FREQUENCY-DISENTANGLED MULTI-PHASE BREAST DCE-MRI SYNTHESIS FROM PRE-CONTRAST MRI

Anonymous authors
Paper under double-blind review

ABSTRACT

Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) captures spatially heterogeneous and temporally evolving tumor enhancement. Synthesizing multi-phase DCE-MRI from pre-contrast MRI therefore requires recovering patient-specific enhancement patterns rather than merely generating realistic images. However, existing methods often oversmooth intratumoral heterogeneity and introduce spurious enhancement outside lesions. We propose Freq-Syn, a frequency-disentangled framework that synthesizes multi-phase breast DCE-MRI from a pre-contrast 3D volume and target acquisition times. Unlike whole-image prediction, Freq-Syn represents enhancement as a progressive Gaussian scale-space hierarchy of global contrast, regional uptake, lesion morphology, and fine texture. Cross-phase interaction jointly models their temporal evolution, while synthesis-guided localization constrains where heterogeneous enhancement is generated. Lesion annotations provide targeted supervision during training but are not required at inference. On the internal and external test sets, Freq-Syn achieves SSIM scores of 0.83 and 0.81, respectively. Compared with the best-performing comparison methods, Freq-Syn reduces the false enhancement rate from 40.47% to 16.93% internally and from 37.49% to 19.59% externally. For downstream lesion segmentation, Freq-Syn consistently achieves higher Dice scores than all comparison synthesis methods on both test sets, demonstrating that the synthesized images preserve task-relevant tumor information. These results highlight the effectiveness of explicitly modeling spatial and temporal enhancement patterns for faithful multi-phase DCE-MRI synthesis. These results demonstrate the effectiveness of explicitly modeling spatial and temporal enhancement patterns for faithful multi-phase DCE-MRI synthesis. The code will be made publicly available upon publication.

1 INTRODUCTION

Multi-time-point breast dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) synthesis aims to generate post-contrast images at different acquisition times from pre-contrast MRI, offering a potential means to model dynamic lesion enhancement without a complete contrast-enhanced acquisition (Schreiter et al., 2025; Osuala et al., 2025; Lang et al., 2025). Beyond generating realistic images, this task requires recovering lesion-specific enhancement patterns that vary across patients, spatial regions, and acquisition times. Such heterogeneous enhancement dynamics are associated with lesion characteristics and tumor biology (Mann et al., 2019; Kim et al., 2022; Zhao et al., 2020). However, existing synthesis objectives remain largely dominated by image-level fidelity, which may attenuate intratumoral heterogeneity or introduce unsupported enhancement outside the lesion (Osuala et al., 2025). As illustrated in Fig. 1(a), faithful multi-time-point DCE-MRI synthesis therefore requires modeling both the spatial organization and temporal evolution of lesion enhancement.

Existing approaches have progressed from single-time-point synthesis (Zhang et al., 2023b) to joint multi-time-point generation (Schreiter et al., 2025; Osuala et al., 2025) and explicit temporal conditioning (Lang et al., 2025; Chung et al., 2025). Nevertheless, synthetic DCE-MRI may still exhibit false-positive enhancement, incomplete tumor enhancement, inaccurate temporal intensity changes, and missing intratumoral features (Osuala et al., 2025). Tumor-aware losses and mask conditioning improve lesion-region synthesis (Ibarra et al., 2025), but such guidance remains relatively coarse and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

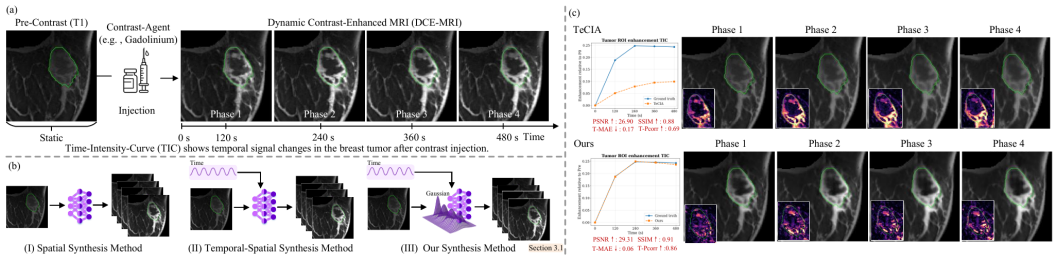


Figure 1: Motivation for multi-time-point breast DCE-MRI synthesis. (a) Complementary characteristics of pre-contrast MRI and DCE-MRI motivate contrast-free dynamic synthesis. (b) Comparison of existing DCE-MRI synthesis paradigms with our approach. (c) Visual comparison with a recent method (Lang et al., 2025).

does not explicitly characterize patient-specific intratumoral enhancement heterogeneity. As illustrated in Fig. 1(b,c), the remaining challenge is therefore not only to determine *when* and *how much* a lesion should enhance, but also *where* heterogeneous enhancement should occur, while avoiding unsupported enhancement in normal tissue.

To address this challenge, we propose **Freq-Syn**, a frequency-disentangled framework for multi-time-point breast DCE-MRI synthesis from a pre-contrast 3D T1-weighted image and target acquisition times. Instead of directly treating each post-contrast phase as a monolithic image target, Freq-Syn models the enhancement residual through a progressive Gaussian scale-space hierarchy that separates global contrast, regional uptake, lesion morphology, and fine texture. Cross-phase interaction models temporal uptake evolution, while synthesis-guided localization progressively focuses heterogeneous enhancement on the lesion. Lesion annotations provide targeted supervision during training but are not required at inference. Together, these components enable explicit modeling of lesion-specific spatial enhancement heterogeneity and its temporal evolution across DCE phases.

Our main contributions are:

- We recast multi-phase DCE-MRI synthesis as an exact additive Gaussian scale-space factorization of the enhancement residual. This formulation replaces a monolithic image target with explicit learning targets for global contrast, regional uptake, lesion morphology, and fine texture, making intratumoral heterogeneity directly optimizable.
- We instantiate this formulation with a progressive 3D architecture that couples voxel-wise cross-phase interaction, an error-aware hard uptake codebook, and one-step synthesis-guided lesion refinement. The resulting coarse-to-fine pathway jointly determines what, when, and where to enhance without requiring a lesion mask at inference.
- Through internal and independent external evaluation, we show that Freq-Syn achieves a stronger overall balance of image fidelity, tumor-pattern preservation, and false-enhancement suppression than the evaluated synthesis methods. Its gains further transfer to tumor segmentation and triple-negative breast cancer classification, providing evidence that the synthesized enhancement retains task-relevant information.

