xMADD: A Unified Diffusion Framework for Conditioned Synthesis of Medical Images and Waveforms

Sam Freesun Friedman[1] SAM@BROADINSTITUTE.ORG
Sana Tonekaboni[2,3] STONEKAB@MIT.EDU
Arash A. Nargesi[4,5] ANARGESI@BROADINSTITUTE.ORG
Caroline Uhler[2,3] CUHLER@MIT.EDU
Mahnaz Maddah[1] MADDAH@BROADINSTITUTE.ORG

[1] *Data Sciences Platform, The Broad Institute of MIT and Harvard, Cambridge, USA*
[2] *Eric and Wendy Schmidt Center, The Broad Institute of MIT and Harvard, Cambridge, USA*
[3] *Massachusetts Institute of Technology, Cambridge, USA*
[4] *Cardiovascular Disease Initiative, The Broad Institute of MIT and Harvard, Cambridge, USA*
[5] *Brigham and Women's Hospital, Boston, Massachusetts, USA*

Abstract

Diffusion models have shown remarkable success in generating high-quality perceptual data, but their use for controlled generation in biomedicine remains limited. We introduce xMADD (cross-Modal cross-Attention Denoising Diffusion), a conditional diffusion framework for producing diverse, high-resolution medical data, including cardiac MRI, brain MRI, and ECG waveforms, guided by clinical phenotypes, demographics, and multimodal signals. By incorporating cross-attention over conditional embeddings, xMADD enables control over generation. Compared to existing generative approaches, xMADD achieves superior image fidelity and stability, while accurately reflecting conditioning phenotypes across modalities. Our results highlight the potential of controlled diffusion-based generation to expand biomedical datasets and facilitate data-sharing without compromising sensitive patient data.

Keywords: Diffusion, Generative models, Dataset augmentation, Modality translation, Digital twins

Data and Code Availability Our study uses data from the UK Biobank imaging study (Littlejohns et al., 2020), a large-scale, prospective biomedical resource containing detailed health, genetic, and imaging data from over 500,000 participants in the United Kingdom (Bycroft et al., 2018). Data were accessed under the UK Biobank application no. 7089.

Institutional Review Board (IRB) UK Biobank has ethical approval from the North West Multi-centre Research Ethics Committee (MREC). All participants provided informed consent, and this approval covers the use of de-identified data for health-related research in the public interest. No additional ethical review was required for this analysis.

1. Introduction

Diffusion models have gained popularity as a powerful approach for synthesizing high-fidelity images across a range of domains, including natural images (Ho et al., 2020; Dhariwal and Nichol, 2021), language (Saharia et al., 2022), and computational biology (Watson et al., 2023; Guo et al., 2024). A key strength of these models lies in conditional generation, which allows outputs to be guided by labels, metadata, or multimodal signals (Kazerouni et al., 2022). In healthcare, this capability has direct relevance to precision medicine where conditional diffusion can be used to synthesize patient-specific data reflecting clinically meaningful phenotypes. Such controlled generation enables counterfactual reasoning in-silico, for example, predicting how a reduction in Body Mass Index (BMI) might alter cardiac morphology, or how an MRI would appear under different therapeutic regimens.

We present xMADD (cross-Modal cross-Attention Denoising Diffusion), a conditional diffusion framework designed to generate a variety of high-resolution biomedical data. Unlike prior generative models that typically condition on limited or unimodal signals, xMADD introduces two key innovations. First, cross-attention over conditional embeddings, enabling integration of heterogeneous signals including demographics, clinical phenotypes, and paired modalities. Second, unified conditional training and post hoc conditioning, allowing the same architecture to flexibly support both modes of guided generation.

We evaluate xMADD on the UK Biobank cohort, spanning three representative modalities: cardiac MRI, brain MRI, and ECG waveforms. Our experiments demonstrate xMADD's superior performance and utility across four key areas:

- Perceptual quality: Generating more realistic synthetic samples compared to GANs, autoencoders, and cross-modal decoders.
- Fidelity to conditioning: Accurately reflecting conditioning guidance signals.
- Cross-modal translation: Capturing meaningful mutual information across paired data modalities and using it to guide generation.
- Utility: Providing high-quality synthetic signals that enhance downstream predictive models through dataset augmentation, while preserving patient privacy by avoiding reliance on real data.

2. Related work

The concept of diffusion has a rich intellectual history with roots in thermodynamics, Brownian motion, and Fourier analysis (Fourier, 1822; Einstein, 1905). A diffusion process for deep generative modeling, adding and removing noise during training, was introduced by Sohl-Dickstein et al. (2015).

2.1. Diffusion-based Generative Models

Diffusion probabilistic models have since been applied in diverse domains ranging from natural images to audio signals (Rombach et al., 2022; Kazerouni et al., 2022). The models effectively create high-quality image samples approximating the true data distribution (Song et al., 2021a). Subsequent improvements such as refined network backbones (Dhariwal and Nichol, 2021) and score-based methods have been proposed, further improving sample quality and training stability (Song et al., 2021b).

2.2. Generative Models in Biomedicine

Generative models in medical imaging tasks have enabled many applications such as data augmentation, modality translation, and image-to-image synthesis (Kazerouni et al., 2022). GAN-based methods have shown impressive results for tasks such as MR-to-CT translation (Mirza and Osindero, 2014), data augmentation (Ahmad et al., 2022), and image fusion applications (Zhou et al., 2023; Fan et al., 2023). However, adversarial training can be unstable and is often subject to mode collapse (Nie and Patel, 2020). Recently, diffusion models have been explored for medical imaging due to their strong generative capabilities and stable training dynamics (Kazerouni et al., 2022; Pinsler et al., 2022). Diffusion-based approaches reduce the reliance on adversarial losses while offering flexibility in incorporating conditional information (Preechakul et al., 2022). These models have been used for image reconstruction (Chung and Ye, 2022), segmentation (Rahman et al., 2023), denoising (Hu et al., 2022), and generation (Pinaya et al., 2022).

2.3. Conditional Generative Models in the Medical Domain

In the medical domain, multimodal data such as MRI, CT, and physiological signals frequently co-occur, prompting interest in cross-modal synthesis (Lee et al., 2021). For instance, methods to synthesize missing imaging modalities from available ones have been explored to handle incomplete datasets or reduce patient exposure to invasive imaging (Tashiro et al., 2021; Gao et al., 2025). Many existing solutions rely on GANs (Hamghalam and Simpson, 2024; Yang et al., 2024; Amirrajab et al., 2022; Xia et al., 2021) or Variational Autoencoders (Pesteie et al., 2019). Diffusion models have been used for counterfactual MRI generation through conditioning on acquisition parameters (Morão et al., 2024) and for image translation with frequency domain filters

to preserve structure (Li et al., 2023). Diffusion autoencoders for post-training conditioning of the c-MRI modality have been used to elucidate genetic architectures (Ometto et al., 2024). xMADD unifies and simplifies these approaches, demonstrating how the same underlying architecture and training methodology supports phenotypic conditioning, modality translation, and post hoc conditioning.

3. Method

We propose a conditional denoising diffusion probabilistic model designed to generate high-resolution data across modalities, such as cardiac MRI (c-MRI), ECG waveforms, and brain MRI (b-MRI) conditioned on clinical control signals or latent space representations from paired modalities.

