Multimodal Generative AI for Personalized Therapeutic Art and Music Experiences

Pedro D. Makinen
School of Computing, Clemson University, Clemson, SC, USA.
contactpedro@clemson.edu

Anton Miles
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
anton499@ucf.edu

Abstract

The convergence of generative artificial intelligence and arts-based therapeutic practice is creating new possibilities for personalized mental health and wellness interventions. This paper presents a system-level analysis of multimodal generative AI platforms designed to deliver individualized therapeutic art and music experiences. Rather than focusing solely on algorithmic performance, the discussion emphasizes structural trade-offs in architecture, personalization, infrastructure, governance, deployment, robustness, fairness, and policy. The paper examines how text-to-image, text-to-audio, and cross-modal latent models can be integrated into coherent therapeutic systems while preserving clinical safety and user agency. It further analyzes the requirements for adaptive user modeling, privacy-preserving data infrastructure, multimodal alignment, and longitudinal engagement. Because therapeutic contexts introduce heightened ethical obligations, the paper considers how bias, accessibility, accountability, and regulatory compliance shape system design. The analysis situates generative therapeutic platforms within broader trends in foundation models, human-centered artificial intelligence, and digital health ecosystems. It argues that sustainable deployment depends not only on generative quality but also on institutional governance, interoperability with clinical workflows, auditability, and equitable access. The paper concludes with forward-looking perspectives on how multimodal generative AI can evolve into a responsible sociotechnical infrastructure for arts-based wellness without displacing human therapeutic judgment.

Keywords

generative artificial intelligence; multimodal systems; therapeutic art; music therapy; personalization; sociotechnical infrastructure; AI governance.

1. Introduction

Generative artificial intelligence has undergone a fundamental transformation over the past decade, driven by advances in adversarial learning, attention-based sequence modeling, and cross-modal representation. Early generative adversarial networks established a powerful framework for synthesizing high-dimensional visual content through competitive learning between generator and discriminator networks [1]. Subsequent developments in transformer architectures expanded generative capabilities across language, vision, and audio by enabling scalable attention over long sequences and large corpora [2]. Multimodal contrastive models such as CLIP introduced a shared embedding space in which visual and linguistic representations could be aligned, enabling zero-shot generation and semantically guided synthesis [3]. Diffusion-based generative models further improved the stability and photorealism of image synthesis by modeling iterative denoising processes [4]. In parallel, autoregressive models such as WaveNet demonstrated that raw audio waveforms could be generated with high fidelity, opening new directions for music and sound synthesis [5]. These algorithmic advances have collectively moved generative AI from experimental demonstration toward practical deployment in creative industries, clinical support tools, and wellness applications.

The arts have long been recognized as a meaningful contributor to health and well-being. A broad review of public health literature links engagement with visual art, music, and creative expression to improvements in emotional regulation, social connection, and perceived quality of life [6]. Within mental health care, music therapy has been examined as an adjunctive treatment for depression, with evidence suggesting beneficial effects on mood and functioning [7]. The emergence of high-quality multimodal generative systems therefore creates an opportunity to personalize therapeutic art and music experiences in ways that were previously difficult to scale. At the same time, the clinical and wellness context imposes requirements that are different from general-purpose creative tools. A therapeutic system must maintain reliability, respect privacy, avoid harmful content, support clinical oversight, and adapt to individual emotional and cognitive states. This paper develops a system-level account of multimodal generative AI for personalized therapeutic art and music experiences, examining architectural choices, personalization mechanisms, governance structures, deployment constraints, and policy implications.

2. Multimodal Therapeutic Experience Landscape and System Requirements

Therapeutic art and music experiences operate across a continuum that includes clinical therapy, structured wellness programs, and informal self-directed creative practice. In clinical settings, therapeutic goals may include emotional expression, reduction of anxiety, improvement of mood, and enhancement of self-awareness. Music therapy interventions can be active, involving improvisation or performance, or receptive, involving listening and guided reflection [7]. Visual art therapy similarly uses drawing, painting, collage, and digital media to support nonverbal expression. When generative AI is introduced into this landscape, the system must do more than produce aesthetically pleasing content. It must produce content that is contextually appropriate, emotionally aligned, and safe for vulnerable users. This raises requirements for interpretability, controllability, and alignment between user intent and generated output.