2 Related Work

2.1 Contrast-Enhanced Breast MRI Synthesis

Early studies on contrast-enhanced breast MRI synthesis mainly focused on predicting a single post-contrast image from non-contrast acquisitions (Zhang et al., 2023b). Recent work has extended this setting to multi-timepoint DCE generation. Lang et al. (2025) synthesize DCE images at different acquisition times from pre-contrast MRI, while Chung et al. (2025) generate delayed DCE images from early acquisitions with explicit temporal conditioning. Related contrast-enhanced CT synthesis further models dependencies across acquisition timepoints to improve temporal consistency Zhu et al. (2025). Despite these advances, existing methods primarily optimize synthesis at the im-

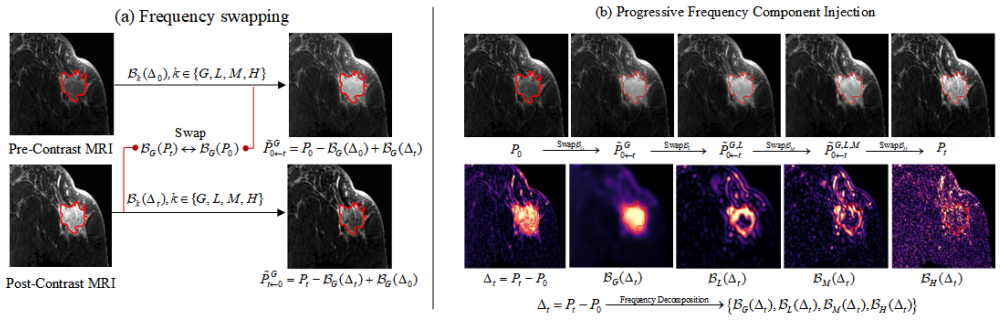


Figure 2: Frequency-specific DCE enhancement analysis. (a) Gaussian scale-space component swapping between P_0 and P_k, where $\mathcal{B}_r$, $r \in G, L, M, H$, represents information from global contrast to fine details. (b) Progressive injection of the decomposed components reconstructs overall contrast, regional uptake, lesion morphology, and fine textures. Red contours delineate the tumor.

age level, leaving regional enhancement variations within the tumor largely implicit. Consequently, intratumoral enhancement heterogeneity is not explicitly constrained during generation. Lesion-focused generation demonstrates the benefit of explicitly modeling pathological regions Zhang et al. (2025), while anatomy-consistent MR synthesis emphasizes preserving underlying structures across contrast changes Shin et al. (2025). However, these approaches are not designed to characterize how the overall enhancement response and heterogeneous regional variations within the same tumor evolve across DCE timepoints. In contrast, our method explicitly exploits tumor-specific information by separately modeling overall lesion enhancement and intratumoral heterogeneous variations over time, while suppressing unsupported enhancement outside the lesion.

2.2 Frequency Learning

Frequency-domain learning has been increasingly explored to disentangle complementary image information across spatial scales. AdaIR adaptively mines and modulates low- and high-frequency representations for image restoration Cui et al. (2025b), while Zhao et al. (2025) progressively decompose restoration into global mapping, coarse-content recovery, and fine-detail refinement from a spectral perspective. Frequency-domain representations have also been introduced into medical imaging. FreMIM exploits complementary low- and high-frequency information to capture global semantics and detailed structures in medical image segmentation Wang et al. (2024), while multi-scale frequency representations have been used to integrate complementary information for pathological image analysis Cui et al. (2025a). These studies show that frequency decomposition provides complementary representations of global image characteristics and localized spatial variations. This property is well aligned with breast DCE-MRI, where tumor enhancement contains both an overall enhancement response and heterogeneous regional variations within the lesion. Motivated by this connection, we employ frequency-aware representations to separately characterize lesion-level enhancement and fine-grained intratumoral variations, while temporal modeling captures how these enhancement patterns evolve across DCE timepoints. As illustrated in Fig. 2, we further examine the spatial information represented by different Gaussian frequency components. Swapping individual components between the pre- and post-contrast images shows that the global- and low-frequency components primarily characterize overall contrast and regional uptake, whereas the mid- and high-frequency components preserve lesion morphology and fine texture. Progressive component injection further demonstrates that these complementary components gradually recover the complete enhancement pattern. This observation motivates our progressive frequency synthesis design.

3 Method

3.1 Overview

Given a pre-contrast volume $P_0 \in \mathbb{R}^{1\times D\times H\times W}$ and the acquisition times $\{t_k\}_{k=1}^{\tau}$ of τ target phases, we synthesize $\{\widehat{P}_k\}_{k=1}^{\tau}$. Here, D, H, and W are the spatial dimensions, k indexes the

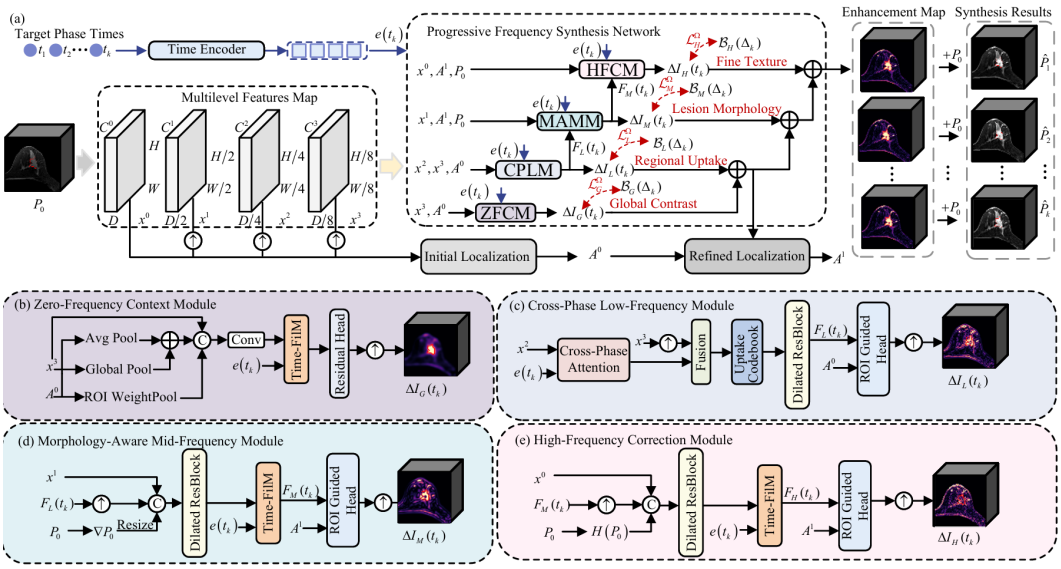


Figure 3: Overview of the proposed Progressive Frequency Synthesis Network. Given a pre-contrast volume P_0 and target times $\{t_k\}_{k=1}^{\tau}$, the network progressively predicts global contrast, regional uptake, lesion morphology, and fine-texture components supervised by $\{\mathcal{B}_G(\Delta_k), \mathcal{B}_L(\Delta_k), \mathcal{B}_M(\Delta_k), \mathcal{B}_H(\Delta_k)\}$, where $\Delta_k = P_k - P_0$. The predicted components are combined with P_0 to synthesize multi-phase DCE-MRI. (b–e) Detailed structures of ZFCM, CPLM, MAMM, and HFCM, respectively. $\oplus$ and C denote element-wise addition and channel-wise concatenation.

target phase, and $\Delta_k = P_k - P_0$ is the ground-truth enhancement at time t_k. Rather than directly predicting the complete enhancement, our Progressive Frequency Synthesis Network estimates four complementary components:

$$\widehat{\Delta}_k = \sum_{r\in\{G,L,M,H\}} \Delta I_r(t_k), \qquad \widehat{P}_k = P_0 + \widehat{\Delta}_k, \tag{1}$$

where $r \in \{G, L, M, H\}$ indexes the global-, low-, mid-, and high-frequency levels, corresponding to global contrast, regional uptake, lesion morphology, and fine texture, respectively.

