EVOLUTION OF GENERATIVE ART SYSTEMS: METHODS, MILESTONES AND OPEN PROBLEMS

Viktor Matanski

Abstract. This article explores how generative art systems have evolved, from the rule-based and algorithmic experiments of the 1960s to today's neural and multimodal models. Looking across visual art, music and literature, it traces three broad stages in this development. Unlike prior surveys that primarily organise the field chronologically or by model family, this article frames the evolution of generative art systems as a representational trade-off: increasing generative fluency is accompanied by recurring losses of interpretability, granular controllability, authorship attribution and cross-modal fidelity. The first stage includes early computational approaches such as procedural generation, L-systems, cellular automata and stochastic methods. The second stage marks the shift toward data-driven techniques, including evolutionary algorithms, generative adversarial networks, variational autoencoders and diffusion models. The third and most recent stage involves the rise of cross-modal and foundation-model-based systems that can generate content across multiple artistic domains. Along the way, the article discusses key milestones, from Harold Cohen's AARON to DALL·E, MusicLM and GPT-4, within their wider methodological and cultural settings. It concludes by highlighting several challenges that remain unresolved, including how to evaluate aesthetic quality reliably, how to think about authorship and copyright, how to balance control with creative freedom and how to build unified frameworks for cross-modal artistic translation.

Key Words: *Generative Art Systems, Computational Creativity, Algorithmic Art, Multimodal Generation, Cross-Modal Artistic Translation, Diffusion Models, Foundation Models, Aesthetic Evaluation, Authorship, Creative AI*

Introduction

The technical evolution of generative art over six decades – from the mainframe-and-plotter experiments of Nees, Nake and Noll in the 1960s to today's multimodal foundation models – is best read as a single sustained move:

from explicit symbolic representation (structure hand-authored as production rules, grammars, fitness functions, or stochastic processes) to latent statistical representation (structure implicit in the weights of a learned model and observable only when the model runs). The move has substantially lowered the technical barrier to creative production. It has also produced four recurring losses, each rarely framed as the unifying problem it is.

1. **Loss of interpretability.** An L-system practitioner in 1986 could read the production rules and predict, in broad strokes, the class of structures the system would generate. A latent-diffusion practitioner in 2026 cannot. The artistic content lives in tens of billions of parameters that resist symbolic explanation.
2. **Loss of granular controllability.** Prompt-conditioned systems offer easy global control through natural language and almost no fine-grained control out of the box. Many of the practically important controllability gains in diffusion-era systems have come from grafting symbolic structure back onto latent generation – ControlNet, T2I-Adapter, IP-Adapter, LoRA, DreamBooth [1, 2].
3. **Loss of authorship attribution.** When AARON painted, Cohen authored the rules. When Stable Diffusion paints, the contribution of any individual training example is statistically diffuse and legally contested. Andersen v. Stability AI [3], the post-Thaler [4] U.S. legal landscape and the EU AI Act [5] GPAI documentation requirements all respond to the same underlying fact: latent representation makes attribution technically difficult at production-model scale.
4. **Loss of cross-modal fidelity.** Cross-modal capability is widely claimed but rarely defined. Section 6 distinguishes five increasingly stringent senses, of which only the weakest – text-conditioned generation in a single output modality – is in widespread use; genuine artistic translation between modalities (a Rothko-to-music system; a poem-to-image system that preserves prosodic structure) remains absent from the production literature.

Where prior surveys treat the move to neural representations as straightforward progress, this article frames it as a trade. Sections 3-5 cover the symbolic, hybrid and latent-statistical eras. Section 6 presents a comparative table of eighteen systems with an explicit coding rubric (Appendix A) and develops the cross-modal distinction in detail. Section 7 revisits open problems through the thesis. Section 8 states four concrete research gaps.

The contribution of this review is therefore not only historical also analytical. Compared with existing surveys of generative art, computational creativity and diffusion-based generation, it contributes four specific elements. First, it interprets the field through a single representational transition, from

explicit symbolic rules to latent statistical representations. Second, it frames this transition as a trade-off rather than as linear technical progress, identifying four recurring losses: interpretability, granular controllability, authorship attribution and cross-modal fidelity. Third, it compares representative systems through a uniform rubric covering modality, representation, controllability, autonomy, evaluation method and cross-modal capability. Fourth, it distinguishes five different senses of "cross-modal" generation, thereby separating weak text-conditioned generation from the stronger and still unresolved problem of genuine cross-modal artistic translation.

