SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Making DBT Accessible Without a DBT Scanner: A Generative AI Approach to Synthesizing Pseudo-Tomosynthesis from Standard Mammograms

Ashok Kumar Sahoo[1], Rajkumar Rajavel[2]

[1]Research Scholar, School of Computer Science and Engineering, Christ University, Bangalore, India.
E-mail: ashok.sahoo123@gmail.com
[2]Associate Professor, Department of AI, ML and Data Science, Christ University, Bangalore, India. E-mail: rajkumarprt@gmail.com
*Corresponding Author: Ashok Kumar Sahoo, E-mail: ashok.sahoo123@gmail.com

Abstract
This study presents MammoTomo, a physics-constrained generative AI framework that synthesizes pseudo-digital breast tomosynthesis (pseudo-DBT) volumes from standard 2D mammograms, addressing the hardware barrier restricting tomosynthesis access in resource-constrained settings. It uses a pseudo-3D latent diffusion model with depth attention for inter-slice coherence and cross-attention with mammogram features, denoising all slices jointly conditioned on the input image. A differentiable projection consistency loss, derived from the Beer-Lambert attenuation model, enforces that the synthesized volume reproduces the input mammogram under forward projection, and lesion-aware attention injection preserves diagnostically critical regions. Benchmarked on T-SYNTH (9,000 paired DM/DBT images), MammoTomo improves image quality and projection fidelity over GAN and independent-slice diffusion baselines, training a DBT lesion detector to near-real-DBT performance (AUC = 0.821 vs. oracle 0.849). Ablations confirm depth attention, physics consistency, and lesion guidance contribute complementary gains, reaching 96.7% of oracle performance. The findings establish a foundation for low-resource, scanner-free breast cancer screening.

Keywords: DBT (Digital Breast Tomosynthesis); pseudo-DBT synthesis; generative AI; latent diffusion model; mammography; cross-modality synthesis; physics-constrained learning; Beer-Lambert projection; depth attention; lesion-aware synthesis; medical image generation; breast cancer screening.

1. Introduction

Breast cancer is the most commonly diagnosed cancer among women worldwide, with 2.30 million new cases and 666,000 deaths recorded globally in 2022, as per Bray et al. (2024); women in lower-resource settings are diagnosed less often but die at disproportionately higher rates, accounting for roughly 70% of the global mortality burden despite lower incidence (Bray et al., 2024). Screening mammography reduces mortality, but standard full-field digital mammography (FFDM) projects three-dimensional breast tissue onto a two-dimensional detector, causing tissue superposition, the overlap of fibroglandular structures that obscures malignancies, and inflates recall rates. Digital breast tomosynthesis (DBT) addresses this limitation by reconstructing a pseudo-3D breast volume from low-dose angular projections; a large randomized trial reported improved invasive cancer detection with DBT over FFDM, as demonstrated by Friedewald et al. (2014), and an independent multicentre trial confirmed this benefit, as per Heindel et al. (2022), while a large ongoing US trial continues to evaluate it prospectively in the TMIST trial (ECOG-ACRIN Cancer Research Group, 2025). The evidence is not unanimous, however: a comparably sized Norwegian randomized trial found no significant detection benefit, as reported by Hofvind et al. (2019), and DBT screening carries a documented cost premium over FFDM even where infrastructure already exists, as per Moger et al. (2019). This mixed but generally favorable evidence, combined with DBT's dedicated-hardware requirement, leaves the modality largely unavailable in the resource-constrained settings where the mortality burden above is concentrated. This access gap motivates a central question: can the diagnostic content of a DBT volume be synthesized from a standard 2D mammogram using generative AI, extending tomosynthesis benefits without new hardware?

Generative AI approaches, particularly denoising diffusion probabilistic models (DDPMs) and latent diffusion models (LDMs), replace the adversarial min-max objective of GANs with an iterative denoising formulation, yielding more stable training and higher sample fidelity across medical image synthesis tasks, as

demonstrated by Ho et al. (2020). In breast imaging, diffusion models have been applied to 2D mammogram synthesis, as per Montoya-del-Angel et al. (2024), and, in the reverse direction, to generating synthetic 2D projections from DBT volumes (Jiang et al., 2019, 2023). However, synthesizing a complete pseudo-DBT volume from a single 2D mammogram remains largely unexplored and presents three fundamental challenges. The synthesis is inherently ill-posed, since the 3D-to-2D projection is many-to-one: infinitely many volumes could explain a given mammogram, leaving an unconstrained generator with no principled way to select among them. Compounding this, generating a coherent volume of 60–80 slices requires maintaining inter-slice consistency across depth, since slices synthesized independently lack shared context and tend to produce visible depth-wise discontinuities. A further, clinically decisive requirement is that critical findings be preserved rather than hallucinated or suppressed; a volume that is visually plausible but diagnostically unfaithful would be worse than no synthesis at all.

In this research, we propose MammoTomo, a physics-constrained latent diffusion model for pseudo-DBT synthesis, introducing three principal contributions. First, a differentiable X-ray projection consistency loss derived from the Beer-Lambert attenuation model is introduced (equations (3) and (4)), enforcing that the synthesized volume reproduces the input mammogram under physical forward projection. This converts the ill-posed synthesis problem into a constrained one, using imaging physics as a free self-supervisory signal. Second, a pseudo-3D denoising U-Net is designed in which all slices of the pseudo-DBT volume are generated jointly in a single DDIM loop, with depth attention enabling inter-slice coherence without prohibitive memory cost. Third, a lesion-aware attention injection mechanism is proposed (equation (5)) that injects predictions from a pretrained lesion detector into the denoising network's cross-attention layers, guiding synthesis toward diagnostically critical regions. Training and evaluation are conducted exclusively on T-SYNTH (Wiedeman et al., 2025), a publicly available dataset of 9,000 paired synthetic DM/DBT images with pixel-level lesion annotations, ensuring full reproducibility.

