. Mitigating fairness issues requires actions along the entire model pipeline. On the data side, developers must curate balanced datasets that include diverse demographic groups, occupations, and cultural contexts. Data augmentation techniques can help but must avoid reinforcing stereotypes. On the algorithmic side, researchers are exploring debiasing methods such as fairness constraints during training, adversarial approaches that penalize differences across protected groups, and selective fine-tuning (e.g., selectively updating model weights to reduce bias). Post-processing techniques, such as adjusting sampling strategies or blending outputs with fairness filters, can also improve fairness. The dialogue with the machine approach described by Qadri et al. (2024) offers a participatory strategy, in which artists and critics deliberately test models for bias and adjust prompts or generation settings accordingly.

4.1.5 Cultural challenges in intelligent creative systems

1. Misappropriation and misuse of cultural symbols. Because generative models learn style and iconography from global image collections, they can "borrow" cultural motifs without understanding context. The exploratory study *Imagining the Far East: Exploring Perceived Biases in AI-Generated Images of East Asian Women* audited three popular models (DALL·E, Midjourney, and Stable Diffusion) using prompts about East Asian women. The researchers invited East Asian users to review outputs and identified 18 perceived biases in four patterns: westernization, overuse or misuse of cultural symbols, sexualization and feminization, and racial stereotypes (Lan et al., 2025). Participants reported that many AI images depicted East Asian women with Western facial features, heavy makeup, and American fashion, erasing distinct regional aesthetics. Models frequently misused cultural elements (e.g., qipao with incorrect patterns or blending unrelated East Asian iconography), trivializing sacred symbols and homogenizing diverse cultures (Lan et al., 2025). Such cultural appropriation has legal ramifications because many indigenous

communities consider traditional designs and religious symbols to be collective cultural property subject to customary law.

Similar concerns appear in broader surveys. Wan et al. (2024b) observed that geo-cultural bias remains under-explored in text-to-image research. They noted that current models default to Western cultures and under-represent others, reinforcing dominant cultural narratives and potentially erasing marginalized communities. A typical bias is over-representation of white masculinity: a version of Stable Diffusion depicted the world as being run by white male CEOs, while dark-skinned men were portrayed committing crimes and dark-skinned women flipping burgers. These biases create representational harms by perpetuating stereotypes and failing to reflect the diversity of global cultures (Wan et al., 2024b).

2. Culturally situated evaluation and art-world engagement. Standard benchmarks for generative models often overlook social and cultural contexts. Qadri et al. (2024) argued that evaluations must be dialogic and involve multiple art worlds. Their method proposes two dialogues: (1) "dialogues with art worlds," in which artists engage with art historians, curators, and archivists to interpret generative outputs; (2) "dialogues with the machine," where artists and critics experiment systematically with AI tools (Qadri et al., 2024). They demonstrated this approach through a case study involving artists from non-Western art worlds in the Persian Gulf. The dialogues highlight how AI can reproduce biases embedded in Western art canons and how artists can repurpose models by adjusting prompts, style guides, or curation to align outputs with their cultural context (Qadri et al., 2024).

Culturally situated evaluation addresses two compliance issues. First, it helps detect cultural appropriation by comparing AI-generated imagery with traditional motifs, identifying misused or stereotyped symbols. Second, it offers a participatory mechanism for communities to influence how AI tools represent them. Qadri et al. (2024) emphasized that involving non-Western art worlds in evaluation can help build "inclusive machines" for diverse art practices. In legal terms, this participatory approach supports free prior and informed consent, a principle in international law that requires obtaining indigenous peoples' consent before using their cultural expressions. Engagement with cultural stakeholders can also inform policies for licensing and data use agreements for culturally sensitive content.

3. Mitigating cultural harms. Mitigation strategies should address both the supply of training data and the generation process. On the supply side, curators need to compile culturally diverse, ethically sourced datasets with metadata describing context and usage restrictions. Collaborative initiatives with museums, indigenous communities and artists can produce training corpora that respect cultural protocols and avoid misappropriation. On the generation side, models can incorporate cultural control variables or style embeddings that allow users to specify cultural contexts transparently. Developers should implement filters that prevent the combination of sacred symbols with unrelated elements. Finally, evaluation metrics should measure cultural authenticity and representation rather than focusing solely on visual fidelity.

4.2 Intelligent creative systems and interactions with AI agents

4.2.1 Milestone generative models and conditioning paradigms

Controllable visual generation refers to the class of generative models that allow users or external inputs to steer the image synthesis process in a desired direction. Rather than producing images arbitrarily, these models accept conditioning signals—such as class labels, text descriptions, semantic masks, poses, or sketches—and generate images that satisfy those conditions. Early approaches to controllable generation were based on GANs. For example, conditional GANs extended the GAN framework by feeding class labels or images to both the generator and discriminator, enabling image-to-image translation tasks like pix2pix and CycleGAN for paired and unpaired image domains, respectively (Isola et al., 2017; Zhu JY et al., 2017). These works demonstrated the power of learned translation for tasks like turning sketches into photos or translating between day and night scenes, but they often required carefully tuned training and still offered limited control over fine-grained image attributes.

Through the late 2010s, GANs were the dominant approach for high-quality image synthesis. Milestone models such as StyleGAN (Karras et al., 2020) achieved unprecedented fidelity and diversity in face and graphics generation by learning a rich intermediate latent space that could be manipulated. Building on StyleGAN, researchers developed methods to interpret and manipulate the latent semantics in a continuous manner: for example, InterFaceGAN (Shen et al., 2020) discovered linear directions in the latent space corresponding to human-interpretable attributes (like age or smile), enabling continuous attribute editing of GAN-generated faces, while GANSpace (Härkönen et al., 2020) used principal component analysis (PCA) in latent feature space to find orthogonal control axes for aspects like pose or lighting. These methods demonstrated the principle of continuous conditioning over generation, where sliding a latent variable causes a smooth change in the output image. Beyond faces, conditional GANs tackled various structured control inputs. For instance, SPADE/GauGAN (semantic image synthesis) (Park et al., 2019) introduced spatially adaptive normalization for conditioning on segmentation maps (rendering photorealistic scenes from layouts), and other conditional GAN frameworks allowed pose-guided person image generation or sketch-to-image synthesis (Zhang JS et al., 2021). However, GAN training is notoriously fragile, and conditional GANs often suffer from mode dropping or require task-specific architectures.

Around 2020, diffusion probabilistic models emerged as a powerful alternative, offering superior coverage of the data distribution and simpler training objectives. The seminal denoising diffusion probabilistic model (DDPM) (Ho et al., 2020) introduced a likelihood-based method that incrementally refines noise into images. Diffusion models initially focused on unconditional generation, but soon incorporated conditions through guidance strategies. Dhariwal and Nichol (2021) showed that diffusion models can surpass GANs in image fidelity and diversity on ImageNet, especially when using classifier-based

guidance (modulating the sampling process with gradients from a classifier to steer the output towards a target class). Subsequent advances like score-based generative models (Song Y et al., 2021) unified diffusion and score-matching perspectives, laying the groundwork for flexible conditioning. A breakthrough in text-conditional generation came with OpenAI's GLIDE and DALL·E systems, and Google's Imagen, which leverage diffusion models with powerful pre-trained text encoders. For example, Imagen (Saharia et al., 2022b) employed a large Transformer language model (T5) to encode text prompts and guided a cascaded diffusion model, achieving astonishing photorealism and semantic alignment with textual descriptions. Another influential approach, latent diffusion model (LDM) by Rombach et al. (2022), showed that diffusion operating in a learned lower-dimensional latent space (coupled with a vision autoencoder) can greatly improve efficiency while retaining quality. Latent diffusion with cross-attention conditioning enabled versatile controllers: a single model can take text, bounding boxes, segmentation maps, or other modalities as inputs by injecting them through cross-attention layers (Rombach et al., 2022). This architecture became the foundation for Stable Diffusion, a widely-used open-source text-to-image model. Extensions of diffusion models have further increased controllability: notably, ControlNet (Zhang L et al., 2023) introduced a trainable network appended to a frozen diffusion model to accept additional inputs (edge maps, human poses, depth maps, etc.) and guide the generation without affecting the original model's weights. By using "zero-convolution" layers and augmenting the diffusion U-Net, ControlNet enabled one-to-one mapping from an input condition (like a sketch) to an output image while preserving the benefits of a pre-trained text-to-image diffuser. Similarly, T2I-Adapter and other related methods (e.g., IP-Adapter and Palette) explored lightweight modules to condition diffusion models on extra modalities, illustrating a growing toolkit for controllable image synthesis in the diffusion era.