3.1. Model Architecture

Our proposed architecture (presented in Figure 1) is a conditional diffusion model with a convolutional U-Net backbone (Nichol and Dhariwal, 2021). It is composed of residually-connected blocks at different resolutions to capture spatial hierarchies and promote gradient flow during training (Ho et al., 2020). A cosine noise schedule is employed to enhance stability and convergence in the denoising process, aligning the noise levels with a smooth progression (Chen, 2023). Encoder and decoder architectures are mirrored with similar parameters and pooling/upsampling blocks, as they have been shown to perform similarly to asymmetric encoders and decoders (Hoogeboom et al., 2024).

To enable conditional modeling on diverse clinical control signals, we integrate a conditioning vector using cross-attention, allowing the model to selectively focus on relevant features within the conditioning context (Vaswani, 2017). Our ablation experiments in the Appendix Table 6 show cross-attention is more effective than other fusion architectures like Feature-wise Linear Modulation (Perez et al., 2018b) and simple concatenation. The architecture allows for many different sources of control signals to be integrated, including:

- **Phenotype Conditioning**: Scalar clinical variables, such as demographics, biomarkers, or diagnostic labels ($p_\theta(x_1|y)$).
- **Autoencoder Conditioning**: Embeddings from the same modality that is being synthesized ($p_\theta(x_1|e(x_1))$).
- **Cross-Modal Conditioning**: Embeddings from paired modalities that provide high-dimensional representation of complementary and mutual information ($p_\theta(x_1|g(x_2))$).

Different forms of conditioning in diffusion models enable distinct use cases with corresponding trade-offs. Phenotype conditioning allows the generation of datasets sampled from user-specified distributions, directly addressing issues such as class imbalance. Autoencoder conditioning requires no labels or paired data during diffusion training; instead, a linear probe is derived from the latent space and synthesis is performed by interpolating along probe weights. This post-training strategy is particularly valuable when phenotype labels are unavailable, as it decouples conditioning from training and enables the same diffusion backbone to generate examples for any phenotype defined later. Cross-modal conditioning requires paired modalities but enables powerful image translation tasks, which are especially relevant for building digital twins and supporting multimodal inference.

3.2. Diffusion Model Training

During training we optimize the sigmoid loss function $\mathcal{L}_{sigmoid}$ proposed by Kingma and Gao (2023). The diffusion process gradually adds noise, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$, to the original image $\mathbf{x}$ as time t increases from 0 to 1. Noise is added according to the cosine noise schedule (Ramesh et al., 2021).

$$\mathbf{z}_t = \sin\left(\frac{t/T + s}{1+s} \cdot \frac{\pi}{2}\right)\mathbf{x} + \cos\left(\frac{t/T + s}{1+s} \cdot \frac{\pi}{2}\right)\boldsymbol{\epsilon}, \tag{1}$$

Where t is the current time step, T is the total number of time steps (50 in all our experiments) and s is a small constant (0.05 in all our experiments) which prevents training on pure noise in the initial time step or pure signal in the last time

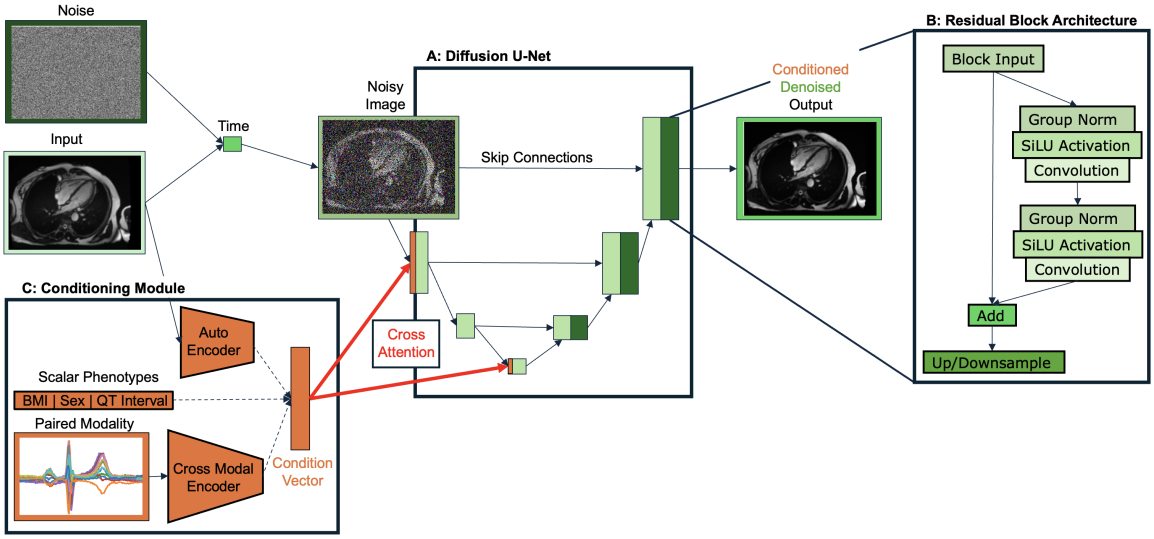

Figure 1: Overview of the xMADD architecture. (A) A diffusion U-Net backbone progressively denoises input images, with skip connections linking encoder and decoder stages. (B) Residual blocks implement group normalization, with up/down-sampling for hierarchical feature processing. (C) The conditioning module encodes different signals, including phenotypic scalars, autoencoder-derived embeddings, and paired modalities, into a condition vector, the dotted lines indicate the framework can handle any or all sources of conditioning. These vectors are integrated into the U-Net via cross-attention, enabling flexible and multimodal guidance during denoising.

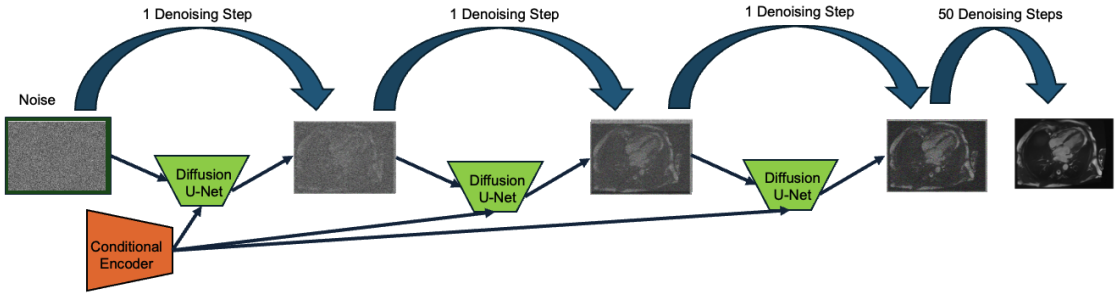

Figure 2: Inference with xMADD requires a random seed, a conditioning vector, and repeated application of the denoising U-Net.

step. The sigmoid loss can then be expressed as a weighting of the MSE loss:

$$\mathcal{L}_{sigmoid} = \exp(-b)\sigma(\lambda_t - b)||\mathbf{x} - \tilde{\mathbf{x}}||^2, \quad (2)$$

Where σ is the sigmoid function, λ_t is the log of the signal to noise ratio and the time, t. In our experiments the weighting bias is $b = 3$ as in Hoogeboom et al. (2024). Appendix Table 6 compares sigmoid with MAE loss, showing the improvements from using sigmoid weighting.

3.3. Diffusion Model Inference

Model inference is performed using a conditioning signal and an image of Gaussian noise, which is progressively denoised (Figure 2). For all syn-

thetic examples in this paper, we perform 50 denoising steps with the denoising diffusion model.