Methodology

Sources were identified through Google Scholar, arXiv, the ACM Digital Library, IEEE Xplore and the proceedings of ICCC, NIME, NeurIPS, ICML, CVPR, ICLR, ISMIR and SIGGRAPH between November 2025 and March 2026, calibrated against three review-of-reviews exercises [6]. Coverage runs from 1960 to April 2026 across visual, music and literary modalities. Each system is characterised across six dimensions – modality, representation, controllability, autonomy, the evaluation method used in the original publication and cross-modal capability – with controllability and autonomy scored Low / Medium / High against the rubric in Appendix A and cross-modal capability decomposed into five senses in Section 6. The rubric was applied by a single rater; replication by independent raters is recommended.

Symbolic Generation (1960s – 1990s)

Early generative-art systems were symbolic in a strong sense: a human wrote the rules and reading those rules was usually enough to know what the system would produce. Nees and Nake exhibited the first algorithmic drawings in 1965; Noll [8] ran the first published preference study comparing human and computer-generated visual output. Lindenmayer's L-systems [9], canonised in Prusinkiewicz and Lindenmayer [10], provide the period's clearest case of complex output from simple local rules. Mandelbrot [11], Conway's Game of Life and Wolfram [12] extended the same idea: a small symbolic substrate, fully inspectable, can yield strikingly intricate visual output.

Harold Cohen's AARON [13] (begun in 1973) is the most analytically important system of this era. AARON ran on hand-coded production rules but did something today's literature often treats as a recent property: it separated controllability from autonomy. Cohen wrote the rules – he could rewrite them – but he could not predict what AARON would draw on a given run. The system made its own composition decisions. Table 1 codes AARON's autonomy as High and its controllability as Medium for that reason.

In music, Xenakis pioneered stochastic and mathematical composition [14]. His UPIC system (1977) is best read not as a stochastic system but as a graphical-to-sonic interactive environment: the composer drew waveforms on a tablet and the system synthesised the sound – an early cross-modal interface. Cope's EMI (begun in 1981) used pattern extraction and Markov-chain recombination over a corpus of human compositions; it was the first widely-discussed system whose novelty came from recombining existing material rather than from hand-written rules – the first concrete step toward the latent-statistical paradigm.

For literature, the OULIPO movement (founded 1960 by Queneau and Le Lionnais) and Queneau's *Cent mille milliards de poèmes* (1961) wrote out an explicit combinatorial grammar; Weizenbaum's ELIZA (1966), Chamberlain and Etter's Racter (1984) and the planning-based story generators TALE-SPIN (1977) and MINSTREL (1994) extended the paradigm. Statistical language models – n-grams, then RNNs and LSTMs [15] – bridged the symbolic and latent eras.

Hybrid Era: Statistical Learning with Symbolic Scaffolding (mid-1990s to late 2010s)

The middle period is best described as statistical learning under symbolic scaffolding: systems learn distributions from data, but the architecture, the training objective and the conditioning signals are all written by hand. Sims's *Evolved Virtual Creatures* [16], strictly an artificial-life project, popularised evolutionary search over generative encodings. The interactive evolutionary-art tradition that followed, Machado's NEvAr and McCormack's ecosystems, is the clearest practical example of high-autonomy and low-individual-output controllability. The user cannot pick any specific image, but can steer the population toward an aesthetic region.

Two systems from 2015 mark the public emergence of deep-learning-based generative aesthetics. DeepDream [17] was a gradient-ascent feature-visualisation method on classifier activations rather than a generative system in the GAN sense, but culturally it produced the first recognisable "neural-network look". Neural Style Transfer [18] separated content and style representations inside VGG-19 and let any photograph be re-rendered in the style of a painting. It was the last major generative-art technique whose insides could be explained in a single paragraph to a non-specialist.