The remainder of this paper is organized as follows: Section 2 surveys related work on cross-modality synthesis, diffusion-based generative models, and physics-informed constraints; Section 3 details the MammoTomo architecture and two-phase training strategy; Section 4 presents the experimental setup, baselines, and quantitative results; and Section 5 concludes with a discussion of principal limitations and future directions.

2. Related Work

Prior work relevant to this study spans DBT's clinical basis, cross-modality image synthesis, diffusion-based generative modeling, and physics- and lesion-guided constraints. Across these areas, no existing method has attempted to synthesize a complete pseudo-DBT volume from a single 2D mammogram at the image level; the subsections below establish this gap directly and identify the specific techniques, drawn from adjacent domains, that MammoTomo adapts to close it.

2.1 Digital Breast Tomosynthesis and Its Clinical Limitations

Digital breast tomosynthesis (DBT) is the imaging modality most directly targeted at resolving tissue superposition in standard full-field digital mammography (FFDM), yet its clinical adoption remains constrained more by cost and hardware availability than by diagnostic performance. DBT acquires low-dose projections at multiple angles and reconstructs a pseudo-3D breast volume, demonstrating improved sensitivity and specificity in large-scale studies, as highlighted by Friedewald et al. (2014). While these clinical advantages are established, adoption remains globally uneven: DBT systems cost considerably more than FFDM units, require dedicated hardware, and deliver a higher radiation dose, leaving most low- and middle-income healthcare settings reliant exclusively on FFDM. This access disparity motivates the present work: synthesizing pseudo-DBT volumes from existing FFDM images can extend tomosynthesis benefits to resource-constrained settings without hardware upgrades.

2.2 Cross-Modality Synthesis in Breast Imaging

Cross-modality image synthesis can be broadly categorized into GAN-based and diffusion-based approaches. Early work employed conditional GANs (cGANs) for MR-to-CT and PET-to-MRI translation, as demonstrated by Nie et al. (2017) and Ben-Cohen et al. (2019). In breast imaging, synthesis has been studied primarily in the reverse direction, generating synthetic 2D mammograms from DBT volumes to reduce radiation exposure. Gradient-guided cGANs improve high-frequency content in synthesized images, as per Jiang et al. (2019), while multi-scale cascaded networks, as introduced in subsequent work by Jiang et al. (2023), address intensity distortion from FFDM-DBT domain differences. A ResNet-based cross-modality feature synthesis (CMFS) model, as proposed by Fan et al. (2025), generates DBT-like features from FFDM images. However, all prior methods target feature-space or 2D synthesis; the present work departs by targeting full pseudo-DBT volume synthesis at the image level, applying a physics-based consistency constraint absent from prior approaches.

2.3 Diffusion Models for Medical Image Synthesis

Denoising diffusion probabilistic models (DDPMs) and their latent variants (LDMs) avoid the adversarial training instability and mode collapse associated with GANs by learning a stable iterative denoising process, establishing them as the dominant generative framework for medical image synthesis, as highlighted by Ho et al. (2020). In breast imaging, MAM-E applies diffusion models conditioned on text descriptions and lesion masks to generate high-quality mammographic images, as demonstrated by Montoya-del-Angel et al. (2024). In volumetric medical imaging, pseudo-3D diffusion architectures, which augment 2D spatial operations with 1D depth attention, enable coherent slice-by-slice generation without the prohibitive memory cost of fully 3D convolution stacks; this factorized spatio-temporal attention design was originally introduced by Ho et al. (2022) for video generation and has not previously been adapted to volumetric medical synthesis. To the best of our knowledge, no prior work has applied latent diffusion models to the synthesis of pseudo-DBT volumes from 2D mammograms, leaving the inter-slice consistency challenge identified above unaddressed for this task.

2.4 Physics-Informed Constraints in Generative Models

The ill-posedness of unconstrained 2D-to-3D synthesis motivates incorporating physical constraints into the generative model itself, which improves geometric fidelity and reduces hallucination. This follows the broader principle established for physics-informed neural networks, wherein governing physical equations are embedded directly into the training objective to regularize otherwise underdetermined learning problems, as established by Raissi et al. (2019). In CT reconstruction, differentiable forward projection operators have been integrated into deep learning pipelines to enforce sinogram consistency, as explored by Wuerfl et al. (2018). The mammographic X-ray imaging process is well-characterized by the Beer-Lambert law: the 2D mammogram is a log-integrated projection of the 3D attenuation volume. This deterministic relationship provides a free and differentiable supervisory signal: any synthesized pseudo-DBT volume must reproduce the input mammogram under projection. To the best of our knowledge, the present work is the first to exploit this physical relationship as a training constraint for breast tomosynthesis synthesis, directly addressing the ill-posedness challenge identified above by regularizing the space of plausible volumes.

2.5 Lesion-Aware Synthesis and Guided Diffusion

Preserving diagnostic content, the third challenge identified above, is a critical requirement for clinically useful synthesis. GAN-based lesion synthesis has been shown to improve diagnostic classification accuracy in dermatology, demonstrating that generative models can preserve clinically salient lesion structure well enough to benefit downstream diagnosis, as per Qin et al. (2020). In diffusion models, spatial guidance mechanisms such as ControlNet, as introduced by Zhang et al. (2023), inject conditioning signals into cross-attention layers without full model retraining. T-SYNTH, a recently released dataset of 9,000 paired synthetic DM and DBT images generated via Monte Carlo X-ray simulation, provides pixel-level lesion masks for both modalities (Wiedeman et al., 2025). The present work exploits these annotations to train a lesion attention predictor whose output is injected into the denoising network's cross-attention layers via a ControlNet-style bias, yielding lesion-aware synthesis that is, to our knowledge, novel in the DBT context.

3. Materials and Methods

This study proposes MammoTomo, a three-stage generative AI pipeline for synthesizing pseudo-DBT volumes from 2D mammograms, comprising: (i) a spatially-rich mammogram encoder, (ii) a pseudo-3D latent diffusion model with depth attention, and (iii) a physics-constrained fine-tuning stage enforced by a differentiable X-ray projection operator. An optional lesion-aware attention injection module further improves preservation of diagnostic findings. Fig. 1 illustrates the full pipeline.

