MIRAGE: Multi-scale Lesion-Informed Representation with Auxiliary Guidance for MRI Contrast Enhancement

Borghesi, Andrea; Wang, Xin; Teuwen, Jonas; Yiasemis, George

Netherlands Cancer Institute
g.yiasemis@nki.nl

Abstract. Inferring contrast enhancement from one pre-contrast breast MRI slice is underdetermined: post-contrast appearance contains physiological information that is not uniquely encoded in baseline anatomy. Optimizing only paired pixel fidelity can suppress uncertain lesion enhancement, whereas adversarial or stochastic generative objectives can favor realistic post-contrast appearance without guaranteeing patient-specific lesion fidelity. We introduce MIRAGE, a residual 2D U-Net that combines global reconstruction and perceptual losses with three forms of lesion-aware supervision available only during training: an asymmetric penalty for missed tumor enhancement, multi-scale auxiliary tumor segmentation, and guidance through a frozen post-contrast tumor segmentation nnU-Net. We evaluate the method on 301 cases from the multicentre MAMA-SYNTH data using eight complementary image-, region-, radiomics-, and segmentation-based metrics. MIRAGE ranks first on six metrics and markedly improves downstream lesion localization over tuned pix2pix, conditional diffusion, and latent bridge-matching baselines. The generative alternatives retain advantages in LPIPS or contrast classification, revealing a clear fidelity-utility trade-off. Leave-one-in and leave-one-out ablations show that the losses are partly redundant for lesion localization but exert distinct effects on appearance, radiomics, and boundary accuracy. These results support task-aware synthesis while also showing that its apparent optimality is conditional on the downstream models and metrics used to define utility. Code and weights are publicly available at https://huggingface.co/NKI-AI/mirage-mama-synth.

Keywords: Breast MRI · DCE MRI · Virtual contrast enhancement · Image synthesis · Task-aware learning · Deep supervision

1 Introduction

Dynamic contrast-enhanced (DCE) MRI is the most sensitive breast-imaging modality and supports lesion detection, disease-extent assessment, treatment planning, and response monitoring [14]. Its value derives from signal changes after administration of a gadolinium-based contrast agent (GBCA). Although contemporary GBCAs have a strong safety record, contrast administration still

requires intravenous access, adds time and complexity, is restricted in selected patients, and contributes to environmental contamination [5, 22]. These considerations motivate contrast-reduced or contrast-free protocols, provided that clinically relevant enhancement information can be recovered reliably.

Virtual contrast enhancement predicts a post-contrast image y from non-contrast input x. Because enhancement depends on physiological factors not fully determined by pre-contrast anatomy, multiple targets may be compatible with the same input, particularly when only a single T1-weighted slice is available. Methods must therefore balance patient-specific fidelity with realistic post-contrast appearance, reflecting the perception–distortion trade-off [1]. In medical imaging, however, plausible synthesis may still add, shift, or suppress clinically relevant features [4].

Breast virtual-contrast studies have used multi-sequence inputs, simulated low-dose acquisitions, diffusion-weighted MRI, lesion-weighted adversarial objectives, and downstream segmentation guidance [3, 11, 15, 19, 25]. Conditional diffusion has also shown that subtraction targets and tumor-aware losses improve lesion fidelity, while mask conditioning further helps but requires lesion localization at inference [8]. These results suggest that global pixel or perceptual criteria alone are insufficient. Existing synthesis families make different compromises: pix2pix uses adversarial learning to encourage sharp target-domain appearance [10]; denoising diffusion models represent flexible conditional distributions through iterative sampling [7]; and latent bridge matching (LBM) enables one-step translation in latent space [2]. Direct paired regression is more tightly anchored to the observed anatomy but may average uncertain detail.

We investigate this tension under the MAMA-SYNTH protocol [18], which maps a single fat-suppressed pre-contrast slice to its peak-enhancement counterpart and evaluates global fidelity, tumor-region similarity, radiomic classification, and segmentation by a fixed post-contrast nnU-Net. We propose **MIRAGE** (**M**ulti-scale Lesion-**I**nformed **R**epresentation with **A**uxiliary **G**uidance for Contrast **E**nhancement), a residual U-Net trained with image-, lesion-, and task-aware supervision. Tumor masks guide training but are not supplied at inference. Our contributions are as follows:

(i) A lesion-informed objective combining asymmetric enhancement preservation, multi-scale segmentation, and frozen downstream-model guidance;
(ii) Complementary leave-one-in and leave-one-out analyses that distinguish isolated benefit from redundancy in the full objective; and
(iii) A comparison with adversarial, diffusion, and latent-bridge alternatives, exposing the trade-off between appearance realism and lesion utility.