GANs [19] redefined image synthesis. The StyleGAN line (with successors through 2021) improved resolution and a limited form of latent-space control. The 2018 sale of *Edmond de Belamy* by the collective Obvious at Christie's and its underlying code, largely Robbie Barrat's public implementation, was the first widely-reported attribution dispute of the neural era. More interesting analytically, CAN [20] tweaks the GAN training objective to reward outputs the discriminator cannot easily place into a known artistic style. CAN is one of the

few neural systems that puts a specific computational-creativity criterion (Boden's [21] notion of transformational creativity) directly into its loss.

Variational autoencoders [22] gave the field a continuous, structured latent space. Somewhere a person could draw novel samples and smoothly interpolate between them. Latent traversal became a recognisable artistic practice. In music, MuseGAN applied GANs to multi-track piano roll and Jukebox [23] used a hierarchical VQ-VAE to generate minute-long audio with vocals – the most ambitious autoregressive audio model of its era. Briot, Hadjeres and Pachet [6] survey deep-learning music generation in detail.

Latent Statistical Generation: Diffusion and Foundation Models (2020s)

Today's generative systems learn distributions over modality-specific latent spaces big enough that the human-authored components: model architecture, loss function and training data, are dwarfed by the learned weights. The system's "rules" are no longer something a person can read.

Diffusion, given its canonical formulation by Ho, Jain and Abbeel [24], has been the dominant paradigm for high-fidelity image generation since around 2021. Rombach et al.'s Latent Diffusion Models [25] cut compute by roughly an order of magnitude by running diffusion in the latent space of a pretrained autoencoder and underpinned Stable Diffusion – a UNet (v1-v2) or transformer (v3) denoising network conditioned on text embeddings via cross-attention. Adjacent systems show the same conditioning idea through different backbones: DALL·E 2 (unCLIP, with a diffusion prior over CLIP image embeddings); Imagen (pixel-space diffusion on a frozen T5-XXL encoder); Parti (autoregression over discrete image tokens). The conditioning signal carries a substantial share of the artistic specification; the backbone does the rest.

Out of the box, a text-conditioned diffusion model gives you easy big-picture control and almost no fine-grained control. The practical fine-control gains of the past three years have come, in many practically important cases, from grafting explicit symbolic structure back onto the diffusion model: ControlNet [26] for structural conditioning on edge, depth and pose maps; T2I-Adapter as a lighter alternative; IP-Adapter for image-conditioned generation; DreamBooth [27] and Textual Inversion for personalising the model on a small set of subject images; LoRA [28] as the de facto adapter for distributing community-trained styles and characters. The pattern is consistent: the latent model gives scale and fluency, while controllability comes from adding back symbolic structure [1, 2]. The post-2023 production developments are, in practice, a hybrid system.

Text-to-music matured between 2022 and 2025. Riffusion (2022) reused Stable Diffusion to generate spectrograms; MusicLM [29] generates music from

text via MuLan-conditioned hierarchical sequences; MusicGen is a single-stage autoregressive transformer with interleaved codebooks; Stable Audio and the consumer-facing Suno and Udio (both 2024), brought text-to-song into mainstream awareness – and into RIAA litigation. Video generation followed a few years behind image: Runway Gen-2 (2023) and Gen-3 (2024), Sora [30] and VideoPoet. Large language models – GPT-3 [31], GPT-4 [32] and their contemporaries – became the principal generative tool for text. Worth noting on GPT-4 specifically: it accepts image and text inputs but produces text only; later OpenAI products extend this to true multimodal generation.

Comparative Analysis and the Cross-Modal Distinction

Table 1 codes eighteen representative systems across six dimensions. The rubric in Appendix A defines controllability as the operator's ability to shape output structure, and autonomy as the extent of the system's independent generative decisions. CLIP is omitted because it is an enabling embedding model, not a generative system; it is discussed instead as cross-modal retrieval in sense 2 below.

Table 1. Comparative characterisation of eighteen generative-art systems.