Alongside GANs and diffusion, Transformer architectures have also driven controllable generation. Ramesh et al. (2021)'s DALL·E was a landmark in this category: it utilized a Transformer to autoregressively generate image tokens (from a learned discrete VAE codebook) from a text prompt, demonstrating zero-shot text-to-image synthesis. This approach treats conditioning as sequence priming—text tokens followed by image tokens—allowing a single model to generate diverse images for a given caption. While powerful, autoregressive Transformers suffer from slow synthesis and sometimes inconsistent global structure. To address these issues, Chang et al. (2022) introduced MaskGIT, a masked generative image Transformer that generates all tokens in parallel and iteratively refines an image by conditioning on progressively fewer masks. MaskGIT's bidirectional design yields significantly faster sampling and was shown to handle editing tasks like inpainting or outpainting by treating known regions as fixed tokens. Another noteworthy model is VQGAN (Esser et al., 2021), which combined a convolutional VQ-VAE with a Transformer to model the codebook indices; although not a pure Transformer approach, it highlighted how combining continuous and discrete representations can facilitate both high fidelity and control (the codebook enables certain editing operations, and the Transformer can incorporate conditioning, including text via contrastive models like CLIP). Transformer-based generators have also been integrated with diffusion pipelines (e.g., Transformer text encoders in guided diffusion, or hybrid architectures in DALL·E 2 (Ramesh et al., 2022)), illustrating that modern controllable generation often blends ideas from these paradigms.

4.2.2 Human-in-the-loop collaboration in generation

The ultimate form of controllable generation is an interactive loop between a human user and a generative model, where the model not only follows initial instructions but can be directed, corrected, or adjusted in real time. Recently, the research community has put increasing emphasis on such human–computer collaboration for image creation. One line of work focuses on instruction-driven editing of images. InstructPix2Pix (Brooks et al., 2023) is a prime example that trains a diffusion model specifically to perform edits on an input image given a written instruction (e.g., "turn this photo into a painting of the same scene" or "remove the person on the right"). This model was built by generating a large synthetic training set of (image, instruction, and edited image) triplets using pre-trained models, and it can generalize to user-provided images at test time, producing the desired edit in one forward pass. Such an approach brings the model closer to a usable assistant, where a researcher or artist can iteratively apply natural language commands to refine an image.

Another avenue is fine-tuning or adapting models to user-specific content—a form of personalization that involves the human providing a small dataset. DreamBooth (Ruiz et al., 2023) pioneered a method for injecting a custom visual concept (e.g., a particular person, object, or style) into a text-to-image diffusion model by fine-tuning on just a few images of that concept. The user supplies 3–5 images of a new subject, and DreamBooth adapts the model so the subject can be regenerated in different contexts under a special identifier. This can be seen as a collaborative process: the human teaches the model a concept through examples, and the model grants control to the human to generate novel images including that concept. Subsequent works have improved efficiency (textual inversion and low-rank adaptation (LoRA)-based fine-tuners) or expanded to styles, showing a trend towards users co-creating content by personalizing models with minimal effort.

Finally, there is a stream of research on real-time interactive tools. These include graphical user interfaces where a human can sketch or erase parts of an image and the AI will regenerate the region (with diffusion models ensuring realism and consistency), or systems where a user and an AI "co-draw" an image. For example, the earlier GANPaint system (Bau et al., 2019) allowed users to add or remove objects in a GAN-generated scene by simply clicking regions (e.g., draw a tree and the GAN completes it realistically); now diffusion-based equivalents achieve even more convincing results and can be conditioned on text for style guidance. Some works propose multi-turn dialogs with image models (where the model asks for clarifications or additional input from the user), although this remains a nascent area. The general goal of these human-in-the-loop approaches is to harness the generative power of modern models while keeping the human's intentions and creativity

at the forefront. As controllable visual generation progresses from simple conditioning to full collaboration, we envision tools that act as intelligent "co-creators," enabling users to continuously steer and refine the generative process through a mix of high-level instructions and low-level adjustments.

In summary, controllable visual generation has evolved rapidly in recent years, propelled by advances in GANs, diffusion models, and Transformers. Researchers have achieved increasingly fine-grained control—from continuous latent sliders in GANs to complex multimodal conditions in diffusion frameworks—and applied these capabilities to a plethora of vision tasks. The trend is clearly toward greater interactivity and alignment with user intent: models are no longer just one-shot image samplers, but components of iterative creative workflows. Ongoing work in human–AI collaboration, feedback mechanisms, and more interpretable controls will further bridge the gap between a user's imaginative intent and the synthesized visuals, opening exciting opportunities for both scientific research and creative applications.

4.2.3 Protecting generated content: from digital watermarking to decentralized approaches

Generative models capable of synthesizing art from textual prompts have rapidly advanced and are now integral to graphic design, advertising, and personalized art. However, these systems raise pressing compliance questions because they learn from and reproduce massive collections of existing images without clear provenance. Copyright, cultural, and fairness considerations are inherently interconnected in the creation and dissemination of AI-generated works. This subsection surveys academic literature, legal analyses, and policy reports to identify the primary compliance challenges and propose future directions.

1. Training-stage data infringement. Most high-quality generative models, including diffusion-based models (e.g., Stable Diffusion) and multimodal models (e.g., DALL·E and Midjourney), are trained on datasets collected from web-scraped images. Stober and Dornis (2025) noted that generative AI systems are typically trained on "extensive amounts of data" collected through web scraping and that many of these images are copyright protected. They argued that developers often rely on the U.S. fair use doctrine or European text and data mining (TDM) exceptions, but generative training differs from traditional TDM. Whereas TDM extracts patterns from data, generative models learn probability distributions and can reproduce or "memorize" copyrighted works. Because generative models reconstruct or approximate the training data rather than simply mine them, Stober and Dornis (2025) contended that existing TDM exceptions may not apply.

The legal uncertainty is compounded by the fact that training corpora are rarely disclosed. In an ongoing lawsuit, the Recording Industry Association of America alleged that the music-generation service Suno trained on "essentially all music files of reasonable quality that are accessible on the open Internet," arguing that Suno used copyrighted recordings behind paywalls and password protections (Stober and Dornis, 2025). Suno defended itself by invoking fair use, claiming that copying protected works as part of a back-end technological process is permissible (Stober and Dornis, 2025). This case illustrates the tension between the need for large training datasets and the rights of creators. Desai and Riedl (2024) emphasized that fair use does not sanction all ways of accessing copyrighted material; they stressed that computer scientists must discipline data practices and that copyright industries may need to relax their assumption that all uses must be licensed.

An additional challenge is "memorization." When a model inadvertently reproduces portions of copyrighted training images, it creates potential liability. Stober and Dornis (2025) noted that sharing models containing memorized training data or generating outputs that include memorized content could infringe copyright. The risk is heightened when training corpora contain proprietary illustrations or photographs whose provenance cannot be traced because the data were scraped from websites without consent. A promising direction involves transparent documentation and data provenance tools. Stober and Dornis (2025) proposed documenting the sources of training data and leveraging music-information-retrieval methods for dataset attribution, which would allow rightsholders to verify whether their works have been used and request removal.

2. Copyrightability and ownership of generated works. Beyond training, ownership of AI-generated outputs is contentious. Generative tools' terms of service often require users to grant an irrevocable license to both input prompts and generated outputs. Ambartsumian and Cannon (2025) noted that while prompts and training data may be copyrighted, AI outputs often lack human authorship and therefore may not be protected. The U.S. Copyright Office reiterated that only material that is the product of human creativity is copyrightable (Ambartsumian and Cannon, 2025); prompts alone do not provide sufficient human control to make the user the author. Courts require that a human exercise creative judgment in the selection or modification of the AI-generated content. Thus, images produced solely through AI in response to text prompts are generally not copyrightable in the United States (Ambartsumian and Cannon, 2025). This situation undermines artists' ability to control and monetize their AI-assisted work.

Comparative studies reveal divergent approaches internationally. The Harvard analysis notes that while the U.S., the EU, and the UK insist on some level of human creativity, China and Japan may be more willing to recognize AI-assisted works. In China, copyright protection may be granted to AI-generated works if a human plays a central role in the creative process; Japan's copyright law is more permissive because it emphasizes the result rather than the creative process. By contrast, U.S. law relies on the constitutional requirement that "authors" produce "writings;" the Copyright Office views the authorship requirement as excluding non-humans (Ambartsumian and Cannon, 2025). These differences create legal uncertainty for artists distributing AI-generated images across borders.