3.1 Problem Formulation

Let I_2D denote a full-field digital mammogram and V the corresponding DBT volume comprising D slices (typically 60–80). The task is to learn a mapping f_θ such that the synthesized volume $\hat{V} = f_\theta(I_2D)$ satisfies geometric consistency with I_2D under the X-ray projection operator P, coherence across the depth dimension D, and preservation of lesion-level diagnostic content. The synthesis problem is fundamentally ill-posed: the 3D-to-2D projection is many-to-one, meaning infinitely many volumes could explain a given mammogram. The physics constraint and lesion guidance introduced in §3.4–§3.5 therefore serve as regularizers, selecting a plausible and clinically faithful solution from this degenerate space.

3.2 Stage 1: Spatially-Rich Mammogram Encoder

A ResNet-50 backbone pretrained on ImageNet and fine-tuned on DM images from T-SYNTH is employed as the mammogram encoder E_φ. To address the shortcomings of prior synthesis methods that pool encoder output to a global embedding vector, the full spatial feature map is retained at the final convolutional stage, yielding a multi-scale feature pyramid $\{F_2D^l\}^1{}_{=1}{}^4$ at strides {2, 4, 8, 16} relative to the input. The coarsest level (stride 16, H/16 × W/16 × 256) captures global glandular context; finer levels (strides 2–8) retain

Cooper's ligament structure and fine calcification detail that global pooling would discard. Retaining full spatial maps is critical, as it enables the denoising network to spatially attend to specific anatomical regions of I_2D at every resolution during cross-attention (§3.3.3).

3.3 Stage 2: Pseudo-3D Latent Diffusion Model

3.3.1 Slice-Level VAE

A variational autoencoder (VAE) is trained on individual DBT slices from T-SYNTH. The encoder comprises four down sampling stages with channel widths [128, 256, 512, 512], each employing 2D strided convolutions (stride 2) followed by two ResNet blocks. The decoder mirrors this architecture with nearest-neighbour upsampling. For a 512×512 input slice, the encoder produces a 4-channel latent of spatial resolution 64×64 (downsampling factor f = 8), reducing memory by 64× relative to pixel-space diffusion. The VAE is trained with a combined loss:

$$L_{VAE} = L_{recon} + \beta \cdot KL\big(q(z|x) \parallel p(z)\big) + \lambda_{perc} \cdot L_{LPIPS} \quad (1)$$

where $\beta = 1\times10^{-6}$ keeps the posterior close to the prior without over-regularizing texture detail and L_{LPIPS} provides perceptual sharpness. The VAE is trained first and frozen before diffusion training commences.

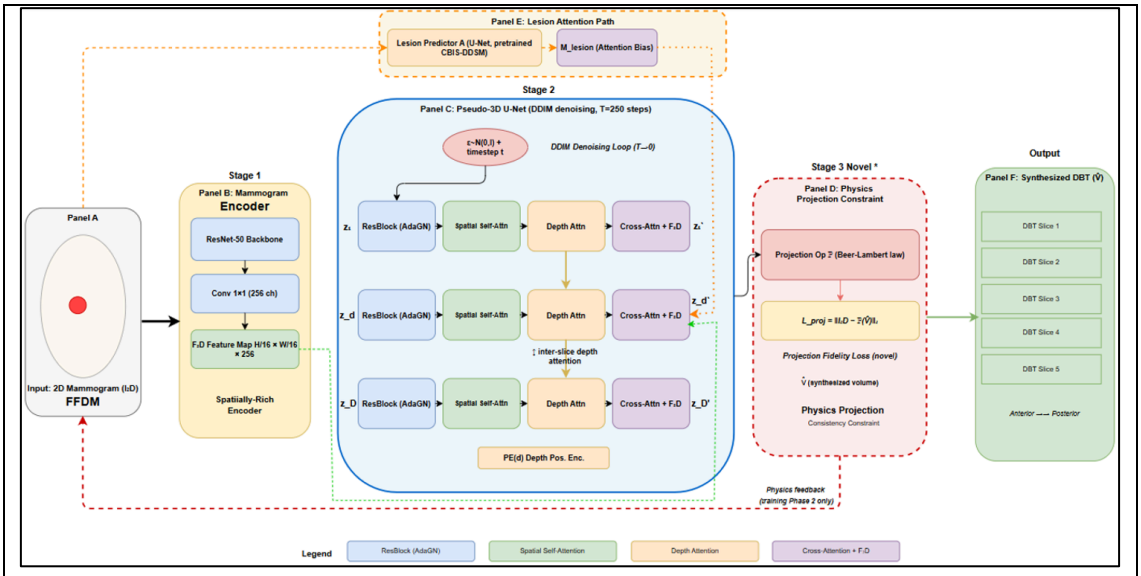

Figure 1. MammoTomo pipeline: (A) 2D mammogram input I_2D; (B) ResNet-50 encoder extracting multi-scale feature pyramid $\{F_2D^l\}$ at strides {2, 4, 8, 16}; (C) Pseudo-3D U-Net jointly denoising D slices via depth attention and cross-attention on F_2D; (D) pretrained lesion predictor injecting attention bias M_{lesion}; (E) differentiable Beer-Lambert projector enforcing L_{proj} during Phase 2; (F) synthesized pseudo-DBT volume $\hat{V}$.

3.3.2 Denoising U-Net with Depth Attention

The core of MammoTomo is a denoising U-Net ε_θ that predicts the noise added to the latent representation of slice d at diffusion timestep t, conditioned on mammogram features F_{2D} and depth position d. Each resolution block comprises four sub-modules applied in sequence: (1) a 2D ResBlock for spatial convolution within slice d; (2) a 2D spatial self-attention layer; (3) a depth attention layer in which slice d attends to its k nearest neighbors in depth (k = 5), implemented as 1D multi-head attention across the D-slice stack; and (4) a cross-attention layer injecting mammogram features F_{2D} at every resolution level (equation (2)). Depth position d is encoded as a sinusoidal positional embedding added to the latent before each resolution block. Channel widths follow [320, 640, 1280, 1280] with 8 attention heads. All D slices share weights and are denoised jointly in a single DDIM loop.