A shared three-dimensional encoder extracts multilevel anatomical features $\{x^i\}_{i=0}^{3}$, where x^i has $1/2^i$ of the input resolution. An initial soft lesion map A^0 derived from these features guides the Global and Low stages. Their coarse multi-phase predictions refine the map to A^1, which guides the Mid and High stages. Ground-truth masks are used only for training supervision and are not required at inference. We use A^j, $j \in \{0, 1\}$, when referring to either localization stage.

3.2 Gaussian Frequency Supervision

We construct a Gaussian scale-space hierarchy from Δ_k. Let $\mathcal{G}^n(X)$ denote n successive applications of a fixed separable Gaussian filter to X. The target components are

$$\begin{aligned} \mathcal{B}_G(\Delta_k) &= \mathcal{G}^8(\Delta_k), & \mathcal{B}_L(\Delta_k) &= \mathcal{G}^4(\Delta_k) - \mathcal{G}^8(\Delta_k), \\ \mathcal{B}_M(\Delta_k) &= \mathcal{G}(\Delta_k) - \mathcal{G}^4(\Delta_k), & \mathcal{B}_H(\Delta_k) &= \Delta_k - \mathcal{G}(\Delta_k). \end{aligned} \tag{2}$$

The four components sum exactly to Δ_k and define scale-specific targets while preserving the complete enhancement. Unlike direct synthesis, this additive hierarchy explicitly supervises weaker lesion morphology and texture components that may otherwise be dominated by smooth enhancement responses during optimization. It serves as a spatial scale-space supervision hierarchy rather than an orthogonal or physical decomposition of contrast kinetics. The filter implementation is provided in Appendix A.

3.3 Temporal Encoding and Cross-Phase Interaction

The time encoder maps the normalized absolute time, inter-phase interval, and relative phase position to a phase-specific embedding $e(t_k)$. To exchange information among target phases, we replicate x^2 across time and apply self-attention only along the phase dimension at each spatial location $\mathbf{p}$:

$$\begin{aligned} Z_{\mathbf{p}} &= \left[x^2(\mathbf{p}) + W_t e(t_k)\right]_{k=1}^{\tau}, \\ U_{\mathbf{p}} &= Z_{\mathbf{p}} + \mathrm{MHA}(\mathrm{LN}(Z_{\mathbf{p}})), \\ F_{\mathbf{p}} &= U_{\mathbf{p}} + \mathrm{FFN}(\mathrm{LN}(U_{\mathbf{p}})). \end{aligned} \tag{3}$$

Here, W_t is a shared temporal projection, while LN, MHA, and FFN denote layer normalization, multi-head self-attention, and a feed-forward network. Reshaping $F_{\mathbf{p}}$ back to volume form gives $\{F_T(t_k)\}_{k=1}^{\tau}$. The exact temporal encoding is provided in Appendix A.

3.4 Progressive Frequency Synthesis

Zero-Frequency Context Module. ZFCM fuses x^3 with local, global, and A^0-weighted lesion context. The feature is conditioned on $e(t_k)$ using feature-wise linear modulation, $\mathrm{FiLM}(F, e) = F \odot [1 + \gamma(e)] + \beta(e)$, and decoded by one residual head to produce $\Delta I_G(t_k)$. The soft map A^0 partitions this prediction into background and lesion contributions for subsequent analysis; it does not introduce two independent Global prediction heads.

Cross-Phase Low-Frequency Module. CPLM concatenates $F_T(t_k)$ with resized x^3 and projects the fused feature z_e into an uptake codebook. With $\bar{z}_e$ and $\bar{c}_m$ denoting the ℓ_2-normalized encoder feature and the m-th codeword, hard cosine quantization with straight-through decoding is

$$m^*(\mathbf{p}) = \arg\max_m \langle \bar{z}_e(\mathbf{p}), \bar{c}_m \rangle, \qquad z_q = \bar{z}_e + \mathrm{sg}(\bar{c}_{m^*} - \bar{z}_e), \tag{4}$$

where sg stops gradients. A dilated residual block decodes z_q into background and lesion heads. Temporal information enters CPLM only through $F_T(t_k)$; no second temporal modulation is applied after quantization.

Morphology-Aware Mid-Frequency Module. MAMM concatenates x^1, the resized Low-stage feature $F_L(t_k)$, the pre-contrast gradient ∇P_0, the coarse enhancement $\widehat{\Delta}_k^{G,L} = \Delta I_G(t_k) + \Delta I_L(t_k)$, and A^1. After residual fusion, the refined localization feature F_{loc} is added and a time-conditioned dilated block produces $F_M(t_k)$. For the Low and Mid stages, the two prediction heads are combined as

$$\Delta I_r(t_k) = (1 - A^j) \odot h_r^{\mathrm{bg}} + A^j \odot \left[h_r^{\mathrm{tu}} - \mu_{A^j}(h_r^{\mathrm{tu}})\right], \quad (r, j) \in \{(L, 0), (M, 1)\}, \tag{5}$$

where h_r^{bg} and h_r^{tu} are the background and lesion-head outputs, and μ_{A^j} is their soft lesion-weighted spatial mean. Centering prevents these stages from repeatedly altering the lesion-wide mean enhancement.

High-Frequency Correction Module. HFCM combines x^0, projected $F_M(t_k)$, A^1, and the local high-pass responses of P_0 and $\widehat{\Delta}_k^{G,L,M} = \sum_{r \in \{G,L,M\}} \Delta I_r(t_k)$, where $\mathcal{H}(X) = X - \mathrm{AvgPool}_{3\times3\times3}(X)$. Separate high-pass background and lesion heads produce $\Delta I_H(t_k)$ without the zero-mean constraint in Eq. 5. All four outputs are supervised by their corresponding targets in Eq. 2.

3.5 Synthesis-Guided Lesion Refinement

Because a tumor may be inconspicuous in P_0, we refine A^0 using emerging multi-phase enhancement. For each phase, $F_L(t_k)$ and $\widehat{\Delta}_k^{G,L}$ are encoded into a lesion feature q_k. A spatial score $s_k(\mathbf{v})$ is normalized across phases at each voxel $\mathbf{v}$:

$$\alpha_k(\mathbf{v}) = \frac{\exp(s_k(\mathbf{v}))}{\sum_{u=1}^{\tau} \exp(s_u(\mathbf{v}))}, \qquad \overline{q}(\mathbf{v}) = \sum_{k=1}^{\tau} \alpha_k(\mathbf{v}) q_k(\mathbf{v}). \tag{6}$$

Here, $\alpha_k(\mathbf{v})$ is the phase contribution and $\bar{q}(\mathbf{v})$ is the aggregated lesion representation. It is concatenated with x^1, projected x^0, and A^0 to produce F_{loc} and A^1. Both A^0 and A^1 are full-resolution sigmoid probability maps used as soft gates. Thus, coarse synthesis refines localization once before Mid and High synthesis; the design is not an iterative feedback loop.