2 Materials and Methods

2.1 Data and task

We use the MAMA-SYNTH release derived from MAMA-MIA [6, 18]. It contains 1,506 pretreatment breast DCE-MRI examinations from four public cohorts (Duke [21], ISPY1 [17], ISPY2 [13], NACT [16]), acquired at 1.5 or 3.0 T

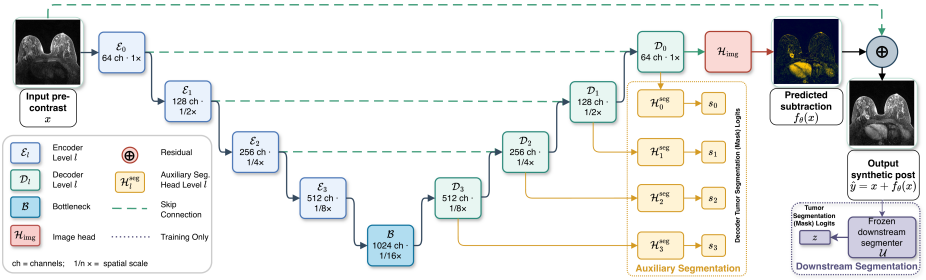

Fig. 1. MIRAGE architecture. A four-scale U-Net with bottleneck $\mathcal{B}$ and skip connections $\mathcal{E}_l \to \mathcal{D}_l$ maps pre-contrast x to synthetic post $\hat{y} = x + f_\theta(x)$, where the image head $\mathcal{H}_{\text{img}}$ predicts the subtraction enhancement $f_\theta(x)$ and a residual connection adds it back to x. At inference, only $\hat{y}$ is returned. During training, segmentation heads on each decoder scale produce tumour-mask logits s_ℓ, supervised against the ground-truth ROI mask m, and a frozen challenge nnU-Net $\mathcal{U}$ segments $\hat{y}$ as z (after z-normalization) using the same mask for supervision. These segmentation paths are outlined with dotted lines and are discarded at inference; m never enters the synthesiser input.

across multiple vendors. For each examination, preprocessing selects the post-contrast phase with the highest mean signal inside the annotated tumor and the 2D slice with the largest tumor area. Each case comprises a fat-suppressed pre-contrast image x, its peak-enhancement target y, and a binary tumor mask m. A global mean and standard deviation estimated from training pre-contrast images are applied to both x and y, preserving their relative intensity change. We use 1,205 cases for training and a fixed 301-case internal validation cohort. Masks are never provided to the synthesizer at inference.

2.2 Architecture and objective

MIRAGE is a four-scale residual 2D U-Net [20] with encoder/decoder widths $[64, 128, 256, 512]$, a 1024-channel bottleneck $\mathcal{B}$, InstanceNorm, and LeakyReLU. Rather than predicting y directly, an 1×1 image head $\mathcal{H}_{\text{img}}$ on the finest decoder stage $\mathcal{D}_0$ reads out the predicted subtraction $f_\theta(x)$ and

$$\hat{y} = x + f_\theta(x), \tag{1}$$

with skip connections from each encoder $\mathcal{E}_l$ to $\mathcal{D}_l$, l=0, . . . , 3 (l=0 full res.).

For training-only supervision, four 1×1 segmentation heads $\mathcal{H}_l^{\text{seg}}$ attach to decoder stage $\mathcal{D}_l$, l=0, . . . , 3, producing tumor-mask logits at each scale s_l. A frozen downstream segmenter (see below) is applied likewise only during training. At inference the network receives only x and returns $\hat{y}$; $\mathcal{H}_l^{\text{seg}}$ and the segmenter path are discarded. The full architecture is shown in Fig. 1.
The full objective is

$$\mathcal{L} = \mathcal{L}_1 + 0.3\,\mathcal{L}_{\text{LPIPS}} + 0.05\,\mathcal{L}_{\text{under}} + 0.1\,\mathcal{L}_{\text{seg}} + 0.2\,\mathcal{L}_{\text{dseg}}, \tag{2}$$

grouped as fidelity, ROI realism, and segmentation. The coefficients were chosen through coarse development sweeps and define one operating point, not a universal optimum. With valid-pixel mask v, $\mathcal{L}_1 = \|v \odot (\hat{y} - y)\|_1 / \|v\|_1$. $\mathcal{L}_{\text{LPIPS}}$ is the VGG-based perceptual loss [24] on the reconstructed post-contrast image.

