Privacy-Preserving On-Device Generative AI for Personalized Healthcare Image Synthesis

Matteo Makinen
School of Computing, Clemson University, Clemson, SC, USA.
matteo.work@clemson.edu

Alessandro J. Allen
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
alessandro.work@uc.edu

Habsen Beyint
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
bryanthudson@ucf.edu

Abstract

Generative artificial intelligence is increasingly positioned as a transformative force in medical imaging, offering to produce patient-specific synthetic views that can support augmentation, simulation, education, and rare pathology representation. At the same time, personalized image synthesis creates acute privacy risks because generative models trained on sensitive health data can leak patient-specific visual information. On-device generative AI has emerged as a promising architectural response, preserving data locality while enabling personalization at the edge. This paper provides a system-level analysis of privacy-preserving on-device generative AI for healthcare image synthesis. It examines the structural relationships among learning paradigms, model architectures, edge hardware, privacy mechanisms, governance, fairness, and sustainability. We argue that technical privacy guarantees alone are insufficient without careful attention to deployment context, lifecycle management, and regulatory alignment. The paper discusses federated learning, differential privacy, model compression, hardware-aware generation, and adversarial threats such as membership inference and reconstruction. It further contrasts edge-native generative models with cloud-centric foundation models and considers the implications for energy use, bias amplification, auditability, and patient consent. By integrating perspectives from machine learning, systems engineering, and health policy, the analysis provides a comprehensive framework for designing and evaluating next-generation personalized medical image synthesis systems. The discussion concludes with future research directions that emphasize standardized evaluation, federated benchmarking, sustainable hardware co-design, and governance structures that preserve public trust.

Keywords

privacy-preserving machine learning; generative adversarial networks; diffusion models; edge computing; healthcare image synthesis; federated learning; differential privacy; medical imaging; data governance; personalization.

1. Introduction

The integration of generative artificial intelligence into healthcare imaging has opened new possibilities for personalized diagnosis, procedural simulation, and medical education.

Generative adversarial networks introduced an adversarial sampling framework that made high-fidelity image synthesis possible [3]. More recently, denoising diffusion probabilistic models have demonstrated that iterative denoising processes can produce highly detailed images with stable training dynamics [4]. Latent diffusion formulations further reduced computational demands by shifting the generative process into a compressed representation space, making high-resolution synthesis more practical across heterogeneous hardware environments [5]. These advances are especially relevant for healthcare because patient-specific synthetic images can be used to model rare anatomical variations, simulate disease progression, augment small local datasets, and support clinician training without requiring additional invasive imaging.

However, personalized image synthesis in healthcare is inseparable from privacy risk. Training or adapting a generative model on patient data can cause the model to memorize and reproduce identifying visual features. Centralized cloud-based pipelines exacerbate this risk because they concentrate sensitive images, model parameters, and optimization signals in a single infrastructure. Federated learning has emerged as a widely studied mechanism for computing model updates locally and sharing only aggregated parameters across institutions [1]. Differential privacy adds formal noise-based protections that limit the influence of any single record on the learned model [2]. When combined with on-device generation, these approaches create a potential pathway toward personalized synthesis without transferring raw patient images to external servers. Nevertheless, realizing this pathway requires navigating deep structural trade-offs involving model capacity, energy consumption, hardware heterogeneity, privacy accounting, fairness, and regulatory compliance.

This paper adopts a systems perspective rather than a purely algorithmic one. It treats on-device generative AI as a socio-technical infrastructure in which architecture, compute, data governance, clinical workflows, and policy interact continuously. The analysis focuses on the structural tensions between personalization quality and privacy preservation, between local autonomy and centralized orchestration, and between model fidelity and sustainable deployment. By examining these tensions, the paper contributes a framework for understanding when and how on-device generative AI can be responsibly deployed in healthcare image synthesis.