System / Year	Modality	Representation	Ctrl	Auto	Evaluation Method	Cross-Modal
AARON (Cohen, begun in 1973)	Visual	Production-rule expert system	Med	High	Qualitative; expert critique	No
UPIC (Xenakis, 1977)	Music	Graphical waveform → DSP synthesis	High	Low	Qualitative; composer assessment	Yes
EMI (Cope, begun in 1981)	Music	Pattern recombination; Markov chains	Low	Med	ABX listening tests	No
L-System Art (Prusinkiewicz, 1986)	Visual	Formal string-rewriting grammar	High	Low	Qualitative; botanical fidelity	No
Evolved Virtual Creatures (Sims, 1994)	Visual / 3D	Genetic programming over node graphs	Low	High	Interactive aesthetic selection	No
DeepDream (Mordvintsev et al., 2015)	Visual	Gradient ascent on CNN activations	Low	Med	Qualitative	No

Neural Style Transfer (Gatys et al., 2016)	Visual	CNN feature statistics (VGG-19)	High	Low	Perceptual user studies	No
CAN (Elgammal et al., 2017)	Visual	GAN with style-deviation reward	Low	High	Human-preference; Turing-style	No
MuseGAN (Dong et al., 2018)	Music	GAN on multi-track piano roll	Med	Med	Pitch / rhythm metrics; user study	No
StyleGAN (Karras et al., 2019)	Visual	Adversarial latent vector; style modulation	Med	Med	FID, IS, PPL; human preference	No
Jukebox (Dhariwal et al., 2020)	Music	Hierarchical VQ-VAE; raw audio tokens	Low	High	Audio-quality ratings	No
GPT-3 (Brown et al., 2020)	Text	Autoregressive transformer; token logits	Med	High	Few-shot benchmarks; human eval	No
Stable Diffusion (Rombach et al., 2022)	Visual	Latent diffusion + text cross-attention	Med	High	FID; HPSv2 human preference	Partial
SD+ ControlNet (Zhang et al., 2023)	Visual	LDM + structural-conditioning adapter	High	High	FID; human preference	Partial
MusicLM (Agostinelli et al., 2023)	Music	Hierarchical LM on MuLan embeddings	Med	High	FAD; MuLan cycle consistency	Partial
MusicGen (Copet et al., 2023)	Music	Single-stage AR transformer; codebooks	Med	High	FAD; subjective listening	Partial
Sora (Brooks et al., 2024)	Video	Transformer over space-time latent patches	Med	High	Qualitative; motion consistency	Partial

GPT-4 (OpenAI, 2023)	Text (img/text in)	AR transformer; RLHF; multimodal input only	High	High	Human preference; benchmarks	Partial (input only)

Three patterns are worth pulling out. (i) Controllability and autonomy are not mirror images of each other. AARON is high-autonomy, medium-controllability; SD with ControlNet is high in both; the same SD base model without ControlNet is medium in controllability. Reading the dimension means looking at the symbolic scaffolding around the base model, not just the base model itself. (ii) Evaluation methods are heterogeneous and not directly comparable across modalities (FID and HPSv2 in vision; FAD and MOS in audio; nothing widely shared in text). (iii) Among the eighteen systems, only UPIC qualifies as cross-modal generation in sense 4 (graphical input → sonic output, with no text in between). To the author's knowledge, no widely adopted production system currently offers verifiable cross-modal artistic translation in this strong sense.

"Cross-modal" gets used in at least five different ways in the literature, of increasing stringency.

- **Sense 1 – Text-conditioned generation.** Text in, single non-text modality out (Stable Diffusion, MusicLM, Sora). The weakest sense; most "multimodal" systems are cross-modal in this sense alone.
- **Sense 2 – Multimodal retrieval.** Inputs from multiple modalities mapped into a shared embedding space for retrieval, similarity, or zero-shot classification (CLIP). Nothing artistic is generated; the system enables generation by other systems.
- **Sense 3 – Multimodal representation learning.** Joint multimodal representation supporting downstream generation in any modality (ImageBind). Technically demanding but not itself generative.
- **Sense 4 – Cross-modal generation.** One modality in, a different non-text modality out - UPIC: graphical → sonic. Image-to-music systems based on shared embedding spaces often turn out, on inspection, to be image → text → music pipelines, with text smuggled in as an intermediary.
- **Sense 5 – Genuine cross-modal artistic translation.** An artwork in one modality, an artwork in another, with the aesthetic content preserved – emotional register, structural shape, prosodic or harmonic rhythm. No production system is currently cross-modal in sense 5.