Inference procedure. At test time, D latent maps $\{z_d^T\}_{d=1}^D$ are initialized independently, each sampled from $z_d^T \sim N(0, I)$. DDIM sampling ($\eta = 0$, deterministic) is then run for T = 250 denoising steps, processing all D slices jointly at every step to leverage depth attention across the volume. Conditioning is provided entirely through cross-attention with $F_{2D} = \text{Encoder}(I_{2D})$ and the lesion-attention bias; no classifier-free guidance is applied. After the final denoising step, each z_d^0 is decoded by the frozen VAE to produce the synthesized volume $\hat{V} = \{\hat{v}_d\}_{d=1}^D$. Synthesis of a complete 72-slice volume requires approximately 18 seconds on a single A100 GPU.

3.3.3 Cross-Attention with Mammogram Features

The mammogram encoder (§3.2) produces a multi-scale feature pyramid $\{F_{2D}^{l}\}_{l=1}^{4}$ at strides {2, 4, 8, 16} relative to the input. At U-Net resolution level l with spatial size $H_l \times W_l$, F_{2D}^{l} is bilinearly resized to $H_l \times W_l$ so that each spatial position in the key-value sequence corresponds directly to the anatomically equivalent location in the mammogram. The cross-attention sub-module then projects the $(H_l \times W_l)$-flattened slice latent to queries Q and the resized mammogram features to keys K and values V:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

This spatial alignment ensures pixel-level correspondence between mammogram anatomy and the denoised slice latent at every scale, propagating fine-grained structural information—e.g., Cooper's ligaments, microcalcification clusters—into the synthesized volume. Without bilinear alignment, features from different anatomical regions would be incorrectly mixed during cross-attention, degrading spatial fidelity.

3.4 Stage 3: Physics-Constrained Fine-Tuning

After Phase 1, a differentiable X-ray projection operator P is introduced. Given the decoded pseudo-DBT volume, P computes X-ray attenuation following the Beer-Lambert law:

$$\hat{I}_{proj} = P(\hat{V}) = -\log\left(\frac{1}{D}\sum_{d=1}^{D}\exp(-\mu \cdot \hat{v}_d)\right) \quad (3)$$

where μ is the mean linear attenuation coefficient of breast tissue and $\hat{v}_d$ is the d-th synthesized slice. The projection consistency loss is:

$$L_{proj} = \left\|I_{2D} - \hat{I}_{proj}\right\|_1 + \lambda_p \cdot L_{LPIPS}(I_{2D}, \hat{I}_{proj}), \quad \lambda_p = 0.1 \quad (4)$$

L_{proj} (equation (4)) is fully differentiable and backpropagated through the VAE decoder during Phase 2. Crucially, it requires no additional annotations—the physics of X-ray projection provides a free self-supervisory signal constraining synthesis to physically plausible volumes. A parallel-beam approximation is adopted, treating projection as Beer-Lambert attenuation perpendicular to the detector; scattered radiation and the focal-spot heel effect are not modelled. This approximation is standard in DBT simulation literature, as per Wiedeman et al. (2025), and adequate for the consistency loss.

3.5 Lesion-Aware Attention Injection

To preserve lesion-level diagnostic content, a lightweight lesion attention predictor A is trained on CBIS-DDSM (Lee et al., 2017) using binary cross-entropy with lesion segmentation masks. CBIS-DDSM (real clinical mammograms) is deliberately employed in place of T-SYNTH annotations, to ensure the predictor generalizes to real-world lesion appearances, improving robustness of the attention bias when deployed on clinical images beyond the simulated domain. At inference time, $M_{lesion} = A(I_{2D})$ is resized to each resolution level and injected into the cross-attention layers (equation (2)) as an additive attention bias:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + \alpha \cdot \log(M_{lesion} + \epsilon)\right)V \quad (5)$$

where α is a learnable scalar and $\varepsilon = 1\times10^{-6}$ prevents log(0). This mechanism is analogous to the attention bias in ControlNet (Zhang et al., 2023), applied at the spatial level of the mammogram's lesion map. The predictor is frozen during diffusion training. The lesion attention bias M_{lesion} (equation (5)) is active in the forward pass of both training phases; however, the supervision term L_{lesion} is introduced only in Phase 2. Phase 1 therefore trains the base denoising objective while already receiving spatial lesion guidance through the frozen predictor's output.

3.6 Two-Phase Training Strategy

Training proceeds in two phases. The VAE is pre-trained and frozen before either phase commences (§3.3.1). Phase 1 then trains the encoder and denoising U-Net with the standard denoising score-matching objective:

$$L_{diffusion} = \mathbb{E}_{z,\epsilon,t}[\|\epsilon - \epsilon_\theta(z_d^t, t, d, F_{2D})\|^2] \quad (6)$$

Phase 2 fine-tunes the trainable components (encoder and denoising U-Net, with the VAE remaining frozen) with the combined objective:

$$L_{total} = L_{diffusion} + \lambda_1 \cdot L_{proj} + \lambda_2 \cdot L_{lesion} \quad (7)$$

where L_{lesion} is the mean squared error between lesion attention maps on synthesized and real DBT slices, and $\lambda_1 = 0.1$, $\lambda_2 = 0.05$. The two-phase strategy ensures the model first learns a stable generation distribution before physics and lesion constraints are applied; training from scratch with all losses simultaneously is empirically unstable.