3.6 Training Objectives

Let A denote the ground-truth lesion mask, Ω the valid breast region, and ∂A its morphological boundary band. For volumes X and Y, we define the regional Charbonnier loss

$$\ell_{\mathcal{R}}(X, Y) = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{v} \in \mathcal{R}} \sqrt{(X(\mathbf{v}) - Y(\mathbf{v}))^2 + \epsilon^2}, \tag{7}$$

where $\mathcal{R}$ is the supervised region, $\mathbf{v}$ is a voxel, $|\mathcal{R}|$ is its volume, and ϵ ensures numerical stability. The frequency outputs are averaged over all target phases:

$$\begin{aligned} \mathcal{L}_{\text{freq}} = \lambda_G \mathcal{L}_G^{\Omega} + \lambda_L \mathcal{L}_L^{\Omega} \\ + s_M \lambda_M \left(\mathcal{L}_M^{\Omega} + 0.5\mathcal{L}_M^{A} + 0.5\mathcal{L}_M^{\partial A} \right) \\ + s_H \lambda_H \left(\mathcal{L}_H^{\Omega} + 0.5\mathcal{L}_H^{A} + 0.5\mathcal{L}_H^{\partial A} \right), \end{aligned} \tag{8}$$

where $\mathcal{L}_r^{\mathcal{R}} = \tau^{-1} \sum_{k=1}^{\tau} \ell_{\mathcal{R}}(\Delta I_r(t_k), \mathcal{B}_r(\Delta_k))$, $\mathcal{R} \in \{\Omega, A, \partial A\}$, and $s_M, s_H \in [0, 1]$ progressively activate the Mid and High terms.

The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{syn}} + \mathcal{L}_{\text{freq}} + \mathcal{L}_{\text{layer}} + \mathcal{L}_{\text{st}} + \mathcal{L}_{\text{loc}}. \tag{9}$$

Here, $\mathcal{L}_{\text{syn}}$ combines region-aware final-output fidelity with a one-sided false-enhancement penalty; $\mathcal{L}_{\text{layer}}$ regularizes the codebook and the Mid and High outputs; $\mathcal{L}_{\text{st}}$ preserves intratumoral heterogeneity and temporal kinetics; and $\mathcal{L}_{\text{loc}}$ supervises A^0 and A^1. Complete definitions, coefficients, spatial regions, and curricula are given in Appendix B.

4 Experiments

4.1 Setup

Dataset. We build our study on a multicenter breast MRI cohort comprising MAMA-MIA (Garrucho et al., 2025) and two private datasets, Center A and Center B. Following recent breast MRI studies (Chen et al., 2024; Luo et al., 2025; Liang et al., 2026), every sample undergoes the same preprocessing pipeline: inter-phase registration with deedsBCV (Chen et al., 2026), N4 bias field correction (Liang et al., 2026), removal of non-breast regions (Zhang et al., 2023a), and resampling to 1 mm^3 isotropic voxels using third-order B-spline interpolation for images and nearest-neighbor interpolation for masks (Liang et al., 2026). All volumes are cropped to $64 \times 96 \times 96$ voxels. We then remove intensity outliers and normalize all phases to $[0, 1]$ using shared intensity statistics to preserve temporal enhancement variations (Zhang et al., 2023a). MAMA-MIA and Center A are randomly split into train/validation/internal-test sets at a 70/10/20 ratio, while Center B serves as an independent external test set.

Implementation Details. The network uses 12 base channels, a 64-dimensional temporal embedding, four attention heads, and a codebook of 64 embeddings with dimension 32. Absolute time and inter-phase intervals are normalized by 480 seconds and encoded with eight sinusoidal frequencies. We train for 300 epochs with a batch size of 2 using AdamW, a learning rate of 10^{-4}, and a weight decay of 10^{-4}. The learning rate is warmed up for five epochs, and gradient norms are clipped to 5. The localization branches are pretrained for three epochs. Remaining optimization details and curricula are given in Appendix B.

Method Comparison. We compare our method with representative DCE-MRI synthesis approaches spanning several generation paradigms. We include a standard baseline and UNet (Ronneberger et al., 2015) as deterministic image-to-image references. AAD-DCE (Bharti et al., 2025) represents attention-based contrast synthesis, while CC-Net (Osuala et al., 2024) represents conditional latent diffusion. MCO-Net (Schreiter et al., 2025) jointly predicts multiple DCE phases

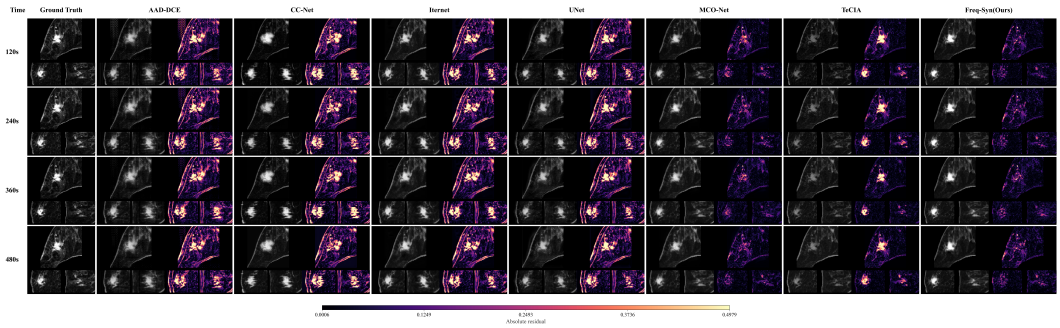


Figure 4: Phase-wise qualitative comparison of multi-phase DCE-MRI synthesis. Synthesized images and absolute residual maps are shown at 120, 240, 360, and 480 seconds, with enlarged tumor views provided below.

Table 1: Quantitative comparison of DCE-MRI synthesis performance on the internal and external test sets. T-MAE denotes tumor-region mean absolute error, T-PCorr denotes tumor enhancement pattern correlation, and FE-Rate denotes the false enhancement rate. ↑ indicates higher is better and ↓ indicates lower is better. Best results are shown in bold.