2. Background and Related Work

Federated learning has matured from a conceptual framework into a broad research field concerned with communication efficiency, heterogeneity, privacy, and robustness [6]. In medical imaging, federated approaches have been applied to multi-institutional collaborations where patient data cannot be shared directly due to institutional or legal constraints [7]. These methods allow participating hospitals to train shared segmentation, classification, or reconstruction models while keeping raw imaging studies within local networks. At the same time, privacy-preserving machine learning in medical imaging has emphasized that federated averaging alone does not guarantee confidentiality because shared gradients or model updates can leak information about local data [8]. Membership inference attacks have shown that even black-box models can reveal whether a particular record was used during training [9]. Formal differential privacy provides a quantitative framework to reason about such leakage by bounding the probabilistic influence of individual data points [10].

Generative models introduce additional challenges because their outputs can directly resemble training samples. This is particularly concerning in medical imaging, where reconstructed images may contain facial features, skin markings, implant geometries, or anatomical

anomalies that are strongly identifying. Early biomedical image segmentation architectures such as U-Net established the value of encoder-decoder networks that combine high-level semantic context with fine spatial detail [11]. The same representational principles have been extended to generative tasks, where the model must preserve anatomical coherence while synthesizing novel variations. An edge-native text-to-image foundation model trained entirely on domestic accelerator hardware has demonstrated that high-fidelity synthesis can be achieved under edge constraints [12]. This type of architecture matters for healthcare because it suggests that on-device generative models need not be weak approximations of large cloud models; they can be designed from the ground up for decentralized inference and personalization.

The related work therefore indicates three converging lines of inquiry: generative modeling for high-quality image synthesis, privacy-preserving distributed learning for clinical data, and edge-native system design for constrained hardware environments. Each line has advanced considerably, but their integration remains underexplored at the system level. Personalized on-device generation is not simply a matter of placing an existing diffusion model on a smartphone. It requires rethinking how patient-specific style, anatomy, and pathology are encoded, how privacy budgets are spent across long lifecycles, and how model updates are governed across clinical institutions.

3. System Architecture for On-Device Personalized Synthesis

A privacy-preserving on-device generative system for healthcare image synthesis typically includes several interacting components. The first component is a learned encoder that maps raw or preprocessed medical images into a compact representation. The second is a generative prior that models the distribution of plausible anatomical structures. The third is a conditioning mechanism that injects patient-specific information, clinical labels, or clinician prompts. The fourth is a decoder or super-resolution stage that reconstructs an image of sufficient resolution for clinical inspection. In a centralized deployment, all components may reside in a large cloud cluster with abundant memory and high-throughput accelerators. In an edge deployment, the same functional decomposition must be compressed, partitioned, or selectively offloaded to accommodate limited memory, thermal envelopes, and energy budgets.

Latent diffusion models offer a useful architectural foundation because their generative process operates in a compressed latent space rather than directly at pixel resolution [5]. This reduces computational effort during iterative sampling and permits smaller on-device denoising networks. However, even latent diffusion models can be too large for many clinical edge devices. Further compression through distillation, quantization, pruning, and sparse attention is often necessary. These techniques reduce the number of arithmetic operations and memory transfers per generation step, but they also introduce trade-offs. Aggressive compression can degrade fine anatomical detail, increase the likelihood of hallucinations, and reduce robustness to rare pathologies. System designers must therefore balance compression against the clinical requirement for trustworthy visual evidence.

Personalization introduces another architectural challenge. A personalized model must adapt to a specific patient or a small set of patients without accumulating excessive sensitivity to their private features. Low-rank adaptation and similar parameter-efficient tuning methods can be used to create small patient-specific modules that modify a shared base model without retraining the entire network. These modules can be stored locally, encrypted, and updated as new imaging data arrive. The resulting architecture supports a separation between a shared

generative prior and local personalization deltas. This separation is structurally valuable because it reduces the amount of patient-specific information that must be transmitted or aggregated. It also creates a natural boundary for privacy auditing, since the personalized modules are the primary location of sensitive visual information.