To penalise insufficient enhancement in the lesion, define $e = \text{relu}(y - x)$ and $\hat{e} = \text{relu}(\hat{y} - x)$. The asymmetric term

$$\mathcal{L}_{\text{under}} = \|m \odot v \odot \text{relu}(e - \hat{e})\|_1 \,/\, \|m \odot v\|_1 \tag{3}$$

penalises only ROI pixels where $\hat{e} < e$; over-enhancement inside the tumor is not penalised by this term and is handled by the global reconstruction losses.

For internal deep supervision, decoder head l predicts logits s_l against level-l targets m_l and v_l ($l = 0$ full resolution; $l > 0$ obtained by adaptive average pooling of m and v followed by binarization). $\mathcal{L}_{\text{seg}}$ is the uniform sum over four scales of masked binary cross-entropy plus soft Dice,

$$\mathcal{L}_{\text{seg}} = \sum_{l=0}^{3} \Big[\text{BCE}(s_l, m_l; v_l) + \text{Dice}(s_l, m_l; v_l) \Big]. \tag{4}$$

For external task guidance, let $\mathcal{U}$ be the released post-contrast nnU-Net [9] from MAMA-SYNTH [18], and let $\mathcal{N}(\cdot; v)$ denote per-image z-scoring over valid pixels v. With tumor logits $z = \mathcal{U}(\mathcal{N}(\hat{y}, v))$,

$$\mathcal{L}_{\text{dseg}} = 0.4\, \mathcal{L}_{\text{BCE}}(z, m; v) + 0.6\, \mathcal{L}_{\text{Dice}}(z, m; v). \tag{5}$$

$\mathcal{U}$ remains frozen, but gradients flow through $\hat{y}$. Training applies a single full-resolution forward pass on the padded canvas, without the evaluator's sliding-window inference, test-time mirroring, or full nnU-Net preprocessing pipeline; $\mathcal{L}_{\text{dseg}}$ is therefore a differentiable surrogate of, not an exact reproduction of, the deployed segmentation score.

2.3 Optimization, baselines, and evaluation

MIRAGE is trained with Adam, initial learning rate 5×10^{-4}, warm-up cosine decay, batch size 4, and 59,000 iterations on a single NVIDIA H100 GPU.

We compare with tuned paired-translation implementations of pix2pix [10], a conditional DDPM [7] predicting the subtraction image, and one-step LBM [2].

All models are evaluated on the validation cohort using the eight metrics defined by the MAMA-SYNTH protocol [18]. The metrics cover four complementary aspects of synthesis quality. Global image fidelity is measured with MSE and LPIPS [24], while tumor-region realism is assessed with tumor-masked SSIM [23] and Fréchet radiomic distance (FRD), which compares radiomic-feature distributions extracted from synthetic and reference tumor regions [12].

Two fixed radiomics classifiers provide task-oriented measures. AUROC_{C} evaluates whether the synthetic images exhibit a post-contrast enhancement phenotype by measuring their separability from pre-contrast images. AUROC_{R}

distinguishes the annotated tumor ROI from its mirrored contralateral counterpart and hence probes whether tumor-associated enhancement is present at the correct region. Both classifiers use the reference tumor mask and consequently assess enhancement characteristics rather than automatic tumor localization.

Downstream utility is evaluated with a released by MAMA-SYNTH single-fold 2D nnU-Net. Predictions on the synthetic images are compared with the reference tumor masks using Dice and the 95th-percentile Hausdorff distance (HD95). This evaluation uses the complete released nnU-Net preprocessing and inference pipeline, in contrast to the simplified differentiable forward pass used to compute $\mathcal{L}_{\text{dseg}}$ during training.

For experiments repeated across random seeds, results are reported as mean $\pm$ std. Differences between configurations are assessed using the two-sample standard error $\text{SE}_\Delta = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$, rather than the variability of either configuration alone. A difference is marked when its magnitude is at least $2, \text{SE}_\Delta$.