	Internal Test Set						External Test Set					
Method	PSNR [dB] ↑	SSIM ↑	MAE ↓	T-MAE ↓	T-PCorr ↑	FE-Rate (%) ↓	PSNR [dB]↑	SSIM ↑	MAE ↓	T-MAE ↓	T-PCorr ↑	FE-Rate (%) ↓
AAD-DDE (Bharti et al., 2025)	18.53	0.36	0.14	0.14	0.18	67.02	16.81	0.31	0.09	0.18	0.17	63.31
CC-Net (Osuala et al., 2024)	15.12	0.46	0.11	0.17	0.06	67.08	16.29	0.48	0.06	**0.14**	0.20	51.01
Iternet (Chung et al., 2025)	22.40	0.65	0.06	**0.12**	0.30	56.91	20.29	0.56	0.06	**0.14**	**0.43**	58.37
UNet (Ronneberger et al., 2015)	**26.29**	0.80	0.04	0.14	0.31	40.47	**25.77**	0.78	**0.04**	**0.14**	0.40	37.49
MCO-Net (Schreiter et al., 2025)	26.11	0.81	0.04	0.14	0.30	44.89	25.65	0.79	**0.04**	**0.14**	0.40	41.55
TeCIA(3D) (Lang et al., 2025)	24.63	0.81	0.05	0.23	0.30	44.47	24.36	0.80	**0.04**	0.25	0.40	37.63
Ours	26.09	**0.83**	**0.03**	0.13	**0.37**	**16.93**	25.41	**0.81**	**0.04**	**0.14**	0.41	**19.59**

using a multi-output architecture. Iternet (Chung et al., 2025) models contrast evolution through iterative prediction, and TeCIA (3D) (Lang et al., 2025) provides an explicit temporal modeling baseline. These methods cover deterministic, attention-based, diffusion-based, multi-output, iterative, and temporal synthesis strategies. They therefore provide a broad comparison for evaluating our spatial-scale heterogeneity modeling and cross-phase temporal design.

Metrics. For synthesis evaluation, PSNR, SSIM, and MAE measure overall image fidelity; tumor-region MAE (T-MAE) and tumor pattern correlation (T-PCorr) assess lesion enhancement intensity and intratumoral spatial patterns, respectively; and false enhancement rate (FE-Rate) quantifies unsupported enhancement in non-enhancing regions. All synthesis metrics are averaged across the target phases. For downstream evaluation, breast tumor segmentation is assessed using Dice, IoU, sensitivity, and precision, while triple-negative breast cancer classification is evaluated using AUC, sensitivity, and specificity.

4.2 Synthesis and Ablation Results

Qualitative results. Figure 4 compares synthesized images and absolute residuals across four acquisition times. AAD-DCE and CC-Net produce diffuse errors beyond the lesion, whereas Iternet and TeCIA tend to underestimate intratumoral enhancement. UNet and MCO-Net better recover the overall appearance but retain noticeable residuals around the lesion and in non-lesion tissue. In contrast, Freq-Syn yields weaker, more spatially confined residuals across phases while retaining heterogeneous enhancement within the tumor, consistent with its explicit separation of regional uptake, morphology, and fine texture.

Quantitative results. Table 1 summarizes the quantitative results. On the internal test set, Freq-Syn achieves the best SSIM (0.83), MAE (0.03), T-PCorr (0.37), and FE-Rate (16.93%), while its PSNR (26.09 dB) remains close to the best result (26.29 dB). On the external set, it obtains the best SSIM (0.81) and FE-Rate (19.59%) and matches the lowest MAE (0.04) and T-MAE (0.14), demonstrating consistent generalization across centers.

Table 2: Ablation study on the internal and external test sets. T-MAE denotes tumor-region mean absolute error, T-PCorr denotes tumor enhancement pattern correlation, and FE-Rate denotes false enhancement rate. Best results are shown in bold.

	Internal Test Set					External Test Set				
Variant	PSNR ↑	SSIM ↑	T-MAE ↓	T-PCorr ↑	FE-Rate ↓	PSNR ↑	SSIM ↑	T-MAE ↓	T-PCorr ↑	FE-Rate ↓
Direct synthesis	21.70	0.60	0.13	0.27	47.29	19.30	0.53	0.15	0.27	56.15
w/o $\mathcal{L}_{\mathrm{freq}}$	24.82	0.80	0.15	0.31	27.83	24.19	0.77	0.17	0.33	32.46
w/o cross-phase attention	25.48	0.81	0.15	0.33	22.64	24.76	0.79	0.16	0.35	26.87
w/o refined localization	25.61	0.82	0.15	0.34	25.92	24.84	0.79	0.17	0.37	30.41
w/o $\mathcal{L}_{\mathrm{st}}$	25.87	0.82	0.14	0.30	18.37	25.16	0.80	0.15	0.34	21.72
w/o uptake codebook	25.76	0.82	0.14	0.35	20.28	25.02	0.80	0.15	0.38	23.65
Full model	**26.09**	**0.83**	**0.13**	**0.37**	**16.93**	**25.41**	**0.81**	**0.14**	**0.41**	**19.59**

Compared with TeCIA, Freq-Syn reduces T-MAE by 43.5%/44.0% and FE-Rate by 61.9%/47.9% on the internal/external sets. Relative to the lowest FE-Rates among all competing methods (40.47%/37.49%), it further reduces erroneous enhancement to 16.93%/19.59%. Importantly, this reduction is accompanied by competitive T-MAE and improved T-PCorr rather than a loss of lesion enhancement, indicating that Freq-Syn localizes enhancement more accurately instead of uniformly suppressing its intensity.

Ablation study. Table 2 evaluates the main components of Freq-Syn. Compared with direct synthesis, the full model improves PSNR from 21.70/19.30 to 26.09/25.41 dB, SSIM from 0.60/0.53 to 0.83/0.81, and T-PCorr from 0.27/0.27 to 0.37/0.41, while reducing FE-Rate from 47.29%/56.15% to 16.93%/19.59% on the internal/external sets.

Removing $\mathcal{L}_{\mathrm{freq}}$ causes the largest degradation in overall synthesis quality, reducing PSNR to 24.82/24.19 dB and increasing FE-Rate to 27.83%/32.46%. Cross-phase attention primarily improves temporal consistency, while refined lesion localization suppresses false enhancement. The spatiotemporal loss improves tumor-pattern preservation, and the uptake codebook provides additional gains in T-PCorr and FE-Rate. Overall, these components provide complementary benefits in image fidelity, tumor-pattern preservation, and false-enhancement suppression.

4.3 Downstream Evaluation

Tumor segmentation. Table 3 evaluates whether the synthesized enhancement remains useful for tumor delineation. Freq-Syn improves Dice from 0.60/0.53 with pre-contrast images to 0.71/0.65 on the internal/external sets, exceeding the strongest synthesis baseline, MCO-Net, by 0.05/0.05. The simultaneous gains in IoU and precision indicate more accurate boundaries and fewer false-positive regions. Although sensitivity is lower, reflecting a more conservative delineation, Freq-Syn substantially narrows the gap to real post-contrast images (0.79/0.71 Dice).

Tumor classification. Triple-negative breast cancer exhibits particularly pronounced inter- and intratumoral heterogeneity among breast cancer subtypes (Kim et al., 2022). We therefore use triple-negative versus non-triple-negative classification to test whether synthesized DCE-MRI preserves subtype-related enhancement phenotypes rather than only image appearance. As shown in Table 4, Freq-Syn improves the internal AUC from 0.61 with pre-contrast images to 0.64, the best result among synthesis methods and close to the post-contrast reference of 0.67. On the external set, it improves AUC from 0.50 to 0.61 and approaches the post-contrast result of 0.63, although CC-Net reaches 0.66. The consistent improvement over pre-contrast input indicates that Freq-Syn recovers complementary tumor information, while the remaining variability across centers highlights the difficulty of preserving subtype-specific heterogeneity.