4. Experiments

We assess perceptual quality and steerability of our proposed framework across a diverse range of medical images, waveforms, conditioned on phenotypes and learned representations. We compare our framework with Wasserstein conditional GAN with gradient penalty (WGAN-GP)(Srivastava et al., 2023), autoencoders (Friedman et al., 2024), and cross-modal generative models (Radhakrishnan et al., 2023). Details of the comparison architectures are provided in Appendix A. For evaluation, we use the UK Biobank cohort (Littlejohns et al., 2020), with the specific modalities summarized in Table 1.

Table 1: Data modalities, shapes, and sample sizes in thousands.

Type	Modality	Shape	Train	Test
b-MRI	T1 MNI	192, 192	39k	5k
c-MRI	4 Chamber	160, 224	37k	5k
c-MRI	3 Chamber	224, 160	37k	5k
c-MRI	2 Chamber	224, 224	37k	5k
ECG	Median	576, 12	51k	6k

4.1. Results: Perceptual Quality

Figure 3 presents examples of images generated by xMADD. We evaluate perceptual quality using Kernel Inception Distance (KID) (Bińkowski et al., 2018) and Fréchet Inception Distance (FID) (Heusel et al., 2017), which measure similarity between the feature distributions of real and synthetic images using features from the Inception network (Szegedy et al., 2016). Although Inception is trained on natural images, prior work has shown that KID correlates more reliably with expert judgment in medical imaging than metrics derived from models trained exclusively on medical datasets (Woodland et al., 2024), making it well-suited for our setting. Lower scores indicate closer alignment with real data, reflecting both higher image quality and diversity. As shown in Table 2, xMADD achieves superior perceptual quality in both modality translation and reconstruction tasks compared to GANs, autoencoders, and cross-modal decoders. Since autoencoders cannot perform modality translation and GANs operate only on real images without reconstruction capability, we omit these baselines from the reconstruction evaluation. Notably, the negative KID observed for xMADD reconstructions arises from U-Net skip connections, which enable near-perfect reconstruction fidelity. In contrast, architectures with lossy bottlenecks, such as autoencoders or DropFuse, are inherently limited in reconstruction accuracy. Figure 4 provides qualitative examples of synthetic images and reconstructions across methods.

4.2. Results: Fidelity to Conditioning

We next assess the fidelity of the generated images to the conditioning signals. This is done by training regression or classification models on real data and then measuring the accuracy of these models' predictions on synthetic data compared with the conditioning target signal used in synthesis. For the categorical control signals, like AFib or sex, we compare the AUROC of the downstream classifier on real versus synthetic data, while for continuous controls, we measure Pearson's correlation coefficient (R). Table 3 shows synthetic ECG median fidelity across a range of phenotypes and Appendix Table 9 shows results for other modalities. The close performance of the downstream models on synthetic samples compared to real images confirms the faithful integration of the conditioning signals.

Visual inspection further confirms the effectiveness of conditional generation. Figure 5 shows examples of synthetic data generated with a fixed random seed but varying conditioning targets. On the right, axial slices of brain MRI are conditioned on Z-slice position, moving upward from the brainstem to the cortex. In the middle, cardiac MRI images are conditioned on BMI, where higher BMI corresponds to increased pericardial fat deposition (bright gray regions), enlarged myocardium, and greater overall body size. On the left, ECG signals are conditioned on QT interval values. The QT interval, spanning from the onset of the QRS complex to the end of the T wave, reflects ventricular repolar-

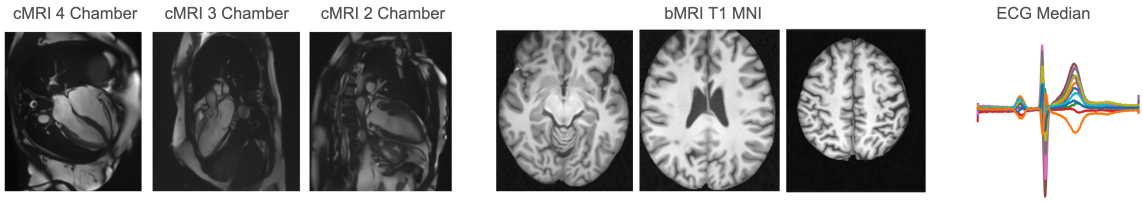

Figure 3: Synthetic medical images (c-MRI, b-MRI) and waveforms (ECG) generated by xMADD.

Table 2: Perceptual quality of c-MRI synthesis in translation and reconstruction for different architectures; GANs do not reconstruct, and autoencoders do not translate.

Model	KID Translate ↓	FID Translate ↓	KID Reconstruct ↓	FID Reconstruct ↓
xMADD	$\mathbf{0.038 \pm 0.015}$	$\mathbf{74.449 \pm 6.013}$	$\mathbf{-0.008 \pm 0.0038}$	$\mathbf{5.465 \pm 0.377}$
DropFuse	0.186 ± 0.0120	111.775 ± 8.028	0.186 ± 0.0120	71.058 ± 5.335
cGAN	0.310 ± 0.0117	129.022 ± 5.633	—	—
Autoencoder	—	—	0.154 ± 0.0156	52.819 ± 3.526

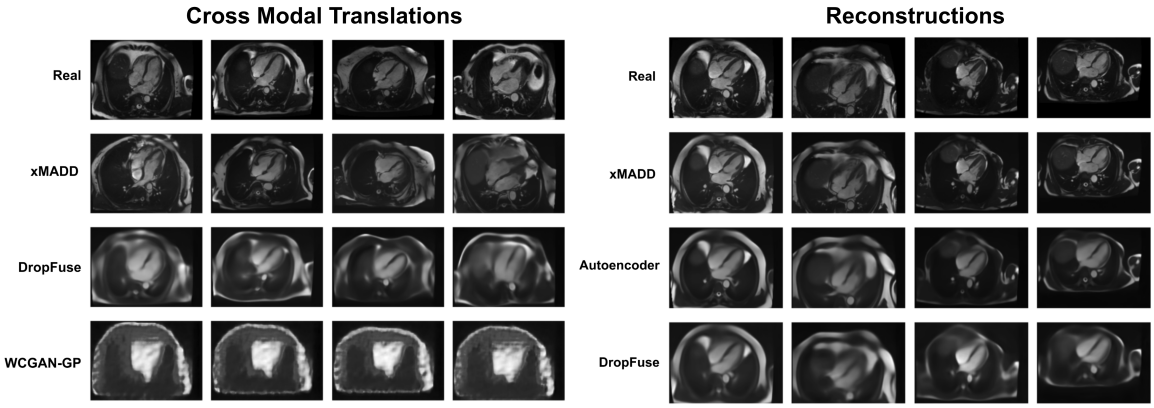

Figure 4: Visual comparison of different generative model architectures for modality translation and reconstruction. Note how the diffusion-based model (xMADD) has higher resolution and fidelity to the real data than adversarial (WCGAN-GP), self-supervised (DropFuse), and unsupervised (Autoencoder) architectures.

ization. The generated waveforms show progressively longer QT intervals, with the expected widening between QRS peak and T wave, while other waveform features remain stable. Importantly, samples generated from different random seeds under the same conditioning value exhibit diversity in secondary features while consistently preserving the controlled attribute, highlighting xMADD's ability to produce both conditionally faithful and diverse outputs.