4.2.4 Representative applications of controllable generation

The ability to control image synthesis has unlocked a wide range of applications, each with tailored models and techniques. Text-to-image synthesis is perhaps the most publicized application, where a user provides a caption and the model generates a corresponding image. Earlier GAN-based methods for this task (AttnGAN (Xu T et al., 2018) and DF-GAN

(Tao et al., 2020)) made progress, but recent diffusion models have dramatically advanced the state of the art. Models like DALL·E 2 (Ramesh et al., 2022) and Imagen (Saharia et al., 2022b) can produce high-resolution, fantastical imagery from free-form text, while adhering closely to the described content. These models often employ classifier-free guidance (Ho and Salimans, 2022) at inference, effectively allowing the user to dial up or down the influence of the text conditioning continuously (trading off between image–text alignment and image realism).

Beyond text, pose-guided image generation is an important application in domains like human motion synthesis and animation. Here, the goal is to generate an image of a person (or object) in a specified pose, often given by a 2D or 3D skeleton. Early approaches used tailored GAN frameworks; for example, PISE (Zhang JS et al., 2021) introduced a decoupled GAN that first generates a coarse person image from the pose and then refines details with an appearance encoder, enabling both pose transfer and attribute editing. More recently, diffusion models have been applied to pose-guided generation: one can condition a latent diffusion model on human skeletal keypoints or dense pose maps by concatenating pose channels into the diffusion model input (Rombach et al., 2022), or by using a ControlNet trained specifically for pose conditions (Zhang L et al., 2023). These diffusion-based methods have been shown to produce sharper and more coherent results, thanks to the mode coverage of diffusion and the ability to integrate with powerful pre-trained text models (e.g., generating a person in a given pose and style as described by a text prompt).

Another popular paradigm is image-to-image translation with structural guidance, which includes tasks like sketch-to-image, segmentation-to-image, and depth-to-image. Controllable generation in this setting means that the model takes an outline or map as input and fills in photorealistic details. The SPADE normalization technique (Park et al., 2019) is a key innovation for GANs, enabling models like GauGAN to produce vivid scenery from semantic label maps. As diffusion models rose to prominence, they quickly proved adept at such tasks as well. For example, the latent diffusion framework (Rombach et al., 2022) demonstrated that simply concatenating a low-resolution segmentation mask to the latent can guide the generation to respect that layout. Subsequent works like Palette (Saharia et al., 2022a) and Pix2Pix-zero (Parmar et al., 2023) leveraged diffusion for general image-to-image tasks including colorization, super-resolution, and inpainting, often outperforming task-specific GANs. With the introduction of ControlNet, sketch-to-image synthesis has become even more accessible: a user's sketch or edge map can be fed into a ControlNet-augmented Stable Diffusion model to yield a faithful but realistically colored and textured image following the outline—essentially an AI-powered "auto-draw" tool. Notably, these models preserve continuous control: by varying the intensity of the condition (for instance, interpolating between two edge maps or adjusting how strictly the model follows the sketch via guidance scale), one can fine-tune the balance between adhering to the input structure and allowing creative freedom.

Image inpainting and outpainting are further applications where controllability is crucial. In inpainting, the model fills in a missing or masked region of an image in a coherent way, guided by the surrounding pixels and optionally by user instructions. GAN-based inpainters (e.g., partial convolution networks, or dedicated conditional GANs like EdgeConnect (Nazeri et al., 2019) and CoModGAN (Zhao SY et al., 2021)) were successful on limited masks, but often struggled with large holes or required post-processing. Diffusion-based inpainting (such as in OpenAI's GLIDE and Stable Diffusion's inpainting model) leverages the iterative refinement property to naturally hallucinate missing content with excellent adherence to context. The user can control what to inpaint by providing a mask and can also supply a text prompt to guide the filling (e.g., "insert a red sofa here"). Paint by Example (Yang BX et al., 2023) further extended diffusion inpainting to use an exemplar image patch as guidance, so that the filled region takes on the texture or style of the example—an advanced form of conditional control mixing visual and spatial inputs. Outpainting (extrapolating beyond the image borders) similarly benefits from diffusion models' flexibility, enabling creative extension of images with user control over the content via text or example images.

Recent years have witnessed a rapid transition of AI-driven artistic image generation from experimental research to large-scale commercial deployment across creative industries. For instance, Adobe integrated generative models into its Firefly and Creative Cloud ecosystem, enabling commercially safe, style-controllable image synthesis for advertising, branding, and digital design workflows. In the beverage industry, Coca-Cola employed generative image models in campaigns such as its AI-assisted "Y3000" branding initiative, where machine-generated futuristic aesthetics were incorporated into product packaging and global marketing materials. In fashion and media, publishers including *Vogue* have experimented with AI-generated visual content for editorial spreads, while independent designers have leveraged tools such as Midjourney and DALL·E to prototype concept art and reduce production cycles. In sportswear and consumer branding, Nike and similar brands have explored generative design pipelines to accelerate visual ideation for promotional assets and product visualization. Collectively, these cases demonstrate that AI-based artistic image creation has evolved from a purely experimental technology into a commercially viable creative infrastructure, supporting cost-efficient content generation, rapid prototyping, and scalable personalization across advertising, entertainment, publishing, and design sectors.

5 Prospects

Amid rapid advances in generative models, interactive art-creation systems still face persistent challenges in reliable cross-modal control, fine-grained and emotionally expressive generation, consistent evaluation, and the responsible handling of safety, privacy, and cultural concerns. Developments in adjacent AI fields point to a broader system-level transition from static, task-specific intelligence toward dynamic architectures characterized by continual state evolution, knowledge acquisition from data, persistent memory, interactive evaluation, and privacy-preserving collaboration (Xu JF et al., 2023; He et al., 2024; Liu et al., 2024; Peng et al., 2024; Zheng QH

et al., 2025). Following this broader shift from static generation toward continuously evolving intelligent systems, we identify four complementary directions for the future development of AI-driven artistic creation: (1) advancing core generative methods beyond passive synthesis by integrating Transformer-centric architectures, multi-agent pipelines, and reinforcement learning from human feedback (RLHF), so that agents can plan, critique, and iteratively refine outputs with improved controllability, intentionality, and stylistic coherence (Wu F et al., 2025); (2) coupling data-driven learning with knowledge-directed generation by incorporating structured art theory, including color and composition principles, emotion ontologies, and knowledge graphs through retrieval and reasoning, thereby improving semantic grounding and reducing representational bias (Pan YH, 2020); (3) orchestrating collaboration between general-purpose foundation models and specialized, domain-tuned models through ensembles, cascades, and distillation to balance open-world understanding with stylistic fidelity and cultural specificity; (4) enabling sustained human–AI co-creation through dialogue, sketch refinement, and continuous preference learning, further supported by spatial intelligence for 3D/augmented reality (AR)/virtual reality (VR) and scene-aware interaction, gesture interaction for embodied and performative control, and on-device real-time creation for privacy, low latency, and personalization (Pan YH, 2019; Wang YW et al., 2024). Together, these directions suggest that future AI art systems will increasingly evolve from isolated image generators into collaborative, knowledge-empowered, memory-aware, and continuously adaptive creative systems, ultimately supporting a human–AI co-agency framework that balances controllability, creativity, and responsible use.

5.1 Future direction of visual art generative model methods

The development of visual art generative models is expected to move beyond mere image synthesis toward interactive, expressive, and controllable artistic creation. While current models, such as GANs and diffusion-based architectures, can produce visually plausible content, they often lack semantic depth, intentionality, and stylistic coherence. Future methods must address challenges of fine-grained control, multimodal guidance, emotional expressiveness, and cultural specificity. With recent advancements in next-generation model architectures (e.g., Mamba (Gu and Dao, 2024) and spiking neural networks (Zhu RJ et al., 2024)), model merging of expert models (Wortsman et al., 2022), multi-agent systems, RLHF, and continuous conditional generative models (Ding et al., 2026), generative art models will evolve from passive generators into collaborative and creative agents.

5.2 Data-driven and knowledge-directed artistic creation

Future generative systems will not only rely on large-scale data but also be enriched with structured domain knowledge. Data-driven learning alone leads to overfitting to stylistic biases and lacks contextual understanding. Introducing knowledge graphs, art theory ontologies, and emotion annotation frameworks can guide generation toward semantically meaningful and aesthetically rich outcomes. For instance, incorporating principles from color theory or composition rules can help models produce work aligned with traditional artistic values. Furthermore, curated datasets annotated with artistic intent, symbolism, or historical context can significantly enhance the interpretability and relevance of generated content.