The conflation matters. Marketing claims of "multimodal" almost always mean sense 1 or sense 2; genuine artistic translation (sense 5) is much harder and is still open. The technical obstacles to sense 5 are real: there is no unified

representational framework that captures aesthetically relevant content across modalities (CLIP-style embeddings capture semantic content but not, for instance, prosody or harmonic motion); there is no large supervision corpus pairing paintings with their "correct" musical translations; and there is no agreed criterion for fidelity. Promising directions include shared aesthetic embedding spaces extending CLIP's contrastive paradigm to additional modalities (CLAP for audio; ImageBind across six modalities), ontology-driven translation frameworks, diffusion bridges between modality-specific latent spaces and neuro-symbolic architectures combining learned representations with explicit aesthetic theory.

Open Problems

There is no consensus on how to judge the aesthetic quality of generated art. The technical metrics in widest use – FID and Inception Score for images, FAD for audio – measure distributional similarity to a reference set, not aesthetic merit. Human-preference protocols (PickScore, ImageReward, HPSv2; MOS-style listening tests) are more ecologically valid but expensive and confounded by novelty. Symbolic-era systems could be evaluated against their own rules; latent statistical systems offer no such reference. The computational-creativity literature – Boden [21], Ritchie [33], Colton [34] and Lamb-Brown-Clarke [7] – has not been substantially absorbed into the generative-art evaluation literature, even though it provides exactly the vocabulary the field needs.

Three threads should be kept separate. *Thaler v. Perlmutter* [4] (2023) was affirmed by the D.C. Circuit on 18 March 2025; the U.S. Supreme Court denied certiorari on 2 March 2026, settling the human-authorship requirement for U.S. copyright. *Andersen et al. v. Stability AI* [3] and *Getty Images v. Stability AI* [35] are still in active litigation as of April 2026 on the upstream question of training-data infringement. The EU AI Act [5] imposes two distinct sets of obligations that should not be conflated: Article 50 transparency disclosures for AI-generated content and Articles 53-55 GPAI-provider duties (technical documentation, downstream-provider information, copyright-policy process and a public summary of training content). The RIAA litigation against Suno and Udio (filed 2024) is the next major training-data test case. The legal disputes are intractable in part because latent statistical representation makes attribution intractable.

Treated here as a recurring loss of the latent-statistical paradigm, partially recovered through symbolic scaffolding (Section 5.2). The active research directions – region-based editing, instruction-following diffusion (InstructPix2Pix), structured-conditioning successors and multi-agent creative pipelines – are all instances of grafting symbolic structure onto latent generation. The next frontier is folding those scaffolds into the base model itself, so per-region or per-attribute control becomes natively available.

Generative systems amplify the biases of their training corpora: Western photorealism in images; Western tonal music over maqam, gamelan, mbira; lower-resource languages in text. Because the bias lives in latent representation rather than inspectable rules, mitigation means dataset rebalancing, debiasing objectives at training time, or post-hoc steering. The cultural-homogenisation concern is partly a function of training-data asymmetry but also partly a function of the latent-statistical paradigm itself, which favours majority-mode generation.

Conclusion: Four Research Gaps

The technical evolution of generative art over six decades is one sustained shift from explicit symbolic representation to latent statistical representation, with one big gain (the technical barrier to creative production has dropped sharply) and four recurring losses. Four concrete research gaps follow.

- **Gap 1.** Bring the computational-creativity evaluation criteria – Boden [21], Ritchie [33], Colton [34] and Lamb-Brown-Clarke [7] – into the standard generative-art evaluation suite alongside FID, FAD and human preference.
- **Gap 2.** Account architecturally for why so much of the practical fine-grained-controllability gain in diffusion-era systems has come from grafting symbolic scaffolding back onto latent generation and where the limit of that strategy lies. This is a research programme, not a single paper.
- **Gap 3.** Build genuine cross-modal artistic translation in the strong sense (Section 6, sense 5), with an evaluation framework that distinguishes it from text-conditioned generation in a single output modality (sense 1). The conflation of these senses in marketing and in the technical literature is the most analytically important confusion in generative art today.
- **Gap 4.** Develop a workable account of authorship attribution under latent statistical representation, sufficient to ground the next decade of legal proceedings post-*Thaler* [4].

The next decade of work in the field will, on this article's reading, be largely about recovering – through symbolic scaffolding, cross-modal grounding, computational-creativity evaluation and attribution research – the properties the shift gave away.