A concrete illustration can be found in portable ultrasound. Clinicians using handheld ultrasound probes may benefit from conditional synthesis that completes or enhances views under limited scanning conditions. An on-device model could generate anatomically plausible longitudinal or cross-sectional views from partial acquisitions while preserving local acoustic measurements. In this setting, low latency and offline operation are important because emergency and rural settings may lack reliable connectivity. The structural trade-off is clear: a large cloud model may produce sharper images, but an edge model can operate without transferring patient data or depending on network availability. This trade-off resembles the evolution of computational photography in consumer smartphones, where on-device neural processing enables real-time enhancement while avoiding cloud upload. Similar architectural logic applies to portable magnetic resonance imaging, capsule endoscopy, and bedside dermatology.

4. Privacy-Preserving Learning and Inference Strategies

Privacy protection in on-device generative AI operates at multiple layers. The first layer is data minimization: keeping raw patient images on the device prevents the most direct form of data exposure. Federated learning extends this principle by allowing local training and periodic aggregation of model updates [1]. However, federated learning alone does not make the system private because model updates can encode information about local data. Differential privacy provides a complementary layer by adding calibrated noise to gradients, weights, or generated outputs [2]. The formal guarantees of differential privacy are especially relevant in healthcare because they provide a quantifiable basis for discussing residual risk [10].

Personalized generative models raise a distinct privacy challenge. When a model is fine-tuned on a small number of patient images, the risk of memorization increases. The resulting model may be able to reconstruct visual details that were present in only one training sample. Membership inference attacks can exploit such memorization to determine whether a particular patient participated in fine-tuning [9]. On-device personalization mitigates some of this risk by keeping the fine-tuned model local, but the model itself may still be vulnerable to extraction if the device is lost, stolen, or inspected by an unauthorized party. Therefore, on-device systems should include protections such as parameter encryption, secure enclaves, and differential privacy during local adaptation.

A further layer of protection concerns the inference pipeline. Even when a generative model is not explicitly trained on a patient, conditional generation based on patient-specific inputs may reveal correlated information. For example, a model that synthesizes a full organ from a partial scan may encode the shape and pathology of that organ in a way that could be reconstructed by an adversary with access to model outputs. Techniques from secure multiparty computation, trusted execution environments, and homomorphic encryption can reduce exposure during inference, but they introduce latency, energy, and compatibility costs. In practice, many clinical edge devices cannot yet support full homomorphic inference at interactive speeds. The field therefore faces a structural trade-off between cryptographic rigor and clinical usability.

This trade-off can be managed through contextual privacy engineering. A device used in a controlled hospital setting may accept higher latency for occasional high-sensitivity synthesis tasks, while continuous monitoring applications may require fast approximate protections. The privacy architecture should also distinguish between anonymized synthetic teaching images and patient-specific clinical images. The former can be subjected to stricter noise and release controls, while the latter may remain inside the clinical device boundary. Such layered strategies are more aligned with real healthcare workflows than an all-or-nothing approach.

5. Infrastructure, Deployment, and Sustainability

The infrastructure required for privacy-preserving on-device generative AI differs significantly from centralized foundation model deployment. Cloud-based generative models depend on large datacenters, high-bandwidth networks, and homogeneous accelerator fleets. On-device generation, by contrast, must operate across a heterogeneous collection of smartphones, tablets, bedside workstations, embedded medical devices, and edge servers. This heterogeneity creates engineering challenges around operator compatibility, memory capacity, processor instruction sets, and thermal management. An edge-native text-to-image foundation model trained entirely on domestic accelerator hardware illustrates that alternative hardware ecosystems can support sophisticated generative workloads [12]. Such developments have implications for healthcare systems seeking to avoid vendor lock-in or to maintain functionality during supply chain disruptions.

Deployment architecture can range from fully offline generation to hybrid split inference. Fully offline generation requires the entire compressed model and all personalization modules to reside on the device. This maximizes privacy and availability but limits the use of larger models or frequent global updates. Hybrid split inference divides the generative process between the device and a trusted edge server. The device may handle the encoder and early denoising steps, while the edge server completes the final high-resolution synthesis. This approach can improve quality, but it expands the data exposure surface and requires careful channel security and minimization. Federated orchestration may be used to update shared model components without collecting patient-specific images [6]. The orchestration layer must track model versions, privacy budgets, consent states, and audit records.