Multimodal generative systems can support therapeutic goals by combining visual and auditory modalities in a shared generative loop. For example, a user might provide a brief description of a current emotional state, and the system might respond with a short musical phrase and a corresponding visual scene that jointly express a validating emotional tone. Hierarchical text-to-image generation models have demonstrated the ability to produce detailed visual content conditioned on natural language, while latent diffusion approaches allow fine-grained control over style and composition [8]. However, therapeutic value depends on more than prompt fidelity. The system must account for the user’s affective trajectory over time, avoid overstimulation, and provide opportunities for reflection rather than passive consumption. These requirements shape architecture, data governance, and evaluation methods.

3. Architectural Foundations for Multimodal Generative Therapy Systems

A multimodal generative platform for therapeutic experiences typically comprises several coordinated subsystems. A sensory input layer may collect text prompts, voice features, physiological signals, or user-selected emotional tags. A representation layer maps these inputs into shared latent spaces that support cross-modal reasoning and content generation. A generative layer synthesizes visual and audio outputs, often using diffusion models, autoregressive models, or hybrid approaches. A therapeutic adaptation layer applies rules and learned policies to adjust content according to user state, session goals, and safety constraints. A delivery layer presents the generated content through web, mobile, or immersive interfaces and collects user feedback for future adaptation.

A central architectural trade-off concerns the degree of coupling between multimodal generative components. Tightly coupled systems can optimize cross-modal coherence by sharing latent representations and gradient signals during training. Transformer-based architectures have facilitated this integration by treating image tokens, text tokens, and audio tokens within unified sequence models [2]. However, tightly coupled training pipelines may be difficult to update independently and can create failure propagation when one modality degrades. Loosely coupled systems, in which a text conditioner drives separate image and audio generators, offer modularity and easier maintenance but may sacrifice cross-modal synchrony. In therapeutic settings, loose coupling can be acceptable when the primary goal is self-expression rather than cinematic synchronization. Nonetheless, multimodal coherence still matters, because mismatched emotional cues between visual and musical content can confuse users and reduce therapeutic trust.

Another architectural consideration is the balance between latent-space generation and template-based personalization. Latent generative models provide open-ended novelty, which may be valuable for sustaining engagement over repeated sessions. However, open-ended generation can produce content that is therapeutically inappropriate or emotionally jarring. Template-based systems offer greater control and predictable clinical safety but may lack the richness needed for personalized expression. Many practical platforms therefore adopt a hybrid design in which generative models produce candidate outputs and rule-based or model-based safety filters constrain the final selection. The personalization layer can then operate over this constrained space, selecting or ranking outputs according to user history, therapeutic goals, and contextual features.

4. Personalization Mechanisms and User Modeling

Personalization in therapeutic generative AI must be understood as a longitudinal process rather than a single inference task. A user’s emotional state, aesthetic preferences, and therapeutic needs change over time, and the system should adapt accordingly. User modeling may draw on self-reported mood ratings, historical interactions, linguistic features of prompts, and optional physiological signals. However, multimodal personalization also introduces significant complexity because each modality carries different information about preference and affect. Visual elements such as color, texture, and composition may evoke distinct emotional responses, while musical elements such as tempo, mode, dynamics, and timbre influence arousal and valence. A robust personalization system must therefore model cross-modal preferences while avoiding overfitting to transient states.

Recent work on AI-assisted artistic musical co-creation involving humans and pets illustrates how generative systems can support wellness-oriented creative expression through multimodal interaction [9]. Such approaches suggest that co-creative engagement can be more therapeutically meaningful than passive generation because users actively shape the creative process. In a personalization framework, the system can learn to present users with controllable generative parameters, such as emotional intensity, energy level, and sensory density, rather than making fully autonomous decisions. This preserves user agency, which is central to many arts-based therapeutic approaches. The system may also use reinforcement signals from explicit feedback or implicit engagement patterns to refine future recommendations. However, personalization must be balanced against the risk of creating closed feedback loops that narrow the user's emotional range over time.

5. Data Governance, Privacy, and Ethical Infrastructure

Therapeutic generative AI systems operate in a sensitive data environment. User prompts, generated content, engagement patterns, and potentially physiological signals can reveal intimate information about mental health, identity, and emotional vulnerability. Data governance therefore cannot be treated as an afterthought. Model cards and structured documentation practices provide one mechanism for clarifying the intended use, limitations, and evaluation criteria of generative models [10]. Such documentation can support institutional review, clinical integration, and user-facing transparency. At the same time, the scale and opacity of large generative models present broader risks, including the possibility that systems trained on uncurated internet data may reproduce harmful stereotypes, inappropriate content, or biased associations [11]. In therapeutic contexts, these risks are amplified because users may be in emotionally sensitive states and more susceptible to harm.