3 Results

Figure 2 shows failing and succeeding representative results from different cohorts. MIRAGE preserves the overall breast anatomy and concentrates the predicted enhancement within the annotated tumor region, although fine differences in intratumor texture and surrounding parenchymal enhancement remain visible relative to the reference.

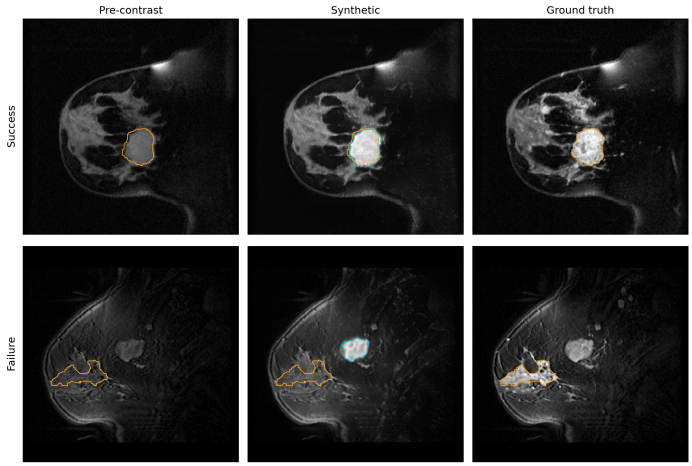

Fig. 2. Qualitative synthesis on two held-out cases from different cohorts, each of single-breast view. The solid contour indicates the expert annotated tumor, while the dashed indicates the segmenter prediction. The failing case lesion area is hardly distinguishible from the surrounding tissue: the synthetizer brightens up the correct area only by $\sim$23%, and prioritizes instead a region that was well defined already from the input.

3.1 Loss ablations

The contribution of each auxiliary objective is examined from two complementary directions. The leave-one-in experiments in Table 1 add each term individually to the masked-$\mathcal{L}_1$ baseline, whereas the leave-one-out experiments in Table 2 remove one term at a time from the complete objective.

Table 1. Leave-one-in ablation over three seeds. Each loss is added to the masked-L_1 baseline. *: mean difference > 2 combined standard errors relative to L_1.

	Image fidelity		ROI realism		Classification		Segmentation	
Recipe	MSE↓	LPIPS↓	FRD↓	SSIM_{tum}↑	AUROC_{C}↑	AUROC_{R}↑	Dice↑	HD95↓
L_1 **only**	$0.597_{\pm 0.009}$	$0.173_{\pm 0.003}$	$10.47_{\pm 0.36}$	$0.482_{\pm 0.017}$	$0.721_{\pm 0.009}$	$0.683_{\pm 0.005}$	$0.495_{\pm 0.038}$	$121.7_{\pm 17.6}$
$+\,\mathcal{L}_{\text{LPIPS}}$	$0.603_{\pm 0.004}$	$0.091^{*}_{\pm 0.002}$	$7.23^{*}_{\pm 0.12}$	$0.498_{\pm 0.013}$	$0.726_{\pm 0.006}$	$0.694^{*}_{\pm 0.007}$	$0.562^{*}_{\pm 0.007}$	$98.7^{*}_{\pm 3.8}$
$+\,\mathcal{L}_{\text{seg}}$	$0.568^{*}_{\pm 0.009}$	$0.172_{\pm 0.005}$	$10.50_{\pm 1.47}$	$0.543^{*}_{\pm 0.018}$	$0.660^{*}_{\pm 0.006}$	$0.712^{*}_{\pm 0.003}$	$0.635^{*}_{\pm 0.010}$	$56.5^{*}_{\pm 7.6}$
$+\,\mathcal{L}_{\text{dseg}}$	$0.596_{\pm 0.011}$	$0.164^{*}_{\pm 0.005}$	$7.77^{*}_{\pm 0.64}$	$0.511^{*}_{\pm 0.008}$	$0.795^{*}_{\pm 0.002}$	$0.681_{\pm 0.007}$	$0.634^{*}_{\pm 0.016}$	$58.1^{*}_{\pm 10.8}$
$+\,\mathcal{L}_{\text{under}}$	$0.599_{\pm 0.004}$	$0.170_{\pm 0.002}$	$9.81_{\pm 0.83}$	$0.554^{*}_{\pm 0.004}$	$0.698^{*}_{\pm 0.009}$	$0.713^{*}_{\pm 0.016}$	$0.632^{*}_{\pm 0.025}$	$54.9^{*}_{\pm 11.3}$

The masked-$\mathcal{L}_1$ model performs poorly on downstream segmentation. Adding any of the four auxiliary terms significantly improves both Dice and HD95. The three mask-aware terms reach Dice scores within 0.003 of one another, showing that each can independently recover a large part of the localization performance.