5.3 Artistic creation through the collaboration of general-purpose and specialized models

A promising research direction is the cooperative use of general-purpose foundation models (e.g., CLIP, DALL·E, and Imagen) with domain-specific fine-tuned models trained on curated artistic corpora. General models provide linguistic understanding and broad visual priors, while specialized models capture stylistic fidelity and creative nuance. This cooperation can be orchestrated through ensemble frameworks, model cascading, or knowledge distillation pipelines, allowing task-specific precision without sacrificing generalization. Such cooperative systems can handle open-ended user prompts while delivering results with artistic coherence and cultural sensitivity.

5.4 AI-aided co-creation agent systems for artistic generation

The vision of a true AI-powered art assistant lies in real-time, interactive co-creation. Current models mostly generate static outputs from one-shot prompts, but future systems should incorporate dialogue understanding, sketch refinement, and iterative feedback handling. By embedding generative agents into digital workspaces (e.g., Photoshop plugins and VR environments), human artists can guide the model in a natural and collaborative workflow. Moreover, modeling user intent through continuous interaction—via sketch, text, voice, or gaze—will enable systems to adaptively learn user preferences and style, lowering barriers for non-experts to participate in creative expression.

Looking ahead, three complementary directions will further shape the landscape of human–AI co-creation in generative art: spatial intelligence, gesture interaction, and on-device real-time creation.

Spatial intelligence enables generative models to transcend 2D outputs by incorporating awareness of environments, spatial relations, and multidimensional contexts. This capability not only facilitates expansion into immersive 3D scenarios but also supports seamless integration of art with physical spaces. For example, in VR and AR environments, models can dynamically generate site-specific installations or interactive artworks that adapt to spatial layouts, lighting, and perspective.

Gesture interaction provides a natural and intuitive interface for generative art. Through gesture recognition and motion tracking, users can directly sculpt artistic elements on an "aerial canvas" or map motion features (e.g., speed, intensity, and posture) to emotional signals that guide the generative process. This interaction paradigm boosts engagement and shifts creative practice from passive control to a performative, embodied experience.

On-device real-time creation emphasizes the deployment of generative models on edge devices to ensure efficient

execution and immediate feedback. Advances in model compression, distillation, and quantization make it feasible for large-scale models to run on mobile platforms, AR/VR headsets, or interactive devices with low latency. Localized generation also preserves user privacy and enables personalized adaptation to individual artistic preferences, making the creative process both secure and inclusive.

These advances signal a paradigm shift in which generative art moves beyond static visual presentation toward spatialized, interactive, and deeply personalized immersive experiences, powered by collaborative large–small model ecosystems, where cloud-scale generative foundation models provide broad creative intelligence and edge-deployed lightweight models enable real-time adaptation, privacy-preserving personalization, and domain specificity. This trajectory highlights the emergence of human–computer cooperative systems and generative agents that not only augment expert creativity but also democratize artistic expression for broader audiences, while demanding robust model compression, secure collaboration mechanisms, and trustworthy generative control to ensure scalable and safe deployment.

5.5 Towards intention-driven interaction paradigms

To build genuinely human-centered collaborative agent systems, future research must move beyond traditional interaction logic and redesign the interaction paradigm itself. This shift is essential for addressing core challenges across the express–generate–edit creative loop. Key directions include:

1. From "prompt engineering" to "intention engineering." New interactive tools should capture implicit knowledge, emotional cues, and high-level semantic intent beyond textual prompts. Examples include gesture-based sculpting interfaces, real-time emotional feedback mechanisms, and systems capable of injecting abstract emotions and cultural background knowledge. These tools aim to achieve precise translation of abstract artistic intentions.

2. Nonlinear creative workflows. Future systems should break from linear editing pipelines by constructing an intelligent creative version tree that supports comparison, merging, and recombination of branching creative ideas. Such nonlinear workflows enable exploratory jumps, better reflect the intertwined logic of artistic thinking, and support efficient management of the creative process.

3. A shift from algorithm-centric to human-centric systems. The creation process will transition from humans adapting to algorithmic constraints to systems that evolve to match human preferences and expressive habits. This requires moving from offline learning with passive fine-tuning to online learning driven by human feedback, enabling continual self-improvement. Multi-user preference data will form an evolving map of human creativity and expression over time, capturing broader cultural and societal changes.

5.6 Generalizable, interpretable, and decentralized detection of AI-generated art

The future development of synthetic content detection will transcend traditional paradigms of binary classification and centralized architectures, moving towards a new stage characterized by collaboration, scalability, and decentralization.

First, a data–knowledge dual-driven framework should be established by integrating large-scale annotated datasets with forensic knowledge graphs, thereby enhancing semantic understanding, logical reasoning, and interpretability of detection models, and improving its applicability in high-trust scenarios. Second, cross-modal detection should be advanced to overcome the limitations of single-modality approaches, enabling models to jointly analyze text, images, audio, and video data to address complex multimodal forgeries generated by large models. Third, inspired by the "scaling law" in large models, research should explore the construction of general-purpose detection models, leveraging scale effects to achieve stronger generalization and robustness. Fourth, detection agents with continual learning and dynamic adaptation capabilities should be developed, allowing them to maintain agile responsiveness and evolutionary capacity in the face of emerging threats. Fifth, decentralized detection mechanisms based on blockchain technology should be introduced, where distributed storage and consensus verification can ensure the traceability and verifiability of synthetic content, thereby enhancing the transparency and tamper-resistance of the detection ecosystem.

5.7 Potential applications of interactive art creating systems

The implications of human–AI collaborative visual art extend beyond traditional fine art and encompass multiple domains:

1. Education. The co-creation system can serve as an interactive tutor for art students, guiding exploration of techniques and styles through structured exercises. It can also generate and visualize content in the style of a specific historical period or artist, deepening understanding of aesthetic principles.

2. Design. Human–AI co-creation supports rapid prototyping in fashion, interior, and product design. It can generate unique textile patterns, novel layouts, or color palettes based on a specific theme or mood, significantly accelerating the ideation and concept development phase.

3. Media and culture. AI for media (AI4Media) is definitely an emerging area with the rise of interactive AI. The co-creation system enables efficient, scalable production of customized visual content across social media, advertising, and e-commerce. Generated visuals can also be tailored to demographic profiles, cultural contexts, and lifestyle preferences.

4. Games and animation. The human–AI collaborative system can facilitate concept art creation for games, films, and animation. This includes procedural generation of dynamic backgrounds, environmental art, and character variations, enriching interactive and immersive digital experiences.

5. Scientific visualization. The system can also be used to generate complex diagrams, data visualizations, or microscopic imagery that can help researchers represent abstract concepts or experimental results more intuitively.

6. Therapy. AI-assisted creative agents can support art and mental health therapy by enabling emotional expression.

They provide safe, non-judgmental spaces where patients can explore and express emotions, engaging in creative activities without requiring advanced artistic skills.

7. Accessibility. The human–AI collaborative creation tools expand opportunities for individuals with physical or cognitive limitations to participate in art-making. Text-to-image interfaces and gesture- or voice-controlled generation allow users to create complex visual works without relying on conventional artistic techniques.

6 Conclusions

This work focuses on generative AI-driven interactive systems for visual and artistic image creation, examining their core tasks, key challenges, and future research directions. We find that the field is undergoing a paradigm shift from technical execution to human–AI iterative co-creation, with central tasks spanning inspiration support, progressive content generation, and aesthetic refinement. We identify three fundamental challenges that shape the development of interactive systems: controllability, integrity, and creativity. Looking forward, we argue that next-generation systems must strike a balance between controllability and creativity, enabling dynamic collaboration while preserving artistic value and responsible system behavior. This task-oriented, interdisciplinary perspective offers a roadmap for advancing research in the area and for building the next generation of AI-driven artistic creation systems.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. U24A20326 and 62502442) and the Zhejiang Provincial Natural Science Foundation of China (No. LQN26F020009).

Author contributions

Huanghuang DENG, Yongwei WANG, Yu ZHAO, Fei WU, and Yunhe PAN initiated the idea and designed the review framework. Huanghuang DENG, Aoran SUN, Zhan MENG, and Yongwei WANG collected the literature for review and drafted the paper. Luyang CHEN, Dunhan LI, Zhiyu LI, and Yue HAN helped organize the paper. Zejian LI, Zemin LIU, Ning ZOU, Kun KUANG, Zhou ZHAO, Kejun ZHANG, and Lingyun SUN reviewed the paper and checked the writing details. All the authors revised and finalized the paper.