A privacy-preserving infrastructure for therapeutic generative AI should incorporate data minimization, purpose limitation, and robust consent mechanisms. Personalization requires historical context, but this does not imply unlimited retention. Systems can use federated learning, differential privacy, or local personalization to reduce centralized data exposure. Governance structures should define how user-generated content is stored, whether it is used for model improvement, and how users can request deletion. In clinical deployments, data governance must also align with health privacy regulations and institutional oversight frameworks. The design of consent processes is particularly important because users may not fully understand how generative models learn from their interactions. Ethical infrastructure should therefore include plain-language explanations, granular permissions, and periodic consent renewal.

6. Fairness, Accessibility, and Sociotechnical Equity

Fairness in therapeutic generative AI extends beyond aggregate statistical parity to consider how different user populations experience creative and emotional systems. Health algorithms have been shown to embed racial and socioeconomic biases when trained on nonrepresentative data or biased labels, leading to unequal care quality [12]. In generative systems, bias may appear in the visual and musical styles that are considered normative, the emotional expressions that are validated, or the cultural references embedded in prompts and outputs. A system that consistently generates therapeutic content aligned with dominant aesthetic traditions may fail to serve users from diverse cultural backgrounds. Fairness therefore requires representation across cultures, languages, artistic traditions, and musical idioms.

Accessibility is a further sociotechnical requirement. Multimodal generative interfaces should accommodate users with visual, auditory, motor, or cognitive impairments. For example, text-to-image systems may be inaccessible to users with visual impairments unless accompanied by rich audio descriptions or haptic feedback. Similarly, music generation tools may exclude users with hearing impairments unless they offer visual or tactile alternatives. Design choices such as adjustable pacing, simplified prompt interfaces, and multimodal redundancy can broaden access. Fairness and accessibility must be treated as system-level properties rather than post hoc add-ons, because they influence model training, interface design, evaluation metrics, and deployment partnerships. Without deliberate attention to these dimensions, therapeutic generative AI risks reinforcing the very inequities that wellness interventions often seek to reduce.

7. Robustness, Safety, and Clinical Risk Management

Robustness in generative therapeutic systems involves both technical performance under distribution shift and contextual safety under emotional and clinical variation. Large language models and generative systems can produce plausible but harmful outputs, including toxic language, graphic imagery, or emotionally invalidating content [13]. In a therapeutic setting, such failures can undermine trust and potentially exacerbate distress. Safety mechanisms must therefore operate at multiple levels, including prompt filtering, output moderation, session-level monitoring, and human escalation pathways. However, overzealous filtering can also reduce the expressive range that makes art and music therapeutic. Striking this balance requires a risk management framework that distinguishes between mild discomfort, which may be part of creative exploration, and genuine harm, which requires intervention.

Accountability structures are essential for clinical risk management. Internal algorithmic auditing can provide a systematic process for documenting model behavior, testing failure modes, and tracking remediation across the development lifecycle [14]. In therapeutic contexts, audits should include domain-specific evaluations such as emotional appropriateness, cultural sensitivity, and alignment with therapeutic objectives. Foundation models introduce additional governance challenges because their broad capabilities can be adapted to many downstream uses, making it difficult to anticipate all failure modes during pretraining [15]. Robust deployment therefore depends on continuous monitoring, staged release, and clear delineation of clinical responsibility. Clinicians, platform operators, and model developers may share overlapping duties, but the system must make these duties explicit to avoid accountability gaps.

8. Deployment Models and Sustainability

Deployment of multimodal therapeutic generative AI can take several forms, including direct-to-consumer wellness applications, clinical adjunct tools, and hybrid models involving human therapists. Each deployment context imposes different requirements for reliability, oversight, and integration. In clinical settings, generative tools should interface with electronic health records, support clinician review, and produce outputs that can be discussed in therapeutic sessions. The broader movement toward high-performance medicine highlights both the promise and the complexity of integrating AI into care pathways, especially when clinical decisions involve subtle patient context [16]. Therapeutic art and music systems may not make clinical diagnoses, but they can still influence emotional states and should therefore be subject to appropriate validation.