4.3. Results: Modality Translation

Our model also enables conditional generation based on information from paired modalities. To achieve this, we leverage pre-trained foundation models to extract embeddings from the conditioning modality, which then serve as a conditioning signal for generating the other modality. The foundational encoders are previously published models for the ECG (Friedman et al., 2025), c-MRI (Radhakrishnan et al., 2023) and b-MRI

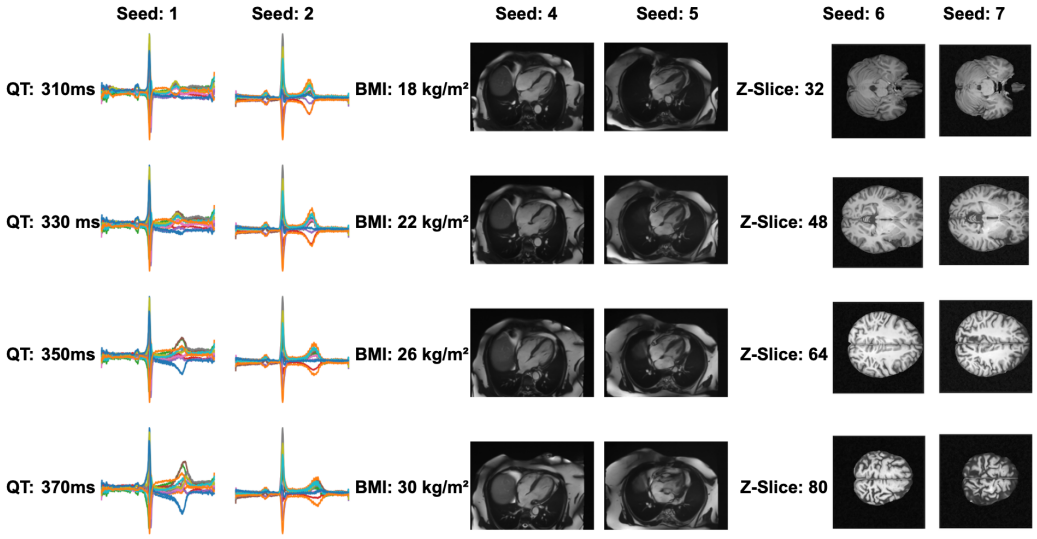

Figure 5: Phenotypic conditioning of ECGs by QT interval (left columns), c-MRI by BMI (center columns), and b-MRI by Z-slice or axial index (right columns). Along the y axis the phenotype is varied while the x-axis shows two examples with different random seeds for each modality.

Table 3: Steerability of synthetic ECG medians with different phenotype conditioning signals.

Phenotype Condition	**R/AUROC ↑ Synthetic**	**R/AUROC ↑ Real**
RR Interval	0.891	0.943
QT Interval	0.957	0.981
QRS Duration	0.869	0.886
AFib	0.734	0.730
Age	0.448	0.563

(Friedman et al., 2024) and provide cross-modal control for the diffusion models. Table 4 presents the perceptual quality from different conditioning signals with the c-MRI 4-Chamber, while Appendix Table 8 includes other cross-modal pairings. Note how all the modalities considered benefit from diverse kinds of conditioning. The reductions in KID serve as an empirical measure of the mutual information between the conditioning modality and the generated data. By this measure, the cardiovascular data modalities of different c-MRI views and ECGs have more synergistic information than c-MRIs and b-MRIs. Cardiac MRIs generated from other c-MRI views yield higher quality than those conditioned on ECG, which in turn outperform c-MRIs conditioned on b-MRI. This conforms to expectations, as ECG and cardiac MRI share more relevant information, providing a stronger conditioning signal, whereas b-MRI and c-MRI have fewer shared features, resulting in less effective guidance.

Table 4: KID results for the c-MRI 4 Chamber. As compared to the unconditioned baseline (top row), autoencoder conditioning yields the best KID (bottom row) followed by cross-modal conditions and scalar phenotype conditions.

Condition	**KID ↓**
–	0.0259
Sex	0.0203
BMI	0.0162
bMRI T1 MNI	0.0194
ECG Median	0.0176
c-MRI 4 Chamber	0.0051

4.4. Results: Utility

In previous sections, we evaluated the quality of images generated by xMADD, demonstrating that it can produce high-fidelity, realistic outputs that accurately reflect the conditioning signal. In this section, we shift our focus to assessing the utility of such generative models, specifically, their application in dataset augmentation and privacy-preserving model derivation.

4.4.1. Dataset augmentation

Dataset augmentation, where synthetic samples are used to expand the size and diversity of available training data, is particularly valuable in healthcare domains where data scarcity or class imbalance limit the utility of conventional datasets (Thambawita et al., 2022). To show this, we use the BMI-conditioned xMADD synthesis of c-MRI 4 Chamber images and generate a new dataset of 59,000 images with target BMI, sampled from a normal distribution. Training solely on these synthetic images defines the *synthetic* regime, where models are never exposed to real data. Combining the 59,000 synthetic images with up to 37,000 real images defines the *augmented* regime, while training exclusively on real images serves as the *supervised* baseline. Using the same CNN backbone across all settings, we train regression models and evaluate them on held-out real data. Table 5 reports the number of synthetic and real images in each regime and the corresponding Pearson correlation for BMI regression. Results show that models trained only on synthetic data achieve competitive performance relative to the supervised baseline, while augmenting real data with synthetic samples yields the best performance overall. This illustrates the importance of such synthetic augmentation, especially in data-scarce settings, where limited real datasets would otherwise constrain model development, as well as the trade-off between privacy and accuracy (Agarwal, 2020; Ziller et al., 2024).

Table 5: Synthetic-only training and dataset augmentation results on BMI regression from c-MRI 4 chamber view.

Regime	Synth N	Real N	R ↑
Synthetic	37k	0	.84 ± .02
Synthetic	59k	0	.88 ± .02
Augmented	37k	2k	.89 ± .01
Augmented	59k	2k	.91 ± .01
Augmented	**59k**	**37k**	**.93** ± .01
Supervised	0	2k	.82 ± .02
Supervised	0	37k	.91 ± .01

4.4.2. Privacy

Prior studies have shown that diffusion models can be prone to memorizing training samples, especially when over-trained on small or low-diversity datasets (Carlini et al., 2023; Dar et al., 2025; Somepalli et al., 2023). To evaluate this risk we adopt the membership inference test proposed in Dar et al. (2025). We compare distances from synthetic samples to real training images and to held-out test images. Figure 6 shows the mean and minimum distance of 800 synthetic images, conditioned on random BMI values, to the train and test samples. The top row shows the l_2 distance between the images, and the bottom row shows the cosine similarity between the embeddings, extracted from a pretrained c-MRI encoder. Although the minimum distance to test images is marginally higher, overall we observe no significant difference between distances to training versus test data. This suggests that membership of a specific image cannot be easily inferred, though the analysis provides only an aggregate measure. Future work should investigate privacy risks at the individual level in greater detail.

5. Conclusion

Our experiments show that the xMADD framework effectively synthesizes high-fidelity medical images and waveforms while preserving clinically relevant phenotypic information. By conditioning on both simple scalar variables and complex cross-modal embeddings, the models exhibit flexible steerability, enabling the generation of realistic, phenotype-targeted synthetic data. Limitations to this approach include the need for iterative diffusion steps which complicates deployment in resource-constrained settings, although recent one-step and accelerated diffusion methods have improved inference speed (Frans et al.;

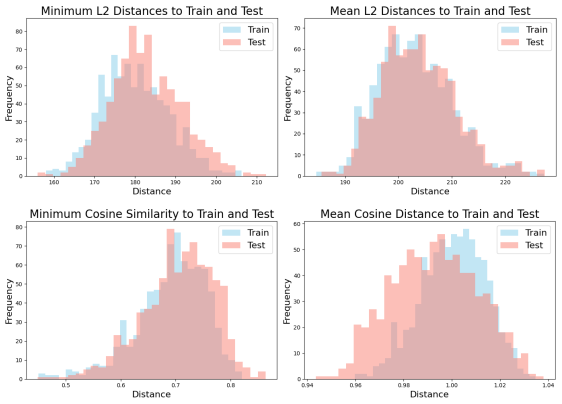

Figure 6: Synthetic to real images distance for the train and test cohort. The top row shows the Minimum and Mean l_2 distances and the bottom row the Cosine distance of the embeddings.