Declaration of AI-Assisted Writing

AI-assisted tools based on ChatGPT 5.4 and Claude Opus 4.6, were used for language refinement, structural feedback and editorial revision. The author

remains fully responsible for the originality, accuracy, references and final content of the manuscript.

Acknowledgements

The research is supported by the project FP25-FMI-010 "Innovative interdisciplinary research in Informatics, Mathematics and Pedagogy of Education" of the Scientific Fund of the Paisii Hilendarski University of Plovdiv, Bulgaria.

Appendix A. Coding Rubric for Table 1

Applied by a single rater; independent replication is recommended.

Controllability – Low: only parameters, curation, or indirect steering. Medium: global conditioning, such as prompts, labels, or fitness functions. High: explicit semantic, spatial, structural, or multimodal control.

Autonomy – Low: mainly reproduces or recombines human-authored content. Medium: makes local generative decisions under user constraints. High: makes global structural decisions beyond the user's explicit input.

Cross-modal capability – No: single-modality input and output. Partial: cross-modality only at input or output. Yes: different modalities at both input and output.

References

[1] R. Po, W. Yifan, V. Golyanik, K. Aberman, J. Barron, A. Bermano, et al., State of the art on diffusion models for visual computing, *Computer Graphics Forum*, 2024, vol. 43, no. 2, e15063, ISSN: 0167-7055, DOI: 10.1111/cgf.15063

[2] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P. Heng, S. Li, A survey on generative diffusion models, *IEEE Transactions on Knowledge and Data Engineering*, 2024, vol. 36, no. 7, pp. 2814-2830, ISSN: 1041-4347, DOI: 10.1109/TKDE.2024.3361474

[3] Andersen, et al. v. Stability AI Ltd. et al., No. 3:23-cv-00201 (N. D. Cal., filed 2023; in active litigation as of April 2026)

[4] Thaler v. Perlmutter, 687 F. Supp. 3d 140 (D.D.C. 2023), aff'd, 130 F.4th 1039 (D.C. Cir. 2025), cert. denied, 2 Mar. 2026

[5] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Official Journal of the European Union, 12 July 2024

[6] J. Briot, G. Hadjeres, F. Pachet, *Deep Learning Techniques for Music Generation*, Springer, Cham, 2020, ISBN: 978-3-319-70162-2

[7] C. Lamb, D. Brown, C. Clarke, Evaluating computational creativity: An interdisciplinary tutorial, *ACM Computing Surveys*, 2018, vol. 51, no. 2, Article 28, pp. 1–34, ISSN: 0360-0300, DOI: https://doi.org/10.1145/3167476

[8] A. Noll, Human or machine: A subjective comparison of Piet Mondrian's 'Composition with Lines' and a computer-generated picture, *The Psychological Record*, 1966, vol. 16, no. 1, pp. 1–10, ISSN: 0033-2933

[9] A. Lindenmayer, Mathematical models for cellular interactions in development, *Journal of Theoretical Biology*, 1968, vol. 18, no. 3, 280–299, ISSN: 0022-5193, DOI: https://doi.org/10.1016/0022-5193(68)90079-9

[10] P. Prusinkiewicz, A. Lindenmayer, *The Algorithmic Beauty of Plants*, Springer-Verlag, New York, 1990, ISBN: 978-0-387-94676-4

[11] B. Mandelbrot, *The Fractal Geometry of Nature*, W. Freeman, New York, 1982, ISBN: 978-0-7167-1186-5

[12] S. Wolfram, *A New Kind of Science*, Wolfram Media, Champaign, IL, 2002, ISBN: 978-1-57955-008-0

[13] H. Cohen, The further exploits of AARON, painter, *Stanford Humanities Review*, vol. 4, no. 2, 1995, pp. 141–158, ISSN: 1063-7095

[14] I. Xenakis, *Formalized Music: Thought and Mathematics in Composition*, Indiana University Press, Bloomington, 1971, ISBN: 978-1-57647-079-4

[15] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Computation*, 1997, vol. 9, no. 8, pp. 1735–1780, ISSN: 0899-7667, DOI: https://doi.org/10.1162/neco.1997.9.8.1735