5 Conclusion

We studied multi-time-point breast DCE-MRI synthesis with a focus on patient-specific intratumoral enhancement heterogeneity, which can be insufficiently captured by image-level reconstruction objectives. By organizing enhancement modeling along spatial frequency and temporal evolution, Freq-Syn separates overall enhancement from localized variations and models their evolution across

Table 3: Downstream breast lesion segmentation performance using synthesized DCE-MRI on the internal and external test sets. The same segmentation model is used for all synthesis methods. Best synthesis-based results are shown in bold.

	Internal Test Set				External Test Set			
Method	Dice ↑	IoU ↑	Sensitivity ↑	Precision ↑	Dice ↑	IoU ↑	Sensitivity ↑	Precision ↑
Pre-Contrast	0.60	0.48	0.83	0.57	0.53	0.40	0.79	0.48
AAD-DCE (Bharti et al., 2025)	0.48	0.35	**0.91**	0.39	0.41	0.30	**0.88**	0.31
CC-Net (Osuala et al., 2024)	0.53	0.40	0.86	0.46	0.46	0.34	**0.88**	0.38
Iternet (Chung et al., 2025)	0.47	0.37	0.88	0.38	0.38	0.27	0.87	0.30
UNet (Ronneberger et al., 2015)	0.55	0.42	0.85	0.48	0.46	0.34	0.82	0.40
MCO-Net (Schreiter et al., 2025)	0.66	0.52	0.85	0.60	0.60	0.46	0.84	0.53
TeCIA3D (Lang et al., 2025)	0.61	0.47	0.89	0.51	0.56	0.43	0.84	0.49
Real Post-Contrast	0.79	0.68	0.81	0.82	0.71	0.59	0.71	0.75
Ours	**0.71**	**0.57**	0.75	**0.75**	**0.65**	**0.51**	0.68	**0.72**

Table 4: Downstream triple-negative versus non-triple-negative breast cancer classification using synthesized DCE-MRI on the internal and external test sets. The same classifier is used for all synthesis methods. Best synthesis-based results are shown in bold.

	Internal Test Set			External Test Set		
Method	AUC ↑	Sensitivity ↑	Specificity ↑	AUC ↑	Sensitivity ↑	Specificity ↑
Pre-Contrast	0.61	0.55	0.62	0.50	0.37	0.69
AAD-DCE (Bharti et al., 2025)	0.56	0.72	0.41	0.62	0.37	**0.88**
CC-Net (Osuala et al., 2024)	0.57	**0.85**	0.28	**0.66**	0.47	0.85
Iternet (Chung et al., 2025)	0.56	0.66	0.47	0.60	0.63	0.62
UNet (Ronneberger et al., 2015)	0.60	0.55	0.67	0.57	0.84	0.45
MCO-Net (Schreiter et al., 2025)	0.62	0.43	**0.79**	0.51	0.68	0.48
TeCIA3D (Lang et al., 2025)	0.58	0.52	0.64	0.52	**0.95**	0.62
Real Post-Contrast	0.67	0.71	0.55	0.63	0.42	0.75
Ours	**0.64**	0.73	0.50	0.61	0.53	0.71

DCE phases. Experiments on internal and external test sets show that Freq-Syn improves tumor-pattern preservation and substantially reduces false enhancement while maintaining competitive image fidelity and downstream utility. These results suggest that explicitly modeling spatial and temporal enhancement patterns can provide a useful complement to conventional whole-image synthesis objectives. More broadly, our findings suggest that DCE-MRI synthesis should be evaluated not only by visual or voxel-wise similarity, but also by the preservation of disease-relevant spatial patterns, temporal dynamics, and downstream utility. Nevertheless, predicting dynamic enhancement from pre-contrast MRI is inherently uncertain, as structural imaging cannot fully determine patient-specific perfusion and vascular behavior. In addition, evaluation on a single external cohort cannot capture the full variability of scanners, acquisition protocols, patient populations, and tumor phenotypes. Future work should therefore investigate uncertainty-aware generation, broader multi-center validation, and prospective evaluation to better establish the reliability and potential clinical utility of synthesized DCE-MRI.

AI USE STATEMENT

Large language model-based tools were used solely for language editing, bibliographic cross-checking, figure-layout suggestions, and LaTeX formatting. They were not used to formulate the research questions, develop the methodology, design or conduct the experiments, analyze the results, or draw the conclusions. All AI-assisted content was carefully reviewed and verified by the authors, who assume full responsibility for the integrity and accuracy of this paper.

REFERENCES

Divya Bharti, Sriprabha Ramanarayanan, Sadhana S, Kishore Kumar M, Keerthi Ram, Harsh Agarwal, Ramesh Venkatesan, and Mohanasankar Sivaprakasam. AAD-DCE: An aggregated

multimodal attention mechanism for early and late dynamic contrast enhanced prostate MRI synthesis. *arXiv preprint arXiv:2502.02555*, 2025. doi: 10.48550/arXiv.2502.02555. URL https://arxiv.org/abs/2502.02555.

Hao Chen, Rui Yin, Yifan Chen, Qi Chen, and Chao Li. Learning patient-specific disease dynamics with latent flow matching for longitudinal imaging generation. In *International Conference on Learning Representations*, volume 2026, pp. 82707–82723, 2026.

Qianqian Chen, Jiadong Zhang, Runqi Meng, Lei Zhou, Zhenhui Li, Qianjin Feng, and Dinggang Shen. Modality-specific information disentanglement from multi-parametric MRI for breast tumor segmentation and computer-aided diagnosis. *IEEE Transactions on Medical Imaging*, 43(5): 1958–1971, 2024. doi: 10.1109/TMI.2024.3352648.

Woojin Chung, Junghwa Kang, Ga Eun Park, Sung Hun Kim, and Yoonho Nam. Synthesizing delayed-phase contrast-enhanced breast mr images from early-phase images using an iterative deep network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 583–592. Springer, 2025.

Xiaoyu Cui, Weixing Chen, and Jiandong Su. A multiscale frequency domain causal framework for enhanced pathological analysis. In *International Conference on Learning Representations*, volume 2025, pp. 13038–13054, 2025a.

Yuning Cui, Syed Waqas Zamir, Salman Khan, Alois Knoll, Mubarak Shah, and Fahad Khan. Adair: Adaptive all-in-one image restoration via frequency mining and modulation. In *International Conference on Learning Representations*, volume 2025, pp. 101306–101327, 2025b.

Lidia Garrucho, Kaisar Kushibar, Claire-Anne Reidel, Smriti Joshi, Richard Osuala, Apostolia Tsirikoglou, Maciej Bobowicz, Javier Del Riego, Alessandro Catanese, Katarzyna Gwoździewicz, Maria-Laura Cosaka, Pasant M. Abo-Elhoda, Sara W. Tantawy, Shorouq S. Sakrana, Norhan O. Shawky-Abdelfatah, Amr Muhammad Abdo Salem, Androniki Kozana, Eugen Divjak, Gordana Ivanac, Katerina Nikiforaki, Michail E. Klontzas, Rosa García-Dosdá, Meltem Gulsun-Akpinar, Oğuz Lafcı, Ritse Mann, Carlos Martín-Isla, Fred Prior, Kostas Marias, Martijn P. A. Starmans, Fredrik Strand, Oliver Díaz, Laura Igual, and Karim Lekadir. A large-scale multicenter breast cancer DCE-MRI benchmark dataset with expert segmentations. *Scientific Data*, 12(1):453, 2025. doi: 10.1038/s41597-025-04707-4.