Sustainability is an essential but often underestimated dimension. Generative sampling is computationally intensive, especially for diffusion models that require multiple iterative steps. Running such models on battery-powered portable devices can reduce device longevity and increase heat generation. Although compression reduces energy per sample, it may require more extensive fine-tuning or more frequent regeneration when outputs fail clinical quality checks. From a lifecycle perspective, the carbon cost of training, compressing, distributing, and updating models must be considered alongside the energy cost of daily inference. Healthcare systems increasingly face pressure to report environmental impact and to avoid wasteful computational practices. A privacy-preserving on-device system should therefore include energy-aware schedulers, adaptive inference strategies that modulate denoising steps according to image complexity, and hardware-aware model selection that matches model size to the capabilities of the device.

The economic sustainability of decentralized generative AI also deserves attention. Centralized foundation models benefit from economies of scale in datacenter procurement and model serving. Edge-native models require investment in heterogeneous hardware support, longer validation cycles, and more complex deployment management. Yet decentralization can reduce network transmission costs, lower dependency on continuous

connectivity, and distribute the burden of hardware refresh across many endpoints. In regions where healthcare infrastructure is fragmented or bandwidth is expensive, on-device generation may offer a more resilient and financially sustainable path than cloud-centric alternatives.

6. Fairness, Robustness, and Governance

Fairness in generative healthcare AI is not reducible to balanced accuracy across demographic groups. Bias can arise from underrepresentation of certain populations in training data, from differences in imaging equipment, from heterogeneous clinical protocols, and from the design choices embedded in generative priors. Research on algorithmic bias in health management has shown that structural inequities can be inadvertently encoded and amplified by predictive models [13]. A survey of fairness in machine learning further demonstrates that bias cannot be removed by a single preprocessing step; it must be addressed across task definition, data curation, model training, and deployment monitoring [14]. For on-device generative models, local personalization may improve fidelity for a specific patient but can also amplify site-specific biases if the local data are narrow or systematically distorted. A generative model that is personalized on images from one scanner type may synthesize images that are less reliable for patients imaged with another technology [15]. Robustness to equipment variation is therefore intertwined with fairness.

Medical image segmentation research has shown that self-configuring pipelines can achieve strong generalization across diverse clinical datasets, but such systems still require rigorous validation across institutions and image conditions [16]. A similar validation philosophy should apply to generative synthesis. The output of a generative model is visually rich and may appear clinically plausible even when the underlying anatomy is incorrect. This creates a governance problem: clinicians may overtrust synthetic images, especially when they are presented in the same interface as real acquisitions. Security and privacy research has also highlighted that machine learning systems can be attacked through training data, model parameters, and inference interfaces [17]. Generative systems add the unique risk of synthetic data contamination, where a model trained on its own outputs may degrade in quality and diversity. In healthcare, this could lead to systematic errors that are difficult to detect in routine use.

Robustness and fairness must therefore be evaluated jointly. A model that is robust on average but fragile for rare anatomies is not clinically fair. A model that is fair in aggregate but vulnerable to adversarial perturbations cannot be considered trustworthy. Deep learning in healthcare has consistently emphasized the gap between impressive benchmark performance and real-world clinical reliability [18]. The same gap applies to personalized image synthesis. Governance mechanisms should require documentation of training data provenance, disclosure of synthetic image status, routine bias audits, and clear procedures for clinicians to report failures. Without such structures, on-device generative AI may become another opaque algorithmic system whose errors are distributed unevenly across patient populations.