3.7 Datasets and Implementation Details

Training and evaluation are conducted exclusively on publicly available data to ensure full reproducibility. The primary dataset is T-SYNTH (Wiedeman et al., 2025): 9,000 paired synthetic DM and DBT images generated via Monte Carlo X-ray simulation across four breast density categories with pixel-level lesion masks, split 7,200/900/900 train/validation/test. The lesion attention predictor is pretrained on CBIS-DDSM and frozen before diffusion training. All models are implemented in PyTorch. The denoising U-Net employs 4 resolution levels with channel dimensions [320, 640, 1280, 1280]. Phase 1 trains for 200 epochs with the Adam optimizer (learning rate = 1×10^{-4}); Phase 2 fine-tunes for 50 additional epochs (learning rate = 5×10^{-5}). A batch size of 8 volumes per GPU is used across 4 A100 GPUs with gradient checkpointing.

4. Results and Discussion

4.1 Evaluation Framework

MammoTomo is assessed across three evaluation tiers of increasing clinical relevance.

Tier 1 — Slice-Level Image Quality: Structural similarity index measure (SSIM), peak signal-to-noise ratio (PSNR), and Fréchet Inception Distance (FID) are computed between synthesized and ground-truth DBT slices. FID captures distributional realism rather than per-pair alignment—appropriate given the ill-posed nature of the synthesis task.

Tier 2 — Projection Fidelity (Novel Metric): Projection fidelity (PF) is introduced as a novel metric: given a synthesized volume $\hat{V}$, the forward projection operator P is applied and the L1 distance to the input mammogram is measured:

$$\mathrm{PF} = \left\| \mathrm{I}_{2D} - \mathrm{P}\left(\hat{\mathrm{V}}\right) \right\|_1 \quad (8)$$

As PF (equation (8)) employs the same parallel-beam approximation as L_{proj} (equation (4)), it quantifies internal physics consistency rather than accuracy under an independent cone-beam model; to the best of our knowledge, no prior synthesis paper reports this metric.

Tier 3 — Lesion-Level Diagnostic Utility: A DBT lesion detector (RetinaNet, ResNet-50 backbone) is trained separately on real T-SYNTH DBT volumes, synthesized pseudo-DBT volumes, and a combined set; detection AUC on held-out real DBT volumes serves as the ultimate clinical test of synthesis quality.

4.2 Baselines

Five baselines are compared against MammoTomo, all trained on the same T-SYNTH split with identical data augmentation. cGAN (Jiang et al. (2019)) employs gradient guidance for 2D-to-slice synthesis, operating slice-by-slice without depth coherence. CycleGAN performs unpaired 2D-to-slice translation—its weakest performance (SSIM = 0.572) confirms the value of paired supervision available in T-SYNTH. LDM (independent) generates each DBT slice independently via a standard latent diffusion model, without cross-slice depth attention. LDM + depth attn adds the pseudo-3D depth attention module to LDM (independent) while omitting the physics constraint and lesion injection, isolating those two contributions. CMFS (Fan et al. (2025)) performs cross-modality feature synthesis in feature space rather than image space, providing a non-generative reference; its published lightweight decoder head is used to render synthesized features into slice-level images for Tier 1 and Tier 3 comparison, since that decoder is not derived from a physical projection model. A structural comparison of all methods is presented in Table 1.

4.3 Ablation Studies

An additive ablation is designed starting from a base LDM (Variant A), progressively incorporating each novel MammoTomo component. Variant B introduces pseudo-3D depth attention, testing inter-slice coherence in isolation; Variant C further adds the Beer-Lambert physics constraint (equation (4)), isolating its impact on projection fidelity (equation (8)); Variant D instead adds lesion guidance atop Variant B—omitting the physics constraint—to isolate its independent contribution to diagnostic utility. Variant E is the full MammoTomo with all three components active. Variants A and B correspond directly to LDM (independent) and LDM + depth attn in Table 3, bridging the baseline comparison and ablation study. Variant configurations and expected impact are summarized in Table 2; quantitative results across all evaluation tiers are presented in Table 4.

Table 1. Structural comparison of MammoTomo against five baseline methods. Depth-aware denotes joint denoising of all slices across the volume; physics constraint denotes Beer-Lambert projection consistency enforced during Phase 2 training.

Baseline	Description	Synthesis Level	Depth-aware	Physics Constraint

cGAN (Jiang et al., 2019)	Gradient-guided cGAN adapted to 2D-to-slice direction	Single slice	No	No
CycleGAN	Unpaired 2D-to-slice translation; no physics constraint	Single slice	No	No
LDM (independent)	Standard LDM; slices generated independently without depth attn	Per-slice	No	No
LDM + depth attn	MammoTomo without physics constraint and lesion guidance	Volume	Yes	No
CMFS (Fan et al., 2025)	Cross-modality feature synthesis (feature space, not image space)	Feature space	No	No
MammoTomo (ours)	Full model: depth attention + physics constraint + lesion guidance	Volume	Yes	Yes

Table 2. Additive ablation design. Each variant adds one component to the previous. Variants A and B correspond to LDM (independent) and LDM + depth attn in Table 3. Quantitative results in Table 4.

Variant	Depth Attn	Physics	Lesion	Expected Impact
A: Base LDM	No	No	No	Worst coherence; baseline reference
B: + Depth Attention	Yes	No	No	Improved SSIM; coherent volumes
C: + Physics Constraint	Yes	Yes	No	Improved projection fidelity (PF metric)
D: + Lesion Guidance	Yes	No	Yes	Improved lesion AUC
E: Full MammoTomo	Yes	Yes	Yes	Best across all three tiers

4.4 Quantitative Results

Table 3 presents quantitative results on the T-SYNTH test set. MammoTomo outperforms all baselines on every metric ($p < 0.05$, paired t-test). Against LDM + depth attn, gains of +0.025 SSIM (0.743 vs. 0.718) and +0.028 Lesion AUC (0.821 vs. 0.793) are recorded; the largest absolute gains, +0.082 SSIM and +3.2 dB PSNR against LDM (independent), establish depth attention as the dominant image-quality contributor. The Projection Fidelity score (equation (8)) falls from 0.151 to 0.098 (35% reduction); Table 4 attributes the majority of this reduction to the Beer-Lambert physics constraint alone (B→C: 32%), with lesion-aware guidance contributing a further small reduction via its regularizing effect on texture quality (§4.4).