Huang et al., 2022). Finally, while diffusion models mitigate some privacy concerns by generating synthetic data, future work should investigate whether sensitive information is inadvertently encoded or reconstructed, underscoring the need for thorough privacy evaluations and red-teaming in medical applications. Despite these limitations, the versatile conditioning mechanism and realistic synthesis by xMADD can empower applications such as bespoke dataset generation and in silico counterfactual experimentation with high-resolution multimodal digital twins.

Acknowledgments

This work was supported in part by the Eric and Wendy Schmidt Center at the Broad Institute of MIT and Harvard.

References

Sushant Agarwal. *Trade-offs between fairness, interpretability, and privacy in machine learning*. PhD thesis, University of Waterloo, 2020.

Waqar Ahmad, Hazrat Ali, Zubair Shah, and Shoaib Azmat. A new generative adversarial network for medical images super resolution. *Scientific Reports*, 12(1):9533, 2022.

Sina Amirrajab, Yasmina Al Khalil, Cristian Lorenz, Jürgen Weese, Josien Pluim, and Marcel Breeuwer. Label-informed cardiac magnetic resonance image synthesis through conditional generative adversarial networks. *Computerized Medical Imaging and Graphics*, 101: 102123, 2022.

Héctor Anaya-Sánchez, Leopoldo Altamirano-Robles, Raquel Díaz-Hernández, and Saúl Zapotecas-Martínez. Wgan-gp for synthetic retinal image generation: Enhancing sensor-based medical imaging for classification models. *Sensors*, 25(1):167, 2024.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.

Ting Chen. On the importance of noise scheduling for diffusion models. 2023.

Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated mri. *Medical image analysis*, 80:102479, 2022.

Salman Ul Hassan Dar, Marvin Seyfarth, Isabelle Ayx, Theano Papavassiliu, Stefan O Schoenberg, Robert Malte Siepmann, Fabian Christopher Laqua, Jannik Kahmann, Norbert Frey, Bettina Baeßler, et al. Unconditional latent diffusion models memorize patient imaging data. *Nature Biomedical Engineering*, pages 1–15, 2025.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794, 2021.

Albert Einstein. Über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen. *Annalen der Physik*, 322 (8):549–560, 1905.

Chao Fan, Hao Lin, and Yingying Qiu. U-patch gan: A medical image fusion method based on gan. *Journal of Digital Imaging*, 36(1):339–355, 2023.

Jean-Baptiste Fourier. *The Analytical Theory of Heat*. Legare Street Press, 1822.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models.

Sam F Friedman, Shaan Khurshid, Rachael A Venn, Xin Wang, Nate Diamant, Paolo Di Achille, Lu-Chen Weng, Seung Hoan Choi, Christopher Reeder, James P Pirruccello, et al. Unsupervised deep learning of electrocardiograms enables scalable human disease profiling. *npj Digital Medicine*, 8(1):23, 2025.

Sam Freesun Friedman, Gemma Elyse Moran, Marianne Rakic, and Anthony Phillipakis. Genetic architectures of medical images revealed by registration of multiple modalities. *Bioinformatics and Biology Insights*, 18: 11779322241282489, 2024.

Zijian Gao, Dong Chen, and Yiqing Shen. A missing multimodal imputation diffusion model for 2d x-ray and 3d ct in covid-19 diagnosis. *Expert Systems with Applications*, 279: 127367, 2025.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Zhen Guo, Yu Wang, Xiao Huang, Yu Li, Lin Zhang, and Jing Zhang. Diffusion models in bioinformatics and computational biology. *Nature Reviews Bioengineering*, 2:136–154, 2024.

Mohammad Hamghalam and Amber L Simpson. Medical image synthesis via conditional gans: Application to segmenting brain tumours. *Computers in Biology and Medicine*, 170:107982, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

Emiel Hoogeboom, Thomas Mensink, Jonathan Heek, Kay Lamerigts, Ruiqi Gao, and Tim Salimans. Simpler diffusion (sid2): 1.5 fid on imagenet512 with pixel-space diffusion. *arXiv preprint arXiv:2410.19324*, 2024.

Dewei Hu, Yuankai K Tao, and Ipek Oguz. Unsupervised denoising of retinal oct with diffusion probabilistic model. In *Medical Imaging 2022: Image Processing*, volume 12032, pages 25–34. SPIE, 2022.

R Huang, MWY Lam, J Wang, D Su, D Yu, Y Ren, and Z Zhao. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. pages 4157–4163, 2022.

Ali Kazerouni, Amir Azizi, and Halimeh Tajmirriahi. Diffusion models in medical imaging: A comprehensive survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–10, 2022.

Diederik P Kingma and Ruiqi Gao. Understanding the diffusion objective as a weighted integral of elbos. *arXiv preprint arXiv:2303.00848*, 2, 2023.

Jennifer Lee et al. A conditional generative model for clinical ecg synthesis. *IEEE Transactions on Biomedical Engineering*, 68(10):3109–3118, 2021.

Yunxiang Li, Hua-Chieh Shao, Xiao Liang, Liyuan Chen, Ruiqi Li, Steve Jiang, Jing

Wang, and You Zhang. Zero-shot medical image translation via frequency-guided diffusion models. *IEEE transactions on medical imaging*, 2023.

Thomas J Littlejohns, Jo Holliday, Lorna M Gibson, Steve Garratt, Niels Oesingmann, Fidel Alfaro-Almagro, Jimmy D Bell, Chris Boultwood, Rory Collins, Megan C Conroy, et al. The uk biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature communications*, 11(1):2624, 2020.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. In *arXiv preprint arXiv:1411.1784*, 2014.

Pedro Morão, Joao Santinha, Yasna Forghani, Nuno Loução, Pedro Gouveia, and Mario AT Figueiredo. Counterfactual mri data augmentation using conditional denoising diffusion generative models. *arXiv preprint arXiv:2410.23835*, 2024.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

Weili Nie and Ankit B Patel. Towards a better understanding and regularization of gan training dynamics. In *Uncertainty in Artificial Intelligence*, pages 281–291. PMLR, 2020.

Sara Ometto, Soumick Chatterjee, Andrea Mario Vergani, Arianna Landini, Sodbo Sharapov, Edoardo Giacopuzzi, Alessia Visconti, Emanuele Bianchi, Federica Santonastaso, Emanuel M Soda, et al. Hundreds of cardiac mri traits derived using 3d diffusion autoencoders share a common genetic architecture. *medRxiv*, 2024.

Aldo Perez et al. Ecgnet: Deep network for arrhythmia classification. *BioMedical Engineering OnLine*, 17(1):1–12, 2018a.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018b.

Mehran Pesteie, Purang Abolmaesumi, and Robert N Rohling. Adaptive augmentation of medical data using independently conditional variational auto-encoders. *IEEE transactions on medical imaging*, 38(12):2807–2820, 2019.

Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, pages 117–126. Springer, 2022.

Sebastian Pinsler et al. Multimodal medical image synthesis with diffusion models. In *Medical Image Computing and Computer-Assisted Intervention*, 2022.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10619–10629, 2022.