7. Policy Implications and Future Research

The deployment of privacy-preserving on-device generative AI in healthcare intersects with multiple regulatory frameworks and ethical expectations. The General Data Protection Regulation establishes principles of data minimization, purpose limitation, and accountability that are directly relevant to decentralized model training and inference [20]. In the United States, the Health Insurance Portability and Accountability Act imposes safeguards for

protected health information and creates expectations around patient rights and breach notification [21]. On-device processing can support compliance by reducing unnecessary data movement, but it does not automatically resolve all legal obligations. For example, a patient-specific generative model may itself be considered a derived representation of protected health information, and the synthetic images it produces may still require governance if re-identification is possible.

High-performance medicine increasingly depends on the convergence of multimodal data, artificial intelligence, and clinical decision support [19]. Generative synthesis adds a new modality of algorithmic output that must be integrated into this convergence without undermining trust. Future research should develop standardized evaluation protocols for on-device generative models that measure privacy leakage, clinical utility, energy use, and fairness simultaneously. Federated benchmarks that span institutions and imaging modalities would help researchers compare systems under realistic heterogeneity. Research on differential privacy in medical imaging has begun to address the tension between formal privacy and clinically meaningful performance, but much remains to be done for generative tasks [22]. Policy frameworks should also evolve to account for synthetic data that is generated, stored, and used entirely on a personal device. The boundary between a medical device, a software as a medical product, and a general-purpose computing platform is still unclear in many jurisdictions.

Future work should also examine the role of patient consent in personalized generative systems. Patients may consent to the use of their images for diagnosis but not for model personalization, or they may consent to local personalization while refusing any participation in federated aggregation. Technical systems must be able to enforce these distinctions granularly and over time. This requires identity management, consent provenance, and model release engineering that are rarely integrated into current machine learning platforms. In addition, the environmental impact of decentralized generative inference should be reported using consistent metrics, enabling healthcare institutions to make sustainability-informed procurement decisions. Research into hardware-software co-design may produce accelerators tailored to privacy-preserving generative tasks, reducing energy consumption while improving the feasibility of stronger cryptographic protections.

8. Conclusion

Privacy-preserving on-device generative AI represents a meaningful but complex trajectory for personalized healthcare image synthesis. It combines the expressive power of generative models with the data locality benefits of edge computing and the formal protections of differential privacy. Yet the system-level challenges are substantial. Architectural compression can degrade clinical fidelity, federated orchestration can introduce new governance burdens, and local personalization can increase memorization risk. Infrastructure heterogeneity and sustainability constraints further complicate deployment. Fairness and robustness cannot be treated as secondary considerations because generative errors in healthcare may directly affect clinical decision-making and patient trust. Policy frameworks must adapt to the unique properties of on-device synthetic data, and future research should develop integrated evaluation methods that bridge machine learning, systems engineering, clinical practice, and regulatory science. A successful privacy-preserving on-device generative AI ecosystem will not be defined solely by the quality of its images, but by its ability to sustain trust, equity, accountability, and clinical safety across distributed healthcare environments.

References

1. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 1273–1282.

2. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308–318.

3. 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, 2672–2680.

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

5. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10684–10695.

6. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210.

7. Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. R., & Bakas, S. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. Scientific Reports, 10, 12598.

8. Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence, 2(6), 305–311.

9. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. Proceedings of the 2017 IEEE Symposium on Security and Privacy, 3–18.

10. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407.

11. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention, 234–241.

12. Chen, Ce, et al. "JuZhou 1.0 Technical Report: The First Edge-Native Text-to-Image Foundation Model Trained Entirely on China-Developed AI Accelerators." arXiv preprint arXiv:2606.28421 (2026).

13. 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.

14. 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.

15. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31–38.

16. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods, 18(2), 203–211.

17. Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018). SoK: Security and privacy in machine learning. Proceedings of the 2018 IEEE Symposium on Security and Privacy, 399–414.

18. Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236–1246.

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

20. European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Official Journal of the European Union, L119, 1–88.

21. U.S. Department of Health and Human Services. (1996). Health Insurance Portability and Accountability Act of 1996. Public Law 104-191.

22. Ziller, A., Usynin, D., Braren, R., Makowski, M., Rueckert, D., & Kaissis, G. (2021). Medical imaging deep learning with differential privacy. Scientific Reports, 11, 13524.