Table 4 traces the additive ablation (Variants A–E): A→B yields +0.057 SSIM and +2.0 dB PSNR; B→C reduces PF from 0.151 to 0.102; B→D raises AUC from 0.793 to 0.814. Beyond its primary effect on PF, the physics constraint (B→C) also raises SSIM (0.718→0.731), PSNR (27.9→28.4 dB), and Lesion AUC (0.793→0.799), mirroring the cross-metric spillover already noted for lesion guidance (B→D) and indicating that both constraints act as general regularizers rather than single-metric fixes. Variant E achieves the best result across all five metrics simultaneously: relative to Variant D, adding the physics constraint (D→E) further reduces PF by 0.050 (0.148→0.098) and raises AUC by 0.007 (0.814→0.821); relative to Variant C, adding lesion guidance (C→E) raises AUC by 0.022 (0.799→0.821) while further lowering PF by 0.004 (0.102→0.098). Each component therefore contributes measurably even after the other two are already present, confirming that the three components are complementary rather than redundant. Table 5 shows graceful degradation across ACR density categories, with SSIM ranging from 0.771 (category A) to 0.712 (category D).

A RetinaNet detector trained on real T-SYNTH DBT volumes achieves AUC = 0.849; MammoTomo synthesized data reaches AUC = 0.821, corresponding to 96.7% of oracle performance and closing 50% of the gap between the strongest baseline (AUC = 0.793) and real-DBT performance. Furthermore, Variant D yields secondary gains beyond its primary AUC benefit: SSIM rises from 0.718 to 0.724 and PF decreases from 0.151 to 0.148, indicating that lesion-aware attention exerts a mild regularizing effect on texture quality.
PF: Projection Fidelity (equation (8); lower is better); all MammoTomo improvements over LDM + depth attn are significant at $p < 0.05$ (paired t-test). †Real DBT oracle: RetinaNet trained on real T-SYNTH DBT volumes (AUC = 0.849); MammoTomo (AUC = 0.821) reaches 96.7% of oracle AUC (0.821/0.849) and closes 50% of the AUC gap between the strongest synthesis baseline (LDM + depth attn, AUC = 0.793) and the real-DBT oracle. ‡CMFS produces feature-space output decoded to slice-level images via its published decoder head (§4.2) for SSIM/PSNR/FID/AUC comparison; because that decoder is not derived from a physical projection model, PF is undefined for this method.

Table 3. Quantitative comparison on the T-SYNTH test set (n = 900 volumes). Bold denotes the best result per metric.

Method	SSIM ↑	PSNR (dB) ↑	FID ↓	PF ↓	Lesion AUC ↑
cGAN (Jiang et al., 2019)	0.608	23.9	91.2	0.193	0.709
CycleGAN	0.572	22.8	99.1	0.208	0.694
LDM (independent)	0.661	25.9	72.4	0.164	0.741
LDM + depth attn	0.718	27.9	57.8	0.151	0.793
CMFS (Fan et al., 2025)	0.634	24.8	83.5	N/A‡	0.724
MammoTomo (ours)	0.743	29.1	51.2	0.098	0.821
Real DBT (oracle)†	—	—	—	—	0.849

Table 4. Additive ablation results on the T-SYNTH test set (n = 900). Depth attention (A→B) yields the largest SSIM gain (+0.057); the physics constraint (B→C) reduces PF by 32% (0.151→0.102); lesion guidance (B→D) raises AUC by +0.021 (0.793→0.814). Variant E achieves the best performance across all metrics simultaneously.

Variant	SSIM ↑	PSNR (dB) ↑	FID ↓	PF ↓	Lesion AUC ↑
A: Base LDM	0.661	25.9	72.4	0.164	0.741
B: + Depth Attention	0.718	27.9	57.8	0.151	0.793
C: + Physics Constraint	0.731	28.4	53.8	0.102	0.799
D: + Lesion Guidance	0.724	28.1	55.4	0.148	0.814
E: Full MammoTomo	0.743	29.1	51.2	0.098	0.821

Table 5. MammoTomo synthesis performance stratified by ACR breast density category; quality degrades gracefully as tissue density increases.

ACR Density Category	SSIM ↑	PSNR (dB) ↑	PF ↓	Lesion AUC ↑
A — Almost entirely fatty	0.771	30.2	0.089	0.834
B — Scattered fibroglandular	0.758	29.6	0.094	0.828
C — Heterogeneously dense	0.731	28.7	0.101	0.815
D — Extremely dense	0.712	27.9	0.108	0.807

4.5 Qualitative Analysis

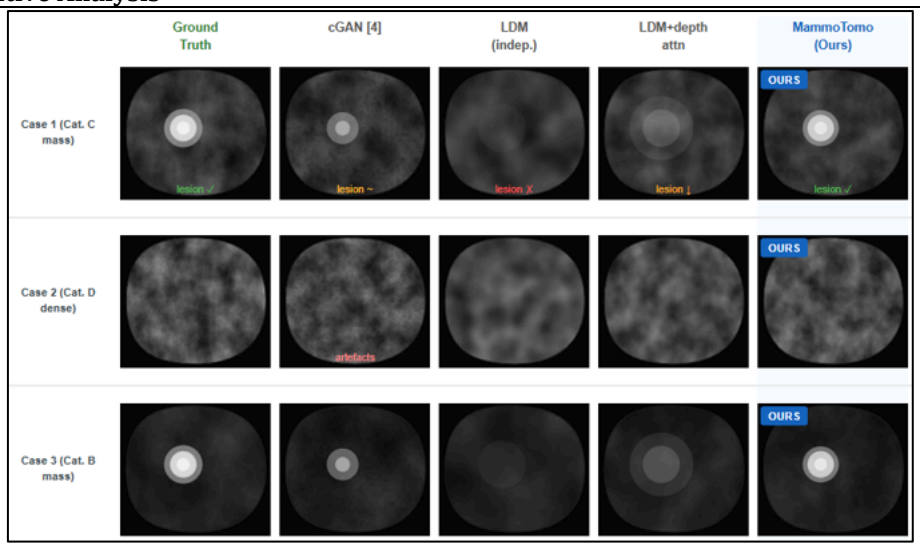

Figure 2: Mid-depth DBT slice comparisons across three representative test cases. MammoTomo preserves trabecular texture and lesion boundaries more faithfully than all baselines; cGAN exhibits checkerboard artefacts on dense tissue, while LDM variants suppress lesion detail