Conflict of interest

Yunhe PAN is an editor-in-chief of *ENGINEERING Information Technology & Electronic Engineering* (*EITEE*); Fei WU is an executive associate editor-in-chief of *EITEE*. They were not involved with the peer review process of this paper. All the authors declare that they have no conflict of interest.

Declaration on the use of generative AI tools

During the preparation of this work, the authors used DeepSeek for the purpose of translation and drafting the prompts for generating illustration examples in Fig. 5. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Achlioptas P, Ovsjanikov M, Haydarov K, et al., 2021. ArtEmis: affective language for visual art. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.11564-11574. https://doi.org/10.1109/CVPR46437.2021.01140

Adamkiewicz K, Woźniak PW, Dominiak J, et al., 2025. PromptMap: an alternative interaction style for AI-based image generation. Proc 30th Int Conf on Intelligent User Interfaces, p.1162-1176. https://doi.org/10.1145/3708359.3712150

Agudo U, Arrese M, Liberal KG, et al., 2022. Assessing emotion and sensitivity of AI artwork. *Front Psychol*, 13:879088. https://doi.org/10.3389/fpsyg.2022.879088

AlDahoul N, Rahwan T, Zaki Y, 2025. AI-generated faces influence gender stereotypes and racial homogenization. *Sci Rep*, 15(1):14449. https://doi.org/10.1038/s41598-025-99623-3

Amabile TM, 1982. Social psychology of creativity: a consensual assessment technique. *J Pers Soc Psychol*, 43(5):997-1013. https://doi.org/10.1037/0022-3514.43.5.997

Ambartsumian Y, Cannon MT, 2025. Why the obsession with human creativity? A comparative analysis on copyright registration of AI-generated works. *Harvard Int Law J*.

Avakri A, Avakri I, 2025. Analysis of the creative mode of intelligent image generation art. Proc Int Conf on Artificial Intelligence, Digital Media Technology and Interaction Design, p.246-250. https://doi.org/10.1145/3726010.3726048

Avrahami O, Lischinski D, Fried O, 2022. Blended diffusion for text-driven editing of natural images. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.18187-18197. https://doi.org/10.1109/CVPR52688.2022.01767

Bahng H, Yoo S, Cho W, et al., 2018. Coloring with words: guiding image colorization through text-based palette generation. Proc 15th European Conf on Computer Vision, p.443-459. https://doi.org/10.1007/978-3-030-01258-8_27

Bau D, Zhu JY, Strobelt H, et al., 2019. GAN dissection: visualizing and understanding generative adversarial networks. Proc 7th Int Conf on Learning Representations.

Bayramli Z, Suleymanzade A, An NM, et al., 2025. Diffusion models through a global lens: are they culturally inclusive? Proc 63rd Annual Meeting of the Association for Computational Linguistics, p.31137-31155. https://doi.org/10.18653/v1/2025.acl-long.1503

Blattmann A, Rombach R, Ling H, et al., 2023. Align your latents: high-resolution video synthesis with latent diffusion models. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.22563-22575. https://doi.org/10.1109/CVPR52729.2023.02161

Boden MA, 2004. The Creative Mind: Myths and Mechanisms. 2nd Edition. Routledge, London, UK. https://doi.org/10.4324/9780203508527

Brooks T, Holynski A, Efros AA, 2023. InstructPix2Pix: learning to follow image editing instructions. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.18392-18402. https://doi.org/10.1109/CVPR52729.2023.01764

Buolamwini J, Gebru T, 2018. Gender shades: intersectional accuracy disparities in commercial gender classification. Proc Conf on Fairness, Accountability and Transparency, p.77-91.

Cai A, Rick SR, Heyman JL, et al., 2023. DesignAID: using generative AI and semantic diversity for design inspiration. Proc ACM Collective Intelligence Conf, p.1-11. https://doi.org/10.1145/3582269.3615596

Canny J, 1986. A computational approach to edge detection. *IEEE Trans Patt Anal Mach Intell*, PAMI-8(6):679-698. https://doi.org/10.1109/TPAMI.1986.4767851

Cao HQ, Tan C, Gao ZY, et al., 2024. A survey on generative diffusion models. *IEEE Trans Knowl Data Eng*, 36(7):2814-2830. https://doi.org/10.1109/TKDE.2024.3361474

Carlini N, Hayes J, Nasr M, et al., 2023. Extracting training data from diffusion models. Proc 32nd USENIX Conf on Security Symp, Article 294.

Chamberlain R, Mullin C, Scheerlinck B, et al., 2018. Putting the art in artificial: aesthetic responses to computer-generated art. *Psychol Aesthet Crea Arts*, 12(2):177-192. https://doi.org/10.1037/aca0000136

Chang HW, Zhang H, Jiang L, et al., 2022. MaskGIT: masked generative image transformer. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.11305-11315. https://doi.org/10.1109/CVPR52688.2022.01103

Chen C, Liu DC, Xu C, 2024. Towards memorization-free diffusion models. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.8425-8434. https://doi.org/10.1109/CVPR52733.2024.00805

Chen LC, Zhu YK, Papandreou G, et al., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proc 15th European Conf on Computer Vision, p.833-851. https://doi.org/10.1007/978-3-030-01234-2_49

Cheng JH, Yin BQ, Cai KX, et al., 2024. TheaterGen: character management with LLM for consistent multi-turn image generation. https://arxiv.org/abs/2404.18919

Chung JJY, Adar E, 2023. PromptPaint: steering text-to-image generation through paint medium-like interactions. Proc 36th Annual ACM Symp on User Interface Software and Technology, Article 6. https://doi.org/10.1145/3586183.3606777

Coalition for Content Provenance and Authenticity, 2024. C2PA Specifications. https://spec.c2pa.org/ [Accessed on Aug. 4, 2026].

Coalition for Content Provenance and Authenticity, 2026a. Content Credentials: C2PA Technical Specification. https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html [Accessed on Aug. 4, 2026].

Coalition for Content Provenance and Authenticity, 2026b. Guiding Principles for C2PA Designs and Specifications. https://c2pa.org/principles/ [Accessed on Aug. 4, 2026].

Costa V, Alonso-Moral JM, Falomir Z, et al., 2023. An art painting style explainable classifier grounded on logical and commonsense reasoning. *Soft Comput*. https://doi.org/10.1007/s00500-023-08258-x

Cui P, Lü Z, Zhang YS, 2025. Construction and design application of plant pattern cultural genealogy in Changsha kiln. *Furnit Interior Des*, 32(10):64-71 (in Chinese). https://doi.org/10.16771/j.cn43-1247/ts.2025.10.010

Cyberspace Administration of China, 2025. Measures for Labelling Artificial-Intelligence-Generated and Synthetic Content (in Chinese). https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm [Accessed on Aug. 4, 2026].

Dalal N, Triggs B, 2005. Histograms of oriented gradients for human detection. Proc IEEE Computer Society Conf on Computer Vision and Pattern Recognition, p.886-893. https://doi.org/10.1109/CVPR.2005.177

Darda KM, Chatterjee A, 2023. The impact of contextual information on aesthetic engagement of artworks. *Sci Rep*, 13(1):4273. https://doi.org/10.1038/s41598-023-30768-9

Demmer TR, Kühnapfel C, Fingerhut J, 2023. Does an emotional connection to art really require a human artist? Emotion and intentionality responses to AI- versus human-created art and impact on aesthetic experience. *Comput Hum Behav*, 148:107875. https://doi.org/10.1016/j.chb.2023.107875

Desai DR, Riedl M, 2024. Between copyright and computer science: the law and ethics of generative AI. https://arxiv.org/abs/2403.14653

Dhariwal P, Nichol A, 2021. Diffusion models beat GANs on image synthesis. Proc 35th Int Conf on Neural Information Processing Systems, Article 672.

D'Incà M, Peruzzo E, Mancini M, et al., 2024. OpenBias: open-set bias detection in text-to-image generative models. https://arxiv.org/abs/2404.07990

Ding X, Wang YW, Zhang K, et al., 2026. CCDM: continuous conditional diffusion models for image generation. *IEEE Trans Multimedia*, 28:4360-4372. https://doi.org/10.1109/TMM.2026.3660128

Esser P, Rombach R, Ommer B, 2021. Taming transformers for high-resolution image synthesis. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.12868-12878. https://doi.org/10.1109/CVPR46437.2021.01268

European Commission, 2025. Explanatory Notice and Template for the Public Summary of Training Content for General-Purpose AI Models. https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models [Accessed on Aug. 4, 2026].