The terms differ more clearly outside the segmentation metrics. Adding either $\mathcal{L}_{\text{LPIPS}}$ or $\mathcal{L}_{\text{under}}$ improves image fidelity, tumor-masked SSIM, and radiomic agreement. Adding $\mathcal{L}_{\text{seg}}$ or $\mathcal{L}_{\text{dseg}}$ produces the largest gains in tumor localization and also improves tumor-region realism.

The complete objective achieves better segmentation performance than any leave-one-in configuration. In the leave-one-out analysis, removing $\mathcal{L}_{\text{seg}}$ is the only ablation that significantly reduces Dice. Removing $\mathcal{L}_{\text{dseg}}$ significantly worsens both FRD and HD95, while Dice remains essentially unchanged. Removing $\mathcal{L}_{\text{LPIPS}}$ worsens LPIPS and also increases FRD.

The effect of $\mathcal{L}_{\text{under}}$ differs between the two analyses. It provides clear improvements when added alone to the masked-$\mathcal{L}_1$ model, but removing it does not materially alter FRD, Dice, or HD95. Its remaining effect is a small improvement in SSIM_{tum} accompanied by a small increase in MSE.

All four leave-one-out configurations produce a higher mean AUROC_{C} than the complete objective. None of these differences reaches the 2σ criterion, however, and the complete model has the largest variability for this metric, with a standard deviation of ± 0.031.

3.2 Comparison across synthesis paradigms

Table 3 compares MIRAGE with pix2pix, conditional DDPM, and LBM on the same validation cohort and evaluation pipeline. The methods occupy different operating points across the metric groups. LBM achieves the best LPIPS, while

Table 2. Leave-one-out ablation over three seeds. Each loss is removed from the full objective. *: mean difference > 2 combined standard errors; †: complete seed separation without meeting that criterion.

	Image fidelity		ROI realism		Classification		Segmentation	
Recipe	MSE↓	LPIPS↓	FRD↓	SSIM_{tum}↑	AUROC_{C}↑	AUROC_{R}↑	Dice↑	HD95↓
Full	$0.607_{\pm 0.008}$	$0.095_{\pm 0.001}$	$6.56_{\pm 0.15}$	$0.540_{\pm 0.007}$	$0.711_{\pm 0.031}$	$0.718_{\pm 0.013}$	$0.679_{\pm 0.009}$	$34.8_{\pm 2.3}$
$-\mathcal{L}_{\text{dseg}}$	$0.590^{*}_{\pm 0.010}$	$0.093^{*}_{\pm 0.001}$	$6.98^{*}_{\pm 0.22}$	$0.553^{*}_{\pm 0.004}$	$0.712_{\pm 0.011}$	$0.712_{\pm 0.013}$	$0.667_{\pm 0.010}$	$45.5^{*}_{\pm 4.1}$
$-\mathcal{L}_{\text{seg}}$	$0.601_{\pm 0.013}$	$0.095_{\pm 0.001}$	$6.78_{\pm 0.17}$	$0.543_{\pm 0.005}$	$0.758_{\pm 0.035}$	$0.708_{\pm 0.003}$	$0.663^{*}_{\pm 0.004}$	$45.2^{*}_{\pm 3.1}$
$-\mathcal{L}_{\text{LPIPS}}$	$0.608_{\pm 0.009}$	$0.167^{*}_{\pm 0.001}$	$8.74^{\dagger}_{\pm 2.13}$	$0.521^{*}_{\pm 0.010}$	$0.732_{\pm 0.008}$	$0.735^{\dagger}_{\pm 0.009}$	$0.671^{\dagger}_{\pm 0.002}$	$36.1_{\pm 3.0}$
$-\mathcal{L}_{\text{under}}$	$0.581^{*}_{\pm 0.004}$	$0.094_{\pm 0.002}$	$6.79_{\pm 0.51}$	$0.530^{*}_{\pm 0.005}$	$0.718_{\pm 0.009}$	$0.717_{\pm 0.011}$	$0.674_{\pm 0.001}$	$34.4_{\pm 2.5}$

Table 3. Cross-paradigm comparison on the validation cohort. Average rank is the mean of the eight per-metric ranks and is descriptive rather than the benchmark's group-wise ranking.