Fig. 2 presents mid-depth slice comparisons across three representative test cases (ACR categories B, C, and D), with MammoTomo producing the sharpest trabecular detail and most accurate lesion boundaries. CycleGAN—trained without paired supervision—yields mode-averaged texture with no lesion localization (SSIM 0.572; AUC 0.694). cGAN (Jiang et al., 2019) exhibits checkerboard artefacts on dense tissue (categories C and D). LDM (independent) achieves plausible single-slice appearance but loses inter-slice coherence and blurs lesion boundaries; LDM + depth attn recovers structural coherence yet lacks projection-fidelity and lesion-level constraints.

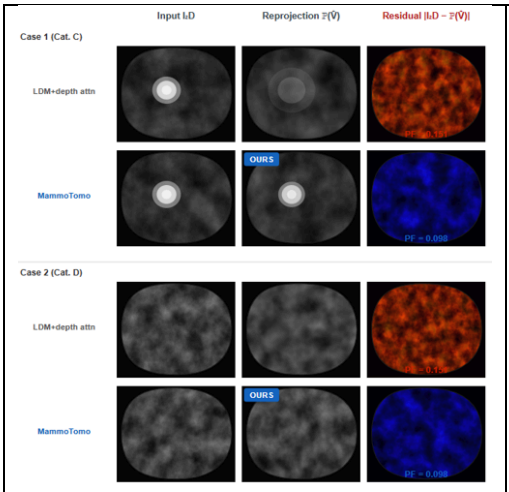

Figure 3: Projection fidelity (PF) residuals for two representative cases. MammoTomo (PF=0.098) yields 35% lower residual than LDM+depth attn (PF=0.151), ΔPF=0.053. Warm=high error, cool=low error.

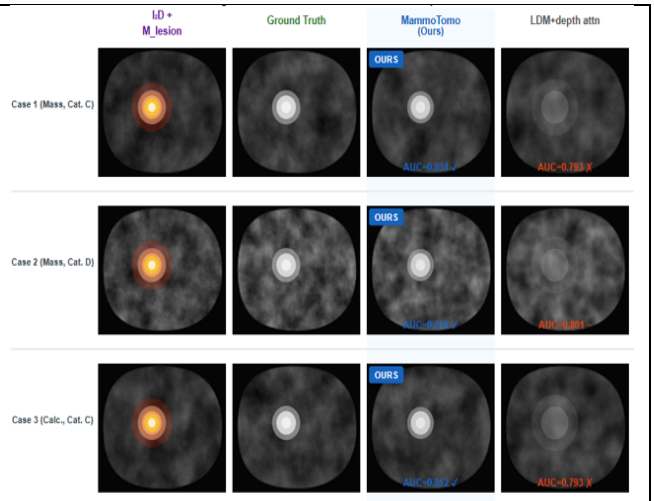

Figure 4: Lesion-aware synthesis across three confirmed malignant mass cases. MammoTomo maintains lesion visibility across the depth axis (ΔAUC = 0.028 vs. LDM + depth attn); M_{lesion} denotes the pretrained lesion predictor attention map, overlaid as a heatmap on I_2D.

Fig. 3 visualizes projection fidelity (equation (8)) for two representative cases via the Beer-Lambert operator (equation (3)) $\mathbb{P}$: I_2D, $\mathbb{P}(\hat{V})$, and the absolute residual $|I_2D - \mathbb{P}(\hat{V})|$ are shown for each method. MammoTomo residuals are substantially lower, particularly in high-attenuation pectoral and fibroglandular regions, corresponding to ΔPF = 0.053 and a 35% reduction in reprojection error.

Fig. 4 demonstrates lesion-aware synthesis across three confirmed malignant mass cases. M_{lesion}, overlaid as an orange heatmap on I_2D, localizes each finding; MammoTomo preserves lesion visibility across the

depth axis in all three cases. LDM + depth attn, which lacks the lesion injection module, suppresses or blurs the lesion in two of three cases (ΔAUC = 0.028 vs. MammoTomo; ΔAUC = 0.080 vs. LDM (independent))

4.6 Limitations and Future Work

Three limitations warrant acknowledgement. Training is conducted exclusively on T-SYNTH (Monte Carlo-simulated images), leaving the domain gap to real clinical acquisitions unquantified. The Beer-Lambert operator (equation (3)) adopts a parallel-beam approximation, neglecting scattered radiation and the focal-spot heel effect. Evaluation is confined to a single dataset; external validation on EMBED or VinDr-Mammo is required before clinical deployment. Furthermore, the multi-scale encoder (§3.2) lacks a dedicated ablation, as its contribution remains entangled with depth attention in Variant B. Future work will address these limitations through real-data fine-tuning, cone-beam projection models, and density-conditional generation.

5. Conclusion

In this study, we present MammoTomo, to our knowledge the first generative framework to synthesize a complete pseudo-DBT volume from a standard 2D mammogram at the image level, without additional imaging hardware. The proposed three-stage pipeline integrates a mammogram encoder, a pseudo-3D latent diffusion model with depth attention, a differentiable Beer-Lambert physics constraint (equations (3) and (4)), and lesion-aware attention (equation (5)) trained on CBIS-DDSM. Benchmarking on T-SYNTH (n = 900) demonstrates SSIM = 0.743, PSNR = 29.1 dB, and FID = 51.2, outperforming the strongest baseline by ΔSSIM = +0.025 and ΔFID = −6.6; Projection Fidelity (equation (8)) confirms physics self-consistency (PF = 0.098, ΔPF = 0.053), and downstream lesion detection reaches AUC = 0.821 (ΔAUC = +0.028 over the strongest baseline), establishing that synthesized volumes preserve clinically actionable diagnostic content. Ablation analysis confirms complementary contributions from depth attention, the physics constraint, and lesion-aware injection.