European Union, 2024. Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:?32024R1689 [Accessed on Aug. 4, 2026].

Frich J, MacDonald Vermeulen L, Remy C, et al., 2019. Mapping the landscape of creativity support tools in HCI. Proc CHI Conf on Human Factors in Computing Systems, Article 389. https://doi.org/10.1145/3290605.3300619

Garcia N, Renoust B, Nakashima Y, 2019. Context-aware embeddings for automatic art analysis. Proc ACM Int Conf on Multimedia Retrieval, p.25-33. https://doi.org/10.1145/3323873.3325028

Goodfellow IJ, Pouget-Abadie J, Mirza M, et al., 2014. Generative adversarial nets. Proc 28th Int Conf on Neural Information Processing Systems, p.2672-2680.

Gu A, Dao T, 2024. Mamba: linear-time sequence modeling with selective state spaces. Proc 1st Conf on Language Modeling.

Han SJ, Fussell SR, 2025. Understanding user perceptions and the role of AI image generators in image creation workflows. Proc CHI Conf on Human Factors in Computing Systems, Article 547. https://doi.org/10.1145/3706598.3713227

Härkönen E, Hertzmann A, Lehtinen J, et al., 2020. GANSpace: discovering interpretable GAN controls. Proc 34th Int Conf on Neural Information Processing Systems, Article 825.

He BX, Guo Y, Zhu YB, et al., 2024. From signal to knowledge: the diagnostic value of raw data in the artificial intelligence prediction of human data for the first time. *Engineering*, 34:60-69. https://doi.org/10.1016/j.eng.2023.02.013

Hertz A, Mokady R, Tenenbaum J, et al., 2022. Prompt-to-prompt image editing with cross attention control. https://arxiv.org/abs/2208.01626

Hessel J, Holtzman A, Forbes M, et al., 2021. CLIPScore: a reference-free evaluation metric for image captioning. Proc Conf on Empirical Methods in Natural Language Processing, p.7514-7528. https://doi.org/10.18653/v1/2021.emnlp-main.595

Heusel M, Ramsauer H, Unterthiner T, et al., 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Proc 31st Int Conf on Neural Information Processing Systems, p.6629-6640.

Higgins I, Matthey L, Pal A, et al., 2017. β-VAE: learning basic visual concepts with a constrained variational framework. Proc 5th Int Conf on Learning Representations.

Ho J, Salimans T, 2022. Classifier-free diffusion guidance. https://arxiv.org/abs/2207.12598

Ho J, Jain A, Abbeel P, 2020. Denoising diffusion probabilistic models. Proc 34th Int Conf on Neural Information Processing Systems, Article 574.

Hu MH, Zheng JB, Liu DQ, et al., 2023. Cocktail: mixing multi-modality controls for text-conditional image generation. https://arxiv.org/abs/2306.00964

Huang LH, Chen D, Liu Y, et al., 2023. Composer: creative and controllable image synthesis with composable conditions. Proc 40th Int Conf on Machine Learning, p.13753-13773.

Huang YK, Ning Z, Cheng M, 2025. LACE: exploring turn-taking and parallel interaction modes in human-AI co-creation for iterative image generation. https://arxiv.org/abs/2504.14827

Isola P, Zhu JY, Zhou TH, et al., 2017. Image-to-image translation with conditional adversarial networks. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.5967-5976. https://doi.org/10.1109/CVPR.2017.632

Jeong S, Choi I, Yun Y, et al., 2025. Culture-TRIP: culturally-aware text-to-image generation with iterative prompt refinement. Proc Conf of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, p.9543-9573. https://doi.org/10.18653/v1/2025.naacl-long.483

Jing PG, Liu XY, Wang J, 2023. StyleEDL: style-guided high-order attention network for image emotion distribution learning. Proc 31st ACM Int Conf on Multimedia, p.853-861. https://doi.org/10.1145/3581783.3612040

Ju X, Zeng AL, Zhao CC, et al., 2023. HumanSD: a native skeleton-guided diffusion model for human image generation. Proc IEEE/CVF Int Conf on Computer Vision, p.15942-15952. https://doi.org/10.1109/ICCV51070.2023.01465

Ju X, Liu X, Wang XT, et al., 2024. BrushNet: a plug-and-play image inpainting model with decomposed dual-branch diffusion. Proc 18th European Conf on Computer Vision, p.150-168. https://doi.org/10.1007/978-3-031-72661-3_9

Jucker JL, Barrett JL, Wlodarski R, 2014. "I just don't get it": perceived artists' intentions affect art evaluations. *Empir Stud Arts*, 32(2):149-182. https://doi.org/10.2190/EM.32.2.c

Karras T, Laine S, Aila T, 2019. A style-based generator architecture for generative adversarial networks. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.4396-4405. https://doi.org/10.1109/CVPR.2019.00453

Karras T, Laine S, Aittala M, et al., 2020. Analyzing and improving the image quality of StyleGAN. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.8107-8116. https://doi.org/10.1109/CVPR42600.2020.00813

Kirstain Y, Polyak A, Singer U, et al., 2023. Pick-a-Pic: an open dataset of user preferences for text-to-image generation. Proc 37th Int Conf on Neural Information Processing Systems, Article 1594.

Klassert T, Ulges A, Fu BY, 2026. BAFIS: dataset + framework to assess occupational bias and human preference in modern text-to-image models. Proc IEEE/CVF Winter Conf on Applications of Computer Vision, p.2168-2177. https://doi.org/10.1109/WACV61042.2026.00215

Lan XY, An JX, Guo YS, et al., 2025. Imagining the Far East: exploring perceived biases in AI-generated images of East Asian women. Proc CHI Conf on Human Factors in Computing Systems, Article 332. https://doi.org/10.1145/3706599.3719678

Leder H, Belke B, Oeberst A, et al., 2004. A model of aesthetic appreciation and aesthetic judgments. *Brit J Psychol*, 95(4):489-508. https://doi.org/10.1348/0007126042369811

Li YH, Liu HT, Wu QY, et al., 2023. GLIGEN: open-set grounded text-to-image generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.22511-22521. https://doi.org/10.1109/CVPR52729.2023.02156

Li YT, Gan Z, Shen YL, et al., 2019. StoryGAN: a sequential conditional GAN for story visualization. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.6322-6331. https://doi.org/10.1109/CVPR.2019.00649

Li ZJ, Zhang Y, Zhou SZ, et al., 2025. RealtimeGen: an intervenable AI image generation system for commercial digital art asset creators. *Int J Hum–Comput Interact*, 41(11):6613-6636. https://doi.org/10.1080/10447318.2024.2382508

Liu QQ, Yan YP, Jin YC, et al., 2024. Secure federated evolutionary optimization—a survey. *Engineering*, 34:23-42. https://doi.org/10.1016/j.eng.2023.10.006

Long J, Shelhamer E, Darrell T, 2015. Fully convolutional networks for semantic segmentation. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.3431-3440. https://doi.org/10.1109/CVPR.2015.7298965

Lv KT, Chen HB, Yuan YJ, et al., 2025. How to inject knowledge efficiently? Knowledge infusion scaling law for pre-training large language models. Proc Conf on Empirical Methods in Natural Language Processing, p.26193-26208. https://doi.org/10.18653/v1/2025.emnlp-main.1331

Ma Z, Li QM, Zhang XH, et al., 2025. An inversion-based measure of memorization for diffusion models. Proc IEEE/CVF Int Conf on Computer Vision, p.16959-16969. https://doi.org/10.1109/ICCV51701.2025.01575

Machajdik J, Hanbury A, 2010. Affective image classification using features inspired by psychology and art theory. Proc 18th ACM Int Conf on Multimedia, p.83-92. https://doi.org/10.1145/1873951.1873965

Maher ML, Grace K, Karimi P, et al., 2018. Evaluating creativity in computational co-creative systems. Proc 9th Int Conf on Computational Creativity, p.104-111.

Man F, Chen X, Wang H, et al., 2025. K-EVER2: knowledge-enhanced visual emotion reasoning and retrieval. Proc 33rd ACM Int Conf on Multimedia.