Sustainability depends on evidence, infrastructure cost, and institutional support. The World Health Organization has called attention to the need for rigorous evidence on arts and health interventions, noting that many studies are heterogeneous in design and outcome measurement [17]. Generative systems introduce additional variability because outputs are probabilistic and often nonreproducible. Longitudinal evaluation should combine quantitative measures of engagement and symptom change with qualitative assessments of user experience. Sustainability also depends on the computational costs of large generative models, which can strain public health and community arts budgets. Distillation, caching, and on-device generation may reduce long-term operational costs, but these techniques require trade-offs in output quality and personalization depth. Health AI systems have demonstrated the value of rigorous validation before clinical translation [18], and multimodal therapeutic platforms should adopt a similarly staged approach.

Multimodal machine learning itself remains a challenging research area, with open questions in representation, alignment, and translation across modalities [19]. These challenges are directly relevant to therapeutic deployment because visual and musical signals carry affective information in culturally and individually specific ways. A system that aligns modalities poorly may produce content that is technically impressive but emotionally empty. Deployment teams should therefore include clinicians, artists, human-computer interaction researchers, and engineers who can collectively evaluate whether generated content supports the intended therapeutic process. Without such interdisciplinary oversight, the system may optimize for engagement metrics or aesthetic novelty rather than therapeutic value.

9. Regulatory and Policy Implications

Regulatory frameworks for generative therapeutic systems are still emerging, but they will likely draw from multiple domains, including medical device regulation, data protection law, consumer safety, and AI-specific legislation. Bias audits and fairness assessments have become important tools for ensuring that automated systems do not disproportionately disadvantage vulnerable groups [20]. Therapeutic generative AI presents a distinct regulatory challenge because its outputs are creative and subjective, yet its context of use may be clinically adjacent. Policymakers must decide whether such systems should be classified as wellness tools, medical devices, or a new category of digital therapeutic. The classification will determine evidentiary requirements, oversight intensity, and liability rules.

Global AI ethics guidelines emphasize principles such as transparency, justice, non-maleficence, and accountability, but they often lack concrete implementation pathways for generative systems [21]. In the therapeutic domain, policy should encourage independent evaluation, adversarial testing, and postmarket surveillance. It should also address the use of user-generated therapeutic data for model improvement, the right to meaningful human review, and the protection of vulnerable populations. International coordination will be necessary because multimodal foundation models are often developed in one jurisdiction and deployed in many others. A policy approach that balances innovation with safety can support the development of therapeutic generative AI while protecting users from exploitative or harmful applications.

10. Conclusion

Multimodal generative AI has the potential to transform personalized therapeutic art and music experiences by enabling adaptive, multimodal, and interactive creative expression at scale. However, realizing this potential requires more than generative model quality. It requires careful architectural design that balances coherence, modularity, controllability, and cost. It requires personalization mechanisms that respect user agency and longitudinal emotional trajectories. It requires governance structures that protect privacy, document limitations, and establish clear accountability. It also requires sustained attention to fairness, accessibility, robustness, and clinical risk. The therapeutic context makes these requirements especially urgent, because users may be emotionally vulnerable and because creative outputs can shape mood, self-perception, and social connection. Future research should develop evaluation frameworks that combine affective science, clinical evidence, and sociotechnical auditing. Interdisciplinary collaboration among AI researchers, clinicians, artists, and policymakers will be necessary to build systems that are both generative and trustworthy. In this vision, generative AI becomes not a replacement for human therapeutic relationships but a flexible, accountable, and culturally responsive infrastructure that supports the expressive and healing potential of art and music.

References

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27.

2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

3. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning, PMLR, 139, 8748-8763.

4. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840-6851.

5. van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

6. Stuckey, H. L., & Nobel, J. (2010). The connection between art, healing, and public health: A review of current literature. American Journal of Public Health, 100(2), 254-263.

7. Aalbers, S., Fusar-Poli, L., Freeman, R. E., Spreen, M., Ket, J. C. F., Vink, A. C., Maratos, A., Crawford, M., Chen, X., & Gold, C. (2017). Music therapy for depression. Cochrane Database of Systematic Reviews, 2017(11), CD004517.

8. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125.

9. Wang, Z., Ma, L., Jin, Y., Feng, Y., Pan, X., Ji, S., & Zhang, K. (2025, August). AI-assisted human-pet artistic musical co-creation for wellness therapy. In Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (pp. 10216-10224).

10. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220-229.

11. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610-623.

12. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

13. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

14. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33-44.

15. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niknami, N., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

16. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.

17. Fancourt, D., & Finn, S. (2019). What is the evidence on the role of the arts in improving health and well-being? A scoping review. WHO Regional Office for Europe.

18. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

19. Baltrusaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443.

20. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35.

21. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.