Adityanarayanan Radhakrishnan, Sam F Friedman, Shaan Khurshid, Kenney Ng, Puneet Batra, Steven A Lubitz, Anthony A Philippakis, and Caroline Uhler. Cross-modal autoencoder framework learns holistic representations of cardiovascular state. *Nature Communications*, 14(1):2436, 2023.

Aimon Rahman, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, and Vishal M Patel. Ambiguous medical image segmentation using diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11536–11546. IEEE Computer Society, 2023.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference*

on machine learning, pages 8821–8831. Pmlr, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494, 2022.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations (ICLR)*, 2021b.

Arpita Srivastava, Ditipriya Sinha, and Vikash Kumar. Wcgan-gp based synthetic attack data generation with ga based feature selection for ids. *Computers & Security*, 134:103432, 2023.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in neural information processing systems*, 34:24804–24816, 2021.

Vajira Thambawita, Pegah Salehi, Sajad Amouei Sheshkal, Steven A Hicks, Hugo L Hammer, Sravanthi Parasa, Thomas de Lange, Pål Halvorsen, and Michael A Riegler. Singan-seg: Synthetic training data generation for medical image segmentation. *PloS one*, 17(5):e0267976, 2022.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

James L. Watson, David Juergens, Nathaniel Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, William Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7974):1089–1100, 2023.

McKell Woodland, Austin Castelo, Mais Al Taie, Jessica Albuquerque Marques Silva, Mohamed Eltaher, Frank Mohn, Alexander Shieh, Suprateek Kundu, Joshua P Yung, Ankit B Patel, et al. Feature extraction for generative medical imaging evaluation: New evidence against an evolving trend. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 87–97. Springer, 2024.

Yan Xia, Nishant Ravikumar, John P Greenwood, Stefan Neubauer, Steffen E Petersen, and Alejandro F Frangi. Super-resolution of cardiac mr cine imaging using conditional gans and unsupervised transfer learning. *Medical Image Analysis*, 71:102037, 2021.

Haowei Yang, Yuxiang Hu, Shuyao He, Ting Xu, Jiajie Yuan, and Xingxin Gu. Applying conditional generative adversarial networks for imaging diagnosis. In *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, pages 1717–1722. IEEE, 2024.

Zijun Zhang. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th*

international symposium on quality of service (IWQoS), pages 1–2. Ieee, 2018.

Tao Zhou, Qi Li, Huiling Lu, Qianru Cheng, and Xiangxiang Zhang. Gan review: Models and medical image fusion applications. *Information Fusion*, 91:134–148, 2023.

Alexander Ziller, Tamara T Mueller, Simon Stieger, Leonhard F Feiner, Johannes Brandt, Rickmer Braren, Daniel Rueckert, and Georgios Kaissis. Reconciling privacy and accuracy in ai for medical imaging. *Nature Machine Intelligence*, 6(7):764–774, 2024.

Appendix A. Architecture Comparison

The generative model architectures which we compare are shown in Figure 7 with their respective loss functions. Wherever possible, architectural hyperparameters were kept consistent across modeling paradigms, such that all the encoders use the same number of residually connected convolutional blocks, the same activation function, and the same width.

A.1. Generative Adversarial Models

Using the same U-Net backbone as in xMADD we train a conditional GAN to minimize the Wasserstein loss (Arjovsky et al., 2017). In comparison to the classical GAN which used Jensen-Shannon divergence (Goodfellow et al., 2020), the Wasserstein loss with gradient penalty is more stable to optimize and less susceptible to mode collapse (Anaya-Sánchez et al., 2024). The discriminator $D(x)$ emits a floating point scalar which represents the 'realness' score of an image. We then minimize the Wasserstein (earth mover's) distance between the distributions of scores for generated and real images. Simultaneously, the generator is optimized to make the realness scores from synthetic images as high as possible by minimizing the generator loss $\mathcal{L}_{Generator}$:

$$\mathcal{L}_{Discriminator} = \tag{3}$$

$$\mathbf{E}_{x\sim P_{generated}}[D(x)] - \mathbf{E}_{x\sim P_{real}}[D(x)] \tag{4}$$

$$\mathcal{L}_{Generator} = -\mathbf{E}_{x\sim P_{generated}}[D(x)] \tag{5}$$

A.2. Autoencoder and DropFuse Models

The autoencoder and DropFuse models make use of the same number of residually-connected convolutional blocks as in xMADD and WGAN-GP but without skip connections. The autoencoder is trained to minimize a reconstruction loss, the mean squared error between real and generated images. DropFuse is a self-supervised model which minimizes a multimodal contrastive loss. This loss encourages embeddings from different modalities of the same individual to be close in latent space while embeddings from different modalities from different individuals are pushed away. DropFuse also optimizes a reconstruction loss with decoders that can be used for cross-modal synthesis from the shared latent space (Radhakrishnan et al., 2023).

A.3. Training Details

All models are trained on a single NVIDIA V100 GPU with maximum batch size possible for each modality, 16 for MRIs and 64 for ECG waveforms. Models are trained with the AdamW optimizer (Zhang, 2018), initial learning rates $4e^{-5}$ and weight decay term of $1e^{-4}$. Models converged in 3-8 hours of training. Inference was performed on a single T4 GPU with 50 denoising steps which takes 3-4 seconds per MRI image and 0.5-second per ECG waveform. Generation of the entire 59K synthetic data set took 3 days on the T4 GPU with cloud cost of $5. Model training code is available at: https://github.com/broadinstitute/ml4h/

Appendix B. Hyperparameters

Experiments in Table 6 on the conditional fusion architecture show how cross-attention outperforms simple concatenation and feature-wise linear modulation (Perez et al., 2018a). Table 7 details hyperparameter experimentation for the different modalities.

Appendix C. Grokking

Diffusion models quickly learn to reconstruct input faithfully as demonstrated by steep declines in training loss, for example those shown in Figure 8. Since the diffusion U-Net contains long