Matsumoto H, Nakamura A, Hamaguchi K, et al., 2025. From sketch to refinement: interactive drawing support with generative AI. Proc 14th Global Conf on Consumer Electronics, p.32-36. https://doi.org/10.1109/GCCE65946.2025.11274974

Millet K, Buehler F, Du GZ, et al., 2023. Defending humankind: anthropocentric bias in the appreciation of AI art. *Comput Hum Behav*, 143:107707. https://doi.org/10.1016/j.chb.2023.107707

Mokady R, Hertz A, Aberman K, et al., 2023. Null-text inversion for editing real images using guided diffusion models. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.6038-6047. https://doi.org/10.1109/CVPR52729.2023.00585

Mou C, Wang XT, Xie LB, et al., 2024. T2I-Adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models. Proc 38th AAAI Conf on Artificial Intelligence, p.4296-4304. https://doi.org/10.1609/aaai.v38i5.28226

Nayak S, Bhatia M, Zhang XF, et al., 2025. CulturalFrames: assessing cultural expectation alignment in text-to-image models and evaluation metrics. Proc Association for Computational Linguistics, p.20918-20953.

Nazeri K, Ng E, Joseph T, et al., 2019. EdgeConnect: generative image inpainting with adversarial edge learning. https://arxiv.org/abs/1901.00212

O'Donovan P, Agarwala A, Hertzmann A, 2011. Color compatibility from large datasets. *ACM Trans Graph*, 30(4):43. https://doi.org/10.1145/2010324.1964958

Ojala T, Pietikainen M, Mäenpää T, 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans Patt Anal Mach Intell*, 24(7):971-987. https://doi.org/10.1109/TPAMI.2002.1017623

Oppenlaender J, Johnston H, Silvennoinen JM, et al., 2025. Artworks reimagined: exploring human-AI co-creation through body prompting. *Proc ACM Hum-Comput Interact*, 9(4):EICS012. https://doi.org/10.1145/3734189

O'Quin K, Besemer SP, 1989. The development, reliability, and validity of the revised creative product semantic scale. *Creativ Res J*, 2(4):267-278. https://doi.org/10.1080/10400418909534323

Pan XG, Tewari A, Leimkühler T, et al., 2023. Drag your GAN: interactive point-based manipulation on the generative image manifold. Proc ACM SIGGRAPH Conf, Article 78. https://doi.org/10.1145/3588432.3591500

Pan YH, 2019. On visual knowledge. *Front Inform Technol Electron Eng*, 20(8):1021-1025. https://doi.org/10.1631/FITEE.1910001

Pan YH, 2020. Multiple knowledge representation of artificial intelligence. *Engineering*, 6(3):216-217. https://doi.org/10.1016/j.eng.2019.12.011

Park T, Liu MY, Wang TC, et al., 2019. Semantic image synthesis with spatially-adaptive normalization. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.2332-2341. https://doi.org/10.1109/CVPR.2019.00244

Parmar G, Singh KK, Zhang R, 2023. Zero-shot image-to-image translation. Proc ACM SIGGRAPH Conf, Article 11. https://doi.org/10.1145/3588432.3591513

Peng YJ, Han JH, Zhang ZL, et al., 2024. The Tong test: evaluating artificial general intelligence through dynamic embodied physical and social interactions. *Engineering*, 34:12-22. https://doi.org/10.1016/j.eng.2023.07.006

Qadri R, Mirowski P, Gabriellan A, et al., 2024. Dialogue with the machine and dialogue with the art world: evaluating generative AI for culturally-situated creativity. https://arxiv.org/abs/2412.14077

Radford A, Kim JW, Hallacy C, et al., 2021. Learning transferable visual models from natural language supervision. Proc 38th Int Conf on Machine Learning, p.8748-8763.

Rafner JF, Zana B, Hansen IB, et al., 2025. Agency in human-AI collaboration for image generation and creative writing: preliminary insights from think-aloud protocols. *Creativ Res J*. https://doi.org/10.1080/10400419.2025.2587803

Ramesh A, Pavlov M, Goh G, et al., 2021. Zero-shot text-to-image generation. Proc 38th Int Conf on Machine Learning, p.8821-8831.

Ramesh A, Dhariwal P, Nichol A, et al., 2022. Hierarchical text-conditional image generation with CLIP latents. https://arxiv.org/abs/2204.06125

Ranftl R, Bochkovskiy A, Koltun V, 2021. Vision transformers for dense prediction. Proc IEEE/CVF Int Conf on Computer Vision, p.12159-12168. https://doi.org/10.1109/ICCV48922.2021.01196

Rezwana J, Maher ML, 2023. Designing creative AI partners with COFI: a framework for modeling interaction in human-AI co-creative systems. *ACM Trans Computer-Human Interact*, 30(5):67. https://doi.org/10.1145/3519026

Rhodes M, 1961. An analysis of creativity. *Phi Delta Kappan*, 42(7):305-310.

Rombach R, Blattmann A, Lorenz D, et al., 2022. High-resolution image synthesis with latent diffusion models. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.10674-10685. https://doi.org/10.1109/CVPR52688.2022.01042

Royal Courts of Justice, 2025. In the High Court of Justice Business and Property Courts of England and Wales Intellectual Property List (ChD). https://www.judiciary.uk/wp-content/uploads/2025/11/Getty-Images-v-Stability-AI.pdf [Accessed on Aug. 4, 2026].

Ruiz N, Li YZ, Jampani V, et al., 2023. DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.22500-22510. https://doi.org/10.1109/CVPR52729.2023.02155

Runco MA, Jaeger GJ, 2012. The standard definition of creativity. *Creativ Res J*, 24(1):92-96. https://doi.org/10.1080/10400419.2012.650092

Saharia C, Chan W, Chang HW, et al., 2022a. Palette: Image-to-image diffusion models. Proc ACM SIGGRAPH Conf, Article 15. https://doi.org/10.1145/3528233.3530757

Saharia C, Chan W, Saxena S, et al., 2022b. Photorealistic text-to-image diffusion models with deep language understanding. Proc 36th Int Conf on Neural Information Processing Systems, Article 2643.

Satkunarajan J, Koch S, Kurzhals K, 2025. The joy of co-painting: creative human-AI collaboration for traceable image-generation workflows. Proc 18th Pacific Visualization Conf, p.318-328. https://doi.org/10.1109/PacificVis64226.2025.00038

Sengar SS, Hasan AB, Kumar S, et al., 2025. Generative artificial intelligence: a systematic review and applications. *Multimed Tools Appl*, 84(21):23661-23700. https://doi.org/10.1007/s11042-024-20016-1

Seshadri P, Singh S, Elazar Y, 2024. The bias amplification paradox in text-to-image generation. Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p.6367-6384. https://doi.org/10.18653/v1/2024.naacl-long.353

Shahriar S, 2022. Gan computers generate arts? A survey on visual arts, music, and literary text generation using generative adversarial network. *Displays*, 73:102237. https://doi.org/10.1016/j.displa.2022.102237

Shan S, Cryan J, Wenger E, et al., 2023. Glaze: protecting artists from style mimicry by text-to-image models. https://arxiv.org/abs/2302.04222

Shen YJ, Gu JJ, Tang XO, et al., 2020. Interpreting the latent space of GANs for semantic face editing. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.9240-9249. https://doi.org/10.1109/CVPR42600.2020.00926

Shi HY, Wang L, Zheng NN, et al., 2022. Loss functions for pose guided person image generation. *Patt Recognit*, 122:108351. https://doi.org/10.1016/j.patcog.2021.108351

Shi WY, Upadhyay S, Quek G, et al., 2025. TalkSketch: multimodal generative AI for real-time sketch ideation with speech. https://arxiv.org/abs/2511.05817

Shneiderman B, 2022. Human-Centered AI. Oxford University Press, Oxford, UK. https://doi.org/10.1093/oso/9780192845290.001.0001

Song Y, Sohl-Dickstein J, Kingma DP, 2021. Score-based generative modeling through stochastic differential equations. Proc 9th Int Conf on Learning Representations.

Song ZR, Yan B, Liu YH, et al., 2025. Injecting domain-specific knowledge into large language models: a comprehensive survey. Proc Association for Computational Linguistics, p.25297-25311.

Stein G, Cresswell JC, Hosseinzadeh R, et al., 2023. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. Proc 37th Int Conf on Neural Information Processing Systems, Article 165.