These results are obtained under a simulated-data setting with a parallel-beam projection approximation (§4.6), and their generalization to real clinical acquisitions and true cone-beam geometry remains to be demonstrated. Building on these results and addressing these limitations, future work will pursue real-data fine-tuning, cone-beam projection models, and density-conditional generation, alongside the broader external validation on EMBED and VinDr-Mammo motivated in Section 4.6. Code, model weights, and evaluation scripts will be released upon acceptance to support reproducibility and broader adoption.

List of Abbreviations: DBT, Digital Breast Tomosynthesis; FFDM, Full-Field Digital Mammography; DM, Digital Mammography; LDM, Latent Diffusion Model; DDPM, Denoising Diffusion Probabilistic Model; DDIM, Denoising Diffusion Implicit Model; VAE, Variational Autoencoder; GAN, Generative Adversarial Network; cGAN, Conditional Generative Adversarial Network; SSIM, Structural Similarity Index Measure; PSNR, Peak Signal-to-Noise Ratio; FID, Fréchet Inception Distance; PF, Projection Fidelity; AUC, Area Under the (ROC) Curve; ACR, American College of Radiology; CBIS-DDSM, Curated Breast Imaging Subset of the Digital Database for Screening Mammography.

Conflict of Interest: The authors declare no conflict of interest.

Acknowledgments: Not applicable.

Funding: This research received no external funding.

Author Contributions: A.K.S.: Conceptualization, Methodology, Software, Investigation, Formal Analysis, Writing – Original Draft. R.R.: Supervision, Writing – Review and Editing. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: The datasets used in this study are publicly available: T-SYNTH at https://github.com/DIDSR/tsynth-release and CBIS-DDSM at the Cancer Imaging Archive (https://www.cancerimagingarchive.net/collection/cbis-ddsm/). Code, model weights, and evaluation scripts will be released upon acceptance.

References

1. Ben-Cohen, A., Klang, E., Raskin, S.P., Soffer, S., Ben-Haim, S., Konen, E., Amitai, M.M., and Greenspan, H. (2019). Cross-modality synthesis from CT to PET using FCN and GAN networks for improved automated lesion detection. Engineering Applications of Artificial Intelligence, 78, 186–194.
2. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R.L., Soerjomataram, I., and Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 74(3), 229–263.

3. ECOG-ACRIN Cancer Research Group. (2025). Tomosynthesis Mammographic Imaging Screening Trial (TMIST) (Clinical trial NCT03233191, enrollment completion report).
4. Fan, M., Wang, L., and Wang, J., et al. (2025). Deep learning synthesis of DBT features from mammography for breast cancer diagnosis. European Journal of Radiology.
5. Friedewald, S.M., Rafferty, E.A., and Rose, S.L., et al. (2014). Breast cancer screening using tomosynthesis in combination with digital mammography. JAMA, 311(24), 2499–2507.
6. Heindel, W., et al. (2022). Digital breast tomosynthesis plus synthesised mammography versus digital screening mammography for the detection of invasive breast cancer (TOSYMA): A multicentre, open-label, randomised, controlled, superiority trial. The Lancet Oncology, 23(5), 601–611.
7. Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In Proceedings of NeurIPS.
8. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D.J. (2022). Video diffusion models. In Proceedings of NeurIPS.
9. Hofvind, S., Holen, Å.S., Aase, H.S., Houssami, N., Sebuødegård, S., Moger, T.A., Haldorsen, I.S., and Akslen, L.A. (2019). Two-view digital breast tomosynthesis versus digital mammography in a population-based breast cancer screening programme (To-Be): A randomised, controlled trial. The Lancet Oncology, 20(6), 795–805.
10. Jiang, G., He, Z., and Zhou, Y., et al. (2023). Multi-scale cascaded networks for synthesis of mammogram to decrease intensity distortion and increase model-based perceptual similarity. Medical Physics, 50(2), 837–853.
11. Jiang, G., Lu, Y., Wei, J., and Xu, Y. (2019). Synthesize mammogram from digital breast tomosynthesis with gradient guided cGANs. In Proceedings of MICCAI (pp. 801–809).
12. Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., and Rubin, D.L. (2017). A curated mammography data set for use in computer-aided detection and diagnosis research. Scientific Data, 4, 170177.
13. Moger, T.A., Swanson, J.O., Holen, Å.S., Hanestad, B., and Hofvind, S. (2019). Cost differences between digital tomosynthesis and standard digital mammography in a breast cancer screening programme: Results from the To-Be trial in Norway. The European Journal of Health Economics.
14. Montoya-del-Angel, R., Sam-Millan, K., Vilanova, J.C., and Martí, R. (2024). MAM-E: Mammographic synthetic image generation with diffusion models. Sensors, 24(7), 2076.
15. Nie, D., et al. (2017). Medical image synthesis with context-aware generative adversarial networks. In Proceedings of MICCAI.
16. Qin, Z., Liu, Z., Zhu, P., and Xue, Y. (2020). A GAN-based image synthesis method for skin lesion classification. Computer Methods and Programs in Biomedicine, 195, 105568.
17. Raissi, M., Perdikaris, P., and Karniadakis, G.E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378, 686–707.
18. Wiedeman, C., Sarmakeeva, A., Sizikova, E., Filienko, D., Lago, M., Delfino, J.G., and Badano, A. (2025). T-SYNTH: A knowledge-based dataset of synthetic breast images. arXiv:2507.04038.
19. Wuerfl, T., et al. (2018). Deep learning CT: Learning projection-domain weights from image domain. IEEE Transactions on Medical Imaging.
20. Zhang, L., Rao, A., and Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In Proceedings of ICCV (pp. 3836–3847).