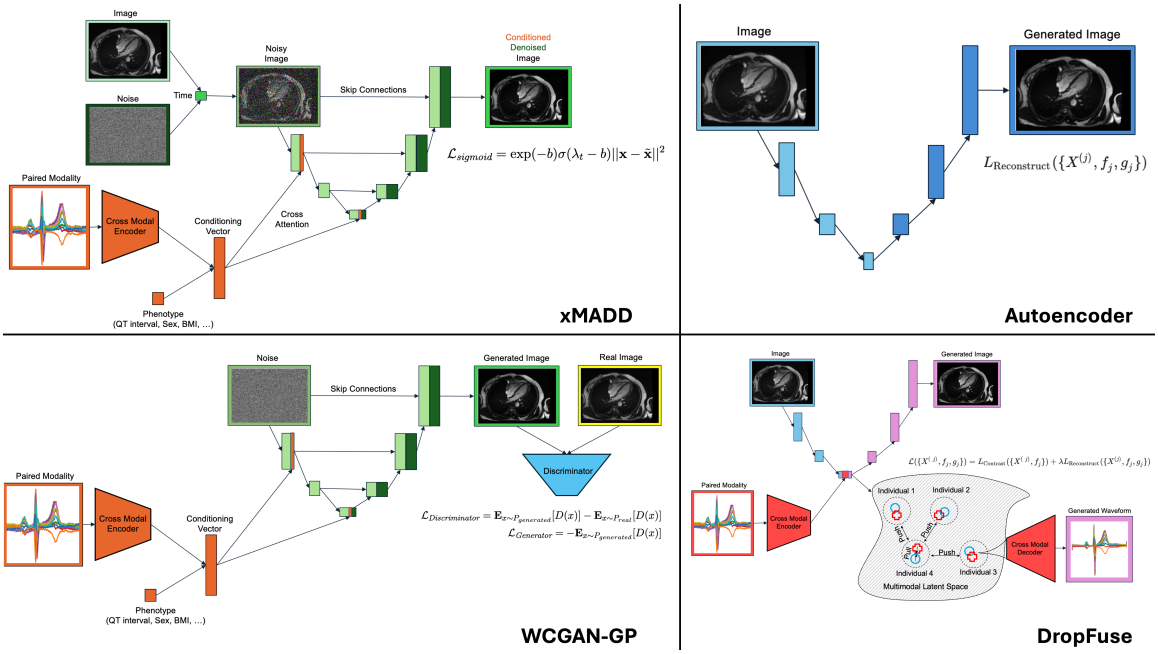

Figure 7: The four generative architectures and their respective loss functions which are compared in this paper. The models are: cross-Modal cross-Attention Denoising Diffusion (xMADD), Wasserstein Conditional Generative Adversarial Network with Gradient Penalty (WCGAN-GP), multimodal contrastive self-supervision with dropout (DropFuse), and an Autoencoder. xMADD and WCGAN-GP use identical convolutional U-Net backbones with skip connections, while DropFuse and the autoencoders use the same number and size of convolutional kernels and downsample to a 256-dimensional bottleneck without skip connections.

Table 6: Conditional vector fusion and loss function comparison for c-MRI and ECG.

Modality	**Condition**	**Blocks**	**Fusion**	**Loss**	**KID** ↓	**auROC** ↑	**R** ↑
c-MRI 4 Chamber	Sex	128x6	Cross-Attention	Sigmoid	.0203	.963	–
c-MRI 4 Chamber	Sex	128x6	Cross-Attention	MAE	.0307	.955	–
c-MRI 4 Chamber	Sex	128x6	Concat	Sigmoid	.1269	.870	–
c-MRI 4 Chamber	Sex	128x6	FiLM	Sigmoid	.1616	.836	–
ECG Median	QT Interval	64x7	Cross-Attention	MAE	–	–	.957
ECG Median	QT Interval	32x7	Concat	MAE	–	–	.914
ECG Median	QT Interval	32x7	FiLM	MAE	–	–	.908
ECG Median	QT Interval	64x6	Cross-Attention	Sigmoid	–	–	.855

range skip connections this is a fairly trivial task, and does not reflect the model truly learning the true data distribution. Rather a grokking phenomenon is observed where the validation loss consistently remains high early in the training and only begins to descend after several epochs of optimization (Power et al., 2022). Figure 9 shows synthetic examples during model training. Only after validation loss starts to fall do the examples become plausible.

Table 7: Comparison of different model configurations and their performance metrics.

Modality	Conditional Signal	Conv Kernel	Dense Blocks	Condition Vector	# Params	Learning Rate	Reconst. MSE	Downstream Eval
MRI 4ch heart	BMI	3	64x6	64	2.3M	5.00E-05	0.0818	0.684
MRI 4ch heart	BMI	3	64x6	64	2.3M	5.00E-04	0.0578	0.718
MRI 4ch heart	BMI	3	128x6	128	9M	5.00E-04	0.0700	0.857
MRI 4ch heart	Sex	3	128x6	2	9M	5.00E-04	0.9130	0.952
MRI 4ch heart	Sex	3	64x6	2	2.2M	5.00E-04	0.9090	0.923
MRI 4ch heart	ECG Latent Space	3	128x6	128	13.7M	5.00E-04	0.0651	-
Brain T1 MNI	Axial Index	3	128x6	64	9M	5.00E-04	0.0452	0.962
Brain T1 MNI	Axial Index	3	64x7	64	2.7M	5.00E-04	0.0475	0.861
ECG Median	QT Interval	11	64x5	64	2.2M	5.00E-04	0.0792	0.857
ECG Median	QT Interval	7	64x7	64	2.1M	5.00E-04	0.0717	0.906
ECG Median	QT Interval	17	64x7	64	4.8M	5.00E-05	0.0776	0.784
ECG Median	QT Interval	11	64x7	64	3.2M	5.00E-04	0.0745	0.902
ECG Median	QT Interval	15	64x7	16	8.8M	1.00E-03	0.0844	0.762
ECG Median	QT Interval	11	128x7	128	12.7M	5.00E-04	0.0713	0.918
ECG Median	RR Interval	9	256x7	64	40M	5.00E-04	0.0818	0.891
ECG Median	RR Interval	9	256x7	64	40M	5.00E-04	0.0908	0.881

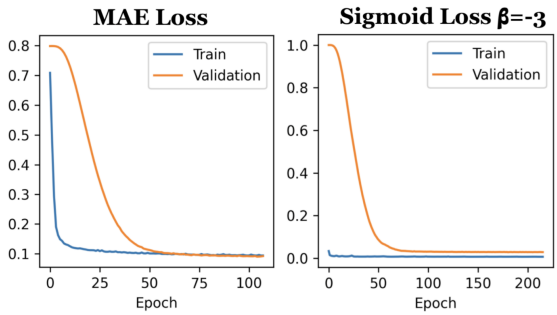

Figure 8: Diffusion models optimization consistently demonstrates grokking across different modalities and loss functions

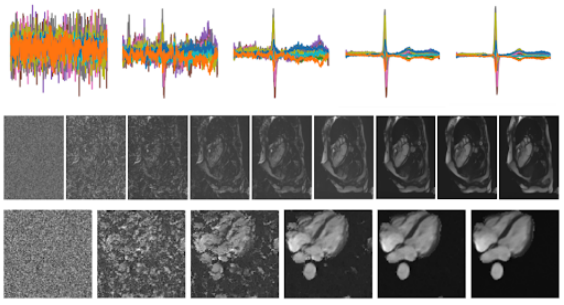

Figure 9: Synthesis during training for ECG and c-MRI.

Appendix D. Fidelity

Conditioning consistently improves perceptual quality across many modalities and conditions considered. Table 8 shows KID improvements with different types of conditioning. Fidelity to the conditional signals is shown in Table 9.

Table 8: Across all modalities considered perceptual quality improves by adding diverse conditioning signals.

Modality	**Condition**	**KID ↓**
c-MRI 2 Chamber	–	0.0962
c-MRI 2 Chamber	ECG Median	0.0795
c-MRI 2 Chamber	c-MRI 2 Chamber	0.0504
b-MRI T1 MNI	–	0.3062
b-MRI T1 MNI	Axial Index	0.2536
b-MRI T1 MNI	b-MRI T2 Flair	0.2504

Table 9: Fidelity to phenotypic conditioning.

Modality	**Phenotype Condition**	**R/AUROC Synthetic**	**R/AUROC Real**
b-MRI T1 MNI	Axial Index	0.963	0.996
b-MRI T2 Flair	Axial Index	0.757	0.816
c-MRI 4 Chamber	BMI	0.857	0.916
c-MRI 4 Chamber	Sex	0.952	0.999
c-MRI 3 Chamber	BMI	0.839	0.895
c-MRI 3 Chamber	Sex	0.928	0.999

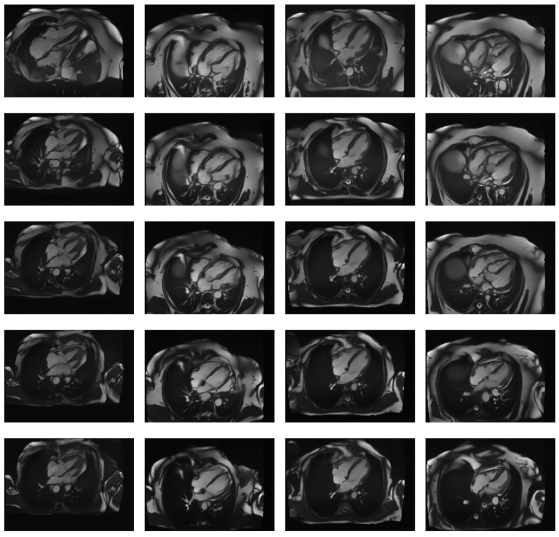

Figure 10: c-MRI 4 Chamber BMI interpolation via post-training conditioning.