Stober S, Dornis TW, 2025. Generative AI training and copyright law. https://arxiv.org/abs/2502.15858

Tan WR, Chan CS, Aguirre HE, et al., 2019. Improved ArtGAN for conditional synthesis of natural image and artwork. *IEEE Trans Image Process*, 28(1):394-409. https://doi.org/10.1109/TIP.2018.2866698

Tan ZT, Chai ML, Chen DD, et al., 2021. Diverse semantic image synthesis via probability distribution modeling. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.7958-7967. https://doi.org/10.1109/CVPR46437.2021.00787

Tao M, Tang H, Wu F, et al., 2020. DF-GAN: a simple and effective baseline for text-to-image synthesis. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.16494-16504. https://doi.org/10.1109/CVPR52688.2022.01602

Tholander J, Jonsson M, 2023. Design ideation with AI - sketching, thinking and talking with generative machine learning models. Proc ACM Designing Interactive Systems Conf, p.1930-1940. https://doi.org/10.1145/3563657.3596014

UNESCO, 2003. Text of the Convention for the Safeguarding of the Intangible Cultural Heritage. https://ich.unesco.org/en/convention [Accessed on Aug. 4, 2026].

UNESCO, 2021. Recommendation on the Ethics of Artificial Intelligence. https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence [Accessed on Aug. 4, 2026].

U.S. Copyright Office, 2025a. Copyright and Artificial Intelligence, Part 2: Copyrightability. U.S. Copyright Office, Washington, DC, USA.

U.S. Copyright Office, 2025b. Copyright and Artificial Intelligence, Part 3: Generative AI Training. U.S. Copyright Office, Washington, DC, USA.

Ungless E, Ross B, Lauscher A, 2023. Stereotypes and smut: the (mis)representation of non-cisgender identities by text-to-image models. Proc Association for Computational Linguistics, p.7919-7942. https://doi.org/10.18653/v1/2023.findings-acl.502

United States Court of Appeals for the District of Columbia Circuit, 2025. Thaler v. Perlmutter (D.C. Cir. 2025). https://law.justia.com/cases/federal/appellate-courts/cadc/23-5233/23-5233-2025-03-18.html [Accessed on Aug. 4, 2026].

United States District Court for the Northern District of California, 2026. 23-201 - Andersen et al. v. Stability AI Ltd. et al. https://www.govinfo.gov/app/details/USCOURTS-cand-3_23-cv-00201 [Accessed on Aug. 4, 2026].

van den Oord A, Vinyals O, Kavukcuoglu K, 2017. Neural discrete representation learning. Proc 31st Int Conf on Neural Information Processing Systems, p.6309-6318.

Vice J, Akhtar N, Sigal L, et al., 2026. On the fairness, diversity and reliability of text-to-image generative models. *Artif Intell Rev*, 59:57. https://doi.org/10.1007/s10462-025-11424-2

von Gioi RG, Jakubowicz J, Morel JM, et al., 2010. LSD: a fast line segment detector with a false detection control. *IEEE Trans Patt Anal Mach Intell*, 32(4):722-732.

Wadinambiarachchi S, Kelly RM, Pareek S, et al., 2024. The effects of generative AI on design fixation and divergent thinking. Proc CHI Conf on Human Factors in Computing Systems, Article 380. https://doi.org/10.1145/3613904.3642919

Wan YX, Wu D, Wang HR, et al., 2024a. The factuality tax of diversity-intervened text-to-image generation: benchmark and fact-augmented intervention. Proc Conf on Empirical Methods in Natural Language Processing, p.9082-9100.

Wan YX, Subramonian A, Ovalle A, et al., 2024b. Survey of bias in text-to-image generation: definition, evaluation, and mitigation. https://arxiv.org/abs/2404.01030

Wang BY, Chen QF, Wang ZY, 2025. Diffusion-based visual art creation: a survey and new perspectives. *ACM Comput Surv*, 57(10):268. https://doi.org/10.1145/3728459

Wang HF, Xing P, Huang RY, et al., 2024. InstantStyle-Plus: style transfer with content-preserving in text-to-image generation. https://arxiv.org/abs/2407.00788

Wang P, Zhang X, Wei LY, et al., 2026. Human-AI co-ideation via combinational generative model. *J Eng Des*, 37(2):458-494. https://doi.org/10.1080/09544828.2025.2504309

Wang RC, Zhang JL, Xie QS, et al., 2024. PainterNet: adaptive image inpainting with actual-token attention and diverse mask control. https://arxiv.org/abs/2412.01223

Wang YW, Shen T, Zhang SY, et al., 2024. Advances in edge-cloud collaboration and evolution for large-small models. *J Image Graph*, 29(6):1510-1534 (in Chinese). https://doi.org/10.11834/jig.240011

Wang ZY, Li AX, Li ZG, et al., 2024. GenArtist: multimodal LLM as an agent for unified image generation and editing. Proc 38th Int Conf on Neural Information Processing Systems.

Wortsman M, Ilharco G, Gadre SY, et al., 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. Proc 39th Int Conf on Machine Learning, p.23965-23998.

Wu F, Shen T, Bäck T, et al., 2025. Knowledge-empowered, collaborative, and co-evolving AI models: the post-LLM roadmap. *Engineering*, 44:87-100. https://doi.org/10.1016/j.eng.2024.12.008

Wu JZ, Ge YX, Wang XT, et al., 2023. Tune-A-Video: one-shot tuning of image diffusion models for text-to-video generation. Proc IEEE/CVF Int Conf on Computer Vision, p.7589-7599. https://doi.org/10.1109/ICCV51070.2023.00701

Xu JF, Liu ZY, Wang SL, et al., 2023. Foundations and applications of information systems dynamics. *Engineering*, 27:254-265. https://doi.org/10.1016/j.eng.2022.04.018

Xu T, Zhang PC, Huang QY, et al., 2018. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.1316-1324. https://doi.org/10.1109/CVPR.2018.00143

Xu ZY, Duan KS, Shen XL, et al., 2025. PixelHacker: image inpainting with structural and semantic consistency. https://arxiv.org/abs/2504.20438

Yang BX, Gu SY, Zhang B, et al., 2023. Paint by Example: exemplar-based image editing with diffusion models. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.18381-18391. https://doi.org/10.1109/CVPR52729.2023.01763

Yang JY, Huang QR, Ding TT, 2023. EmoSet: a large-scale visual emotion dataset with rich attributes. Proc IEEE/CVF Int Conf on Computer Vision, p.20326-20337. https://doi.org/10.1109/ICCV51070.2023.01864

Yang JY, Feng JW, Huang H, 2024. EmoGen: emotional image content generation with text-to-image diffusion models. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.6358-6368. https://doi.org/10.1109/CVPR52733.2024.00608

Yang ZY, Wang JF, Li LJ, et al., 2024. Idea2Img: iterative self-refinement with GPT-4V(ision) for automatic image design and generation. https://arxiv.org/abs/2310.08541

Zeiler MD, Fergus R, 2014. Visualizing and understanding convolutional networks. Proc 13th European Conf on Computer Vision, p.818-833. https://doi.org/10.1007/978-3-319-10590-1_53

Zhang JS, Li K, Lai YK, et al., 2021. PISE: person image synthesis and editing with decoupled GAN. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.7978-7986. https://doi.org/10.1109/CVPR46437.2021.00789

Zhang L, Rao AY, Agrawala M, 2023. Adding conditional control to text-to-image diffusion models. Proc IEEE/CVF Int Conf on Computer Vision, p.3813-3824. https://doi.org/10.1109/ICCV51070.2023.00355

Zhang SF, Huang JF, Zhou Q, et al., 2024. Continuous-multiple image outpainting in one-step via positional query and a diffusion-based approach. Proc 12th Int Conf on Learning Representations.

Zhang TY, Xie XX, Du XS, et al., 2024. Sketch-guided scene image generation. https://arxiv.org/abs/2407.06469

Zhao SH, Chen DD, Chen YC, et al., 2023. Uni-ControlNet: all-in-one control to text-to-image diffusion models. Proc 37th Int Conf on Neural Information Processing Systems, Article 491. https://doi.org/10.52202/075280-0491

Zhao SY, Cui J, Sheng YL, et al., 2021. Large scale image completion via co-modulated generative adversarial networks. Proc 9th Int Conf on Learning Representations.

Zheng HZ, 2025. Translation methods of AIGC in graphic creativity of intangible cultural heritage and art design strategies. *East China Pulp Pap Ind*, 55(10):82-84 (in Chinese).

Zheng QH, Liu H, Zhang XQ, et al., 2025. Machine memory intelligence: inspired by human memory mechanisms. *Engineering*, 55:24-35. https://doi.org/10.1016/j.eng.2025.01.012

Zhu JY, Park T, Isola P, et al., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proc IEEE Int Conf on Computer Vision, p.2242-2251. https://doi.org/10.1109/ICCV.2017.244

Zhu RJ, Zhao QH, Li GQ, et al., 2024. SpikeGPT: generative pre-trained language model with spiking neural networks. https://arxiv.org/abs/2302.13939