Sebastian Ibarra, Javier del Riego, Alessandro Catanese, Julian Cuba, Julian Cardona, Nataly Leon, Jonathan Infante, Karim Lekadir, Oliver Diaz, and Richard Osuala. Comparing conditional diffusion models for synthesizing contrast-enhanced breast mri from pre-contrast images. In *Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care*, pp. 226–236. Springer, 2025.

Jin Joo Kim, Jin You Kim, Hie Bum Suh, Lee Hwangbo, Nam Kyung Lee, Suk Kim, Ji Won Lee, Ki Seok Choo, Kyung Jin Nam, Taewoo Kang, et al. Characterization of breast cancer subtypes based on quantitative assessment of intratumoral heterogeneity using dynamic contrast-enhanced and diffusion-weighted magnetic resonance imaging. *European Radiology*, 32(2):822–833, 2022.

Daniel M Lang, Richard Osuala, Veronika Spieker, Karim Lekadir, Rickmer Braren, and Julia A Schnabel. Temporal neural cellular automata: Application to modeling of contrast enhancement in breast mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 604–614. Springer, 2025.

Yanting Liang, Zhitao Wei, Yi Dai, Xiaobo Chen, Siyao Du, Chinting Wong, Zeyan Xu, Weibo Gao, Chu Han, Kexin Chen, et al. An interpretable ai system reduces false-positive mri diagnoses by stratifying high-risk breast lesions. *Nature Communications*, 17(1):2263, 2026.

Luyang Luo, Mingxiang Wu, Mei Li, Yi Xin, Qiong Wang, Varut Vardhanabhuti, Winnie C. W. Chu, Zhenhui Li, Juan Zhou, Pranav Rajpurkar, and Hao Chen. A large model for non-invasive and personalized management of breast cancer from multiparametric MRI. *Nature Communications*, 16(1):3647, 2025. doi: 10.1038/s41467-025-58798-z.

Ritse M Mann, Nariya Cho, and Linda Moy. Breast mri: state of the art. *Radiology*, 292(3):520–536, 2019.

Richard Osuala, Daniel M. Lang, Preeti Verma, Smriti Joshi, Apostolia Tsirikoglou, Grzegorz Skorupko, Kaisar Kushibar, Lidia Garrucho, Walter H. L. Pinaya, Oliver Diaz, Julia A. Schnabel, and Karim Lekadir. Towards learning contrast kinetics with multi-condition latent diffusion models. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15005 of *Lecture Notes in Computer Science*, pp. 713–723. Springer, 2024. doi: 10.1007/978-3-031-72086-4_67.

Richard Osuala, Smriti Joshi, Apostolia Tsirikoglou, Lidia Garrucho, Walter HL Pinaya, Daniel M Lang, Julia A Schnabel, Oliver Diaz, and Karim Lekadir. Simulating dynamic tumor contrast enhancement in breast mri using conditional generative adversarial networks. *Journal of Medical Imaging*, 12(S2):S22014–S22014, 2025.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Hannes Schreiter, Jessica Eberle, Lorenz A. Kapsner, Dominique Hadler, Sabine Ohlmeyer, Ramona Erber, Julius Emons, Frederik B. Laun, Michael Uder, Evelyn Wenkel, Sebastian Bickelhaupt, and Andrzej Liebert. Virtual dynamic contrast enhanced breast mri using 2d u-net architectures. In *Lecture Notes in Computer Science*, volume 15451, pp. 85–95. Springer, 2025. doi: 10.1007/978-3-031-77789-9_9.

Yejee Shin, Yeeun Lee, Hanbyol Jang, Geonhui Son, Hyeongyu Kim, and Dosik Hwang. Anatomical consistency and adaptive prior-informed transformation for multi-contrast mr image synthesis via diffusion model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30918–30927. IEEE, 2025.

Wenxuan Wang, Jing Wang, Chen Chen, Jianbo Jiao, Yuanxiu Cai, Shanshan Song, and Jiangyun Li. Fremim: Fourier transform meets masked image modeling for medical image segmentation. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7845–7855. IEEE, 2024.

Hantao Zhang, Yuhe Liu, Jiancheng Yang, Shouhong Wan, Xinyuan Wang, Wei Peng, and Pascal Fua. Lefusion: Controllable pathology synthesis via lesion-focused diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 13232–13253, 2025.

Jiadong Zhang, Zhiming Cui, Zhenwei Shi, Yingjia Jiang, Zhiliang Zhang, Xiaoting Dai, Zhenlu Yang, Yuning Gu, Lei Zhou, Chu Han, Xiaomei Huang, Chenglu Ke, Suyun Li, Zeyan Xu, Fei Gao, Luping Zhou, Rongpin Wang, Jun Liu, Jiayin Zhang, Zhongxiang Ding, Kun Sun, Zhenhui Li, Zaiyi Liu, and Dinggang Shen. A robust and efficient AI assistant for breast tumor segmentation from DCE-MRI via a spatial-temporal framework. *Patterns*, 4(9):100826, 2023a. doi: 10.1016/j.patter.2023.100826.

Tianyu Zhang, Luyi Han, Anna D'Angelo, Xin Wang, Yuan Gao, Chunyao Lu, Jonas Teuwen, Regina Beets-Tan, Tao Tan, and Ritse Mann. Synthesis of contrast-enhanced breast mri using t1-and multi-b-value dwi-based hierarchical fusion network with attention mechanism. In *International conference on medical image computing and computer-assisted intervention*, pp. 79–88. Springer, 2023b.

Chen Zhao, Zhizhou Chen, Yunzhe Xu, Enxuan Gu, Jian Li, Zili Yi, Qian Wang, Jian Yang, and Ying Tai. From zero to detail: Deconstructing ultra-high-definition image restoration from progressive spectral perspective. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17935–17935. IEEE, 2025.

R Zhao, WJ Ma, J Tang, YZ Chen, LN Zhang, H Lu, and PF Liu. Heterogeneity of enhancement kinetics in dynamic contrast-enhanced mri and implication of distant metastasis in invasive breast cancer. *Clinical Radiology*, 75(12):961–e25, 2020.

Qikui Zhu, Hao Wu, Yanyan Zhang, and Shuo Li. Causality-driven spatio-temporal generator for multi-phase contrast-enhanced ct synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 110–119. Springer, 2025.

A Architecture and Implementation Details

We follow the notation introduced in Sec. 3. Each application of $\mathcal{G}$ uses the one-dimensional binomial kernel $g = [1, 4, 6, 4, 1]/16$ along all three spatial dimensions. The resulting $5 \times 5 \times 5$ separable filter is applied with two-voxel replicate padding, so all scale-space components preserve the input size.