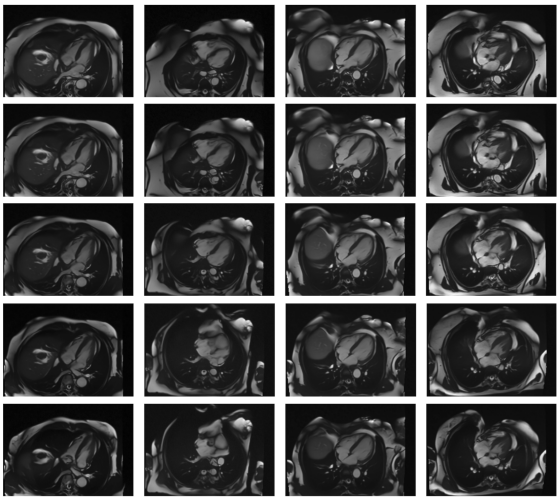

Figure 11: c-MRI 4 Chamber BMI interpolation via phenotype conditioning.

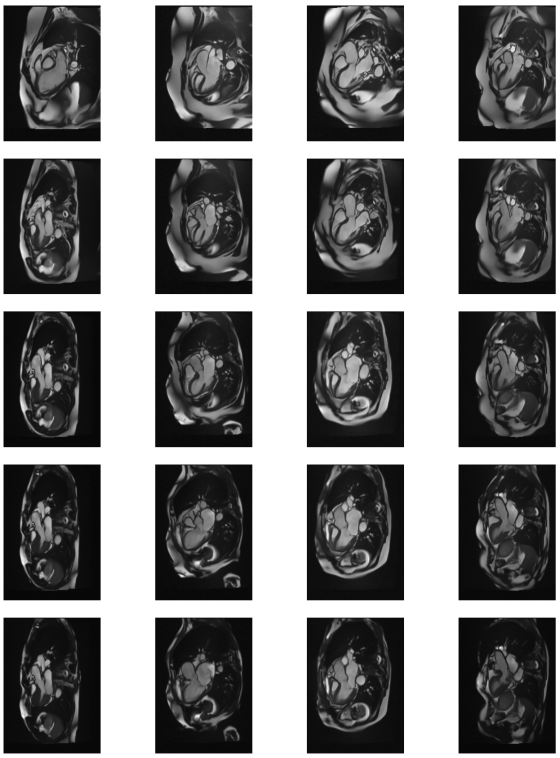

Figure 12: c-MRI 3 Chamber BMI interpolation via post-training conditioning.

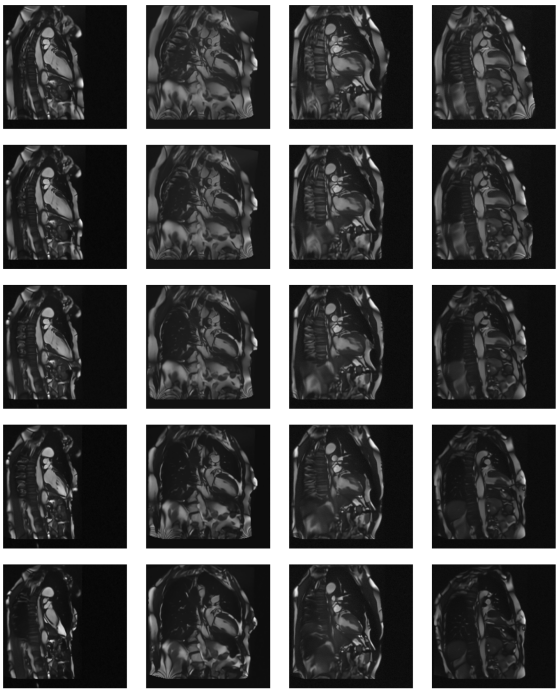

Figure 13: c-MRI 2 Chamber BMI interpolation via post-training conditioning.

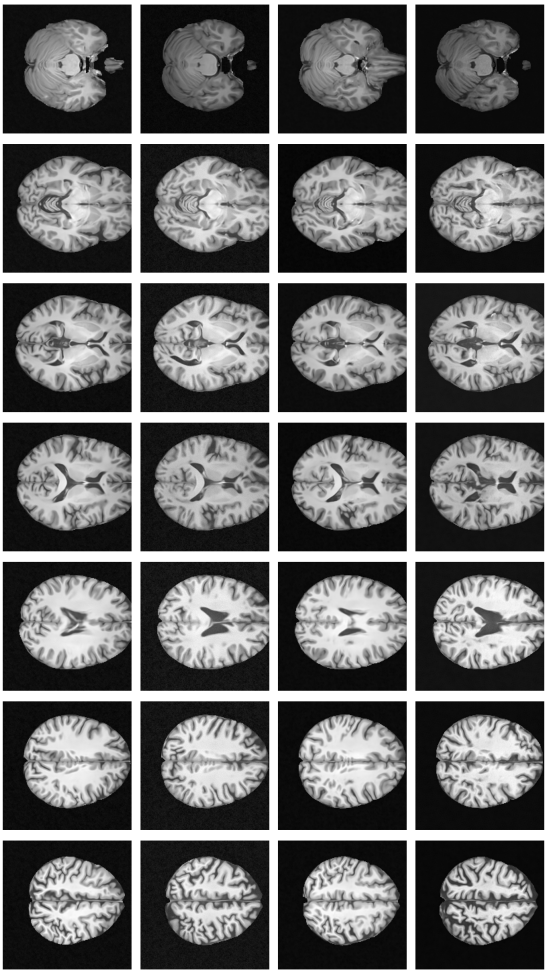

Figure 14: b-MRI T1 MNI axial index (Z-slice) interpolation via phenotype conditioning

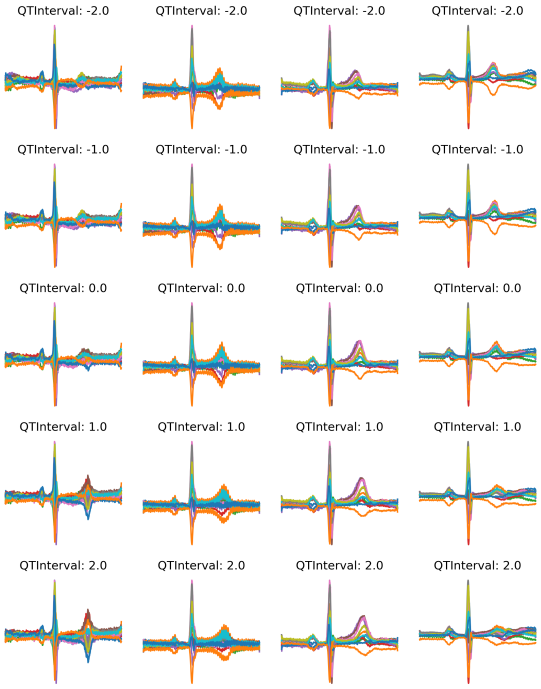

Figure 15: ECG Median QT Interval interpolation via phenotype conditioning