[16] K. Sims, Evolving virtual creatures, *Proc. of SIGGRAPH 1994*, ACM, New York, 1994, pp. 15–22, ISBN: 978-0-89791-667-7, DOI: https://doi.org/10.1145/192161.192167

[17] A. Mordvintsev, C. Olah, M. Tyka, Inceptionism: Going deeper into neural networks, Google Research Blog, 17 June 2015, URL: https://research.google/blog/inceptionism-going-deeper-into-neural-networks/

[18] L. Gatys, A. Ecker, M. Bethge, Image style transfer using convolutional neural networks, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2414–2423, DOI: https://doi.org/10.1109/CVPR.2016.265

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, *Advances in Neural Information Processing Systems*, 2014, vol. 27, pp. 2672–2680, DOI: https://doi.org/10.48550/arXiv.1406.2661

[20] A. Elgammal, B. Liu, M. Elhoseiny, M. Mazzone, CAN: Creative adversarial networks, generating 'art' by learning about styles and deviating from style norms, *Proc. International Conf. on Computational Creativity (ICCC)*, 2017, pp. 96–103, DOI: https://doi.org/10.48550/arXiv.1706.07068

[21] M. Boden, *The Creative Mind: Myths and Mechanisms*, 2nd ed., Routledge, London, 2004, ISBN: 978-0-415-31453-7

[22] D. Kingma, M. Welling, Auto-encoding variational Bayes, *arXiv preprint*, arXiv:1312.6114, 2013, DOI: https://doi.org/10.48550/arXiv.1312.6114

[23] P. Dhariwal, H. Jun, C. Payne, J. Kim, A. Radford, I. Sutskever, Jukebox: A generative model for music, *arXiv preprint*, arXiv:2005.00341, 2020, DOI: https://doi.org/10.48550/arXiv.2005.00341

[24] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 6840–6851, DOI: https://doi.org/10.48550/arXiv.2006.11239

[25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10684–10695, DOI: https://doi.org/10.1109/CVPR52688.2022.01042

[26] L. Zhang, A. Rao, M. Agrawala, Adding conditional control to text-to-image diffusion models, Proc. IEEE/CVF International Conf. on Computer Vision (ICCV), 2023, pp. 3836–3847, DOI: https://doi.org/10.1109/ICCV51070.2023.00355

[27] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, K. Aberman, DreamBooth: Fine-tuning text-to-image diffusion models for subject-driven generation, *Proc. of IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 22500–22510, DOI: https://doi.org/10.1109/CVPR52729.2023.02155

[28] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, *Proc. International Conf. on Learning Representations (ICLR)*, 2022, DOI: https://doi.org/10.48550/arXiv.2106.09685

[29] A. Agostinelli, T. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, et al., MusicLM: Generating music from text, *arXiv preprint*, arXiv:2301.11325, 2023, DOI: https://doi.org/10.48550/arXiv.2301.11325

[30] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, et al., Video generation models as world simulators, *arXiv preprint*, arXiv:2402.17177, 2024, DOI: https://doi.org/10.48550/arXiv.2402.17177

[31] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al., Language models are few-shot learners, *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 1877–1901, DOI: https://doi.org/10.48550/arXiv.2005.14165

[32] OpenAI, GPT-4 technical report, *arXiv preprint*, arXiv:2303.08774, 2023, DOI: https://doi.org/10.48550/arXiv.2303.08774

[33] G. Ritchie, Some empirical criteria for attributing creativity to a computer program, *Minds and Machines*, 2007, vol. 17, no. 1, pp. 67–99, ISSN: 0924-6495, DOI: https://doi.org/10.1007/s11023-007-9066-2

[34] S. Colton, Creativity versus the perception of creativity in computational systems, *AAAI Spring Symposium on Creative Intelligent Systems*, 2008, pp. 14–20

[35] Getty Images (US), Inc. v. Stability AI, Inc., No. 1:23-cv-00135 (D. Del., filed 2023; in active litigation as of April 2026)

Viktor Matanski

Paisii Hilendarski University of Plovdiv,

Faculty of Mathematics and Informatics,

236 Bulgaria Blvd., 4027 Plovdiv, Bulgaria

Corresponding author: matanski@uni-plovdiv.bg