For temporal encoding, the absolute time and interval are normalized as $\widetilde{t}_k = t_k/480$ and $\widetilde{\delta t}_k = (t_k - t_{k-1})/480$, with $t_0 = 0$. Each is expanded by $\{\sin(2^a \pi x), \cos(2^a \pi x)\}_{a=0}^{7}$ and concatenated with the raw normalized values and $(k-1)/(\tau-1)$. A two-layer multilayer perceptron maps the resulting vector to the 64-dimensional $e(t_k)$. The scale and shift in FiLM are produced by a linear projection of this embedding.

To obtain A^0, x^1, x^2, and x^3 are projected to a common channel dimension, trilinearly upsampled to the resolution of x^0, concatenated with x^0, and decoded with a sigmoid. For refinement, each pair $(F_L(t_k), \widehat{\Delta}_k^{G,L})$ is encoded before the phase aggregation in Eq. 6. The aggregate is concatenated at the resolution of x^1 with x^1, projected x^0, and resized A^0; the output is then upsampled and passed through a sigmoid to obtain A^1. Both maps remain soft during synthesis.

The uptake codebook in Eq. 4 contains 64 learnable 32-dimensional codewords. In both training and inference, each normalized encoder vector is replaced by its maximum-cosine-similarity codeword. Soft assignment probabilities with temperature 0.07 are used only to regularize codeword usage and do not replace the hard forward assignment.

B Complete Training Objectives

B.1 Spatial Regions and Schedules

The valid support is $\Omega = \{\mathbf{v} : |P_0(\mathbf{v})| > 10^{-6}\}$. For phase k, the normal enhancing and non-enhancing non-lesion regions are

$$\mathcal{E}_k = (\Omega \setminus A) \cap \{\Delta_k > 0.02\}, \qquad \mathcal{N}_k = (\Omega \setminus A) \cap \{\Delta_k \le 0.02\}. \tag{10}$$

The boundary band used in Eq. 8 is $\partial A = \mathrm{Dilate}_5(A) - \mathrm{Erode}_5(A)$, where the subscript denotes a $5 \times 5 \times 5$ structuring element.

All curricula use

$$s(\eta; a, b) = \mathrm{clip}\left(\frac{\eta - a}{b - a}, 0, 1\right), \tag{11}$$

where η is the training epoch. Specifically, $s_M = s(\eta; 8, 13)$, $s_H = s(\eta; 8, 15)$, $s_{\mathrm{het}} = s(\eta; 10, 25)$, and $s_F = s(\eta; 15, 30)$. The error-aware codebook weighting is ramped from epochs 10 to 40. During the first three epochs, only localization objectives are optimized.

B.2 Synthesis and Module-Specific Objectives

The synthesis objective is

$$\mathcal{L}_{\mathrm{syn}} = \mathcal{L}_\Omega + \mathcal{L}_{\Omega \setminus A} + \mathcal{L}_\mathcal{E} + 2\mathcal{L}_A + 0.5 s_F \mathcal{L}_{\mathrm{FE}}, \tag{12}$$

where the first four terms apply Eq. 7 to the final enhancement over the indicated regions and average across phases. To define the one-sided false-enhancement term, let $o_k(\mathbf{v}) = \mathrm{ReLU}(\widehat{\Delta}_k(\mathbf{v}) - \Delta_k(\mathbf{v}) - 0.01)$ for $\mathbf{v} \in \mathcal{N}_k$. $\mathcal{L}_{\mathrm{FE}}$ is the mean of o_k over all valid voxels plus the mean of its largest 10% values, averaged across phases.

The error-aware vector-quantization loss is

$$\mathcal{L}_{\mathrm{VQ}} = \mathcal{L}_{\mathrm{code}} + 0.25\mathcal{L}_{\mathrm{commit}} + 0.01\mathcal{L}_{\mathrm{usage}}, \tag{13}$$

where $\mathcal{L}_{\mathrm{code}}$ and $\mathcal{L}_{\mathrm{commit}}$ are spatially weighted squared distances from the codeword to the detached encoder feature and vice versa. The weights are derived from detached synthesis error over Ω plus half the intratumoral deviation from the phase-wise lesion mean, Gaussian-smoothed, mean-normalized, clipped to 4, and multiplied by a maximum factor of 2 according to the codebook ramp.

The usage term is the Kullback–Leibler divergence between the weighted marginal code distribution and the uniform distribution.

The remaining module-specific terms are grouped as

$$\begin{aligned}\mathcal{L}_{\text{layer}} = 0.25\mathcal{L}_{\text{VQ}} + s_M\left(0.25\mathcal{L}_{\text{MP}} + 0.25\mathcal{L}_{\text{MV}}\right) \\ + s_H\left(0.5\mathcal{L}_{\text{HG}} + 0.02\mathcal{L}_{\text{HM}}\right).\end{aligned} \tag{14}$$

$\mathcal{L}_{\text{MP}}$ is the tumor-region correlation loss between the predicted and target Mid components, and $\mathcal{L}_{\text{MV}}$ matches their local variance computed in a $5 \times 5 \times 5$ window. $\mathcal{L}_{\text{HG}}$ matches three-dimensional gradients of the High component over Ω and A with relative weight 1:0.25, while $\mathcal{L}_{\text{HM}}$ matches its lesion-wise mean.

B.3 Spatiotemporal and Localization Objectives

The spatiotemporal objective is

$$\mathcal{L}_{\text{st}} = 0.25 s_{\text{het}}\mathcal{L}_{\text{het}} + \mathcal{L}_{\text{rate}} + \mathcal{L}_{\text{curve}}, \tag{15}$$

where $\mathcal{L}_{\text{het}}$ is the sum of lesion-wise standard-deviation error, robust dynamic-range error, and normalized multiscale spatial-pattern error. Let $\widehat{P}_0 = P_0$ and $d_k^t = (t_k - t_{k-1})/60$. Then $\mathcal{L}_{\text{rate}}$ applies the Charbonnier penalty to the difference between $(\widehat{P}_k - \widehat{P}_{k-1})/d_k^t$ and $(P_k - P_{k-1})/d_k^t$, while $\mathcal{L}_{\text{curve}}$ matches the lesion-mean enhancement at every target phase.

For each localization stage j, the loss is

$$\mathcal{L}_{\text{loc}}^j = \mathcal{L}_{\text{BCE}}^j + \mathcal{L}_{\text{Dice}}^j + \mathcal{L}_{\text{FP}}^j + 0.25\mathcal{L}_{\text{bd}}^j, \qquad \mathcal{L}_{\text{loc}} = 0.5\mathcal{L}_{\text{loc}}^0 + 0.5\mathcal{L}_{\text{loc}}^1. \tag{16}$$

Here, the false-positive term averages the predicted probability outside A, and the boundary term matches morphological gradients computed with a $3 \times 3 \times 3$ neighborhood. The frequency coefficients in Eq. 8 are $\lambda_G = 1$, $\lambda_L = 1$, $\lambda_M = 2$, and $\lambda_H = 1$.

AdamW uses a ReduceLROnPlateau scheduler with factor 0.5 and patience 8, gradient clipping at 5, and early stopping after 30 validation epochs without improvement.