	Image fidelity		ROI realism		Classification		Segmentation		
Model	MSE↓	LPIPS↓	FRD↓	SSIM_{tum}↑	AUROC_{C}↑	AUROC_{R}↑	Dice↑	HD95↓	Avg rank↓
MIRAGE (ours)	**0.607**	0.095	**6.56**	**0.540**	0.711	**0.718**	**0.679**	**34.8**	**1.50**
DDPM	0.795	0.122	9.07	0.389	**0.816**	0.632	0.495	122.4	3.13
pix2pix	0.628	0.169	11.75	0.497	0.755	0.677	0.509	106.2	2.88
LBM	0.612	**0.079**	8.12	0.529	0.792	0.657	0.493	126.6	2.50

DDPM obtains the highest AUROC_{C}. MIRAGE achieves the best downstream segmentation performance and the lowest FRD. All three alternative synthesis models perform worse than MIRAGE on Dice and HD95. They also produce higher FRD values, despite obtaining stronger results on selected appearance-oriented metrics. Thus, the models producing the most perceptually favorable or strongly contrast-like outputs do not produce the most accurate tumor localization or the closest tumor-region radiomic distribution.

4 Discussion and Conclusion

The ablations show that the auxiliary objectives provide overlapping but distinct constraints. All improve the masked-$\mathcal{L}_1$ baseline individually, while the complete objective achieves better downstream segmentation than any single-term addition. The final performance therefore cannot be attributed to one dominant loss.

The segmentation objectives are particularly complementary. Removing $\mathcal{L}_{\text{seg}}$ is the only ablation that significantly reduces Dice, indicating that direct multiscale supervision remains important for overlap-based localization. Removing $\mathcal{L}_{\text{dseg}}$, by contrast, primarily worsens HD95 and FRD while leaving Dice largely unchanged, suggesting that guidance from the frozen segmenter additionally constrains lesion boundaries and tumor-region appearance. Removing $\mathcal{L}_{\text{LPIPS}}$ likewise worsens both LPIPS and FRD. This association does not imply that LPIPS directly optimizes radiomic fidelity, but shows that perceptual and radiomic agreement are coupled in the present setting.

The clearest redundancy concerns $\mathcal{L}_{\text{under}}$. Although effective when added alone, it has little effect on FRD, Dice, or HD95 once the other losses are present. Its remaining improvement in SSIM_{tum} is accompanied by a small increase in MSE. The other lesion-aware objectives may therefore already provide much of the pressure needed to preserve tumor enhancement, making $\mathcal{L}_{\text{under}}$ the first candidate for removal when favoring a simpler objective.

The classification and cross-paradigm results reveal a broader trade-off. Removing any auxiliary term slightly increases AUROC_{C}, although none of these differences reaches the predefined 2σ threshold. Similarly, LBM achieves the best LPIPS and DDPM the highest AUROC_{C}, yet pix2pix, DDPM, and LBM all perform worse than MIRAGE on segmentation and FRD. Images that appear sharper or more characteristically post-contrast are therefore not necessarily better aligned with the patient-specific tumor location or radiomic distribution. The fixed classifier may reward a more stereotyped enhancement phenotype, whereas the lesion-aware losses emphasize spatial correspondence, although the observed AUROC_{C} trend remains tentative.

These differences are consistent with the respective objectives. Adversarial and generative distribution-learning methods encourage outputs that resemble the post-contrast domain, but realistic appearance alone does not guarantee correctly localised enhancement. The results concern the evaluated implementations and should not be interpreted as showing that adversarial, diffusion, or bridge-matching methods are intrinsically unsuitable.

The selected operating point is nevertheless conditional. Loss weights were chosen through coarse development sweeps on one internal validation cohort rather than exhaustive multi-objective optimisation, and different applications may prioritise different compromises between sharpness, localization, and downstream utility. Moreover, changes to the downstream model for $\mathcal{L}_{\text{dseg}}$ computation could alter the optimal objective and weighting.

Finally, none of the reported metrics establishes physiological correctness. Pixelwise metrics compare against one observed post-contrast acquisition, radiomic metrics capture selected image properties, and fixed classifiers and segmenters inherit their own biases. External cohorts, multiple acquisition protocols and downstream models, and reader studies are therefore needed to determine whether synthetic enhancement preserves diagnostically relevant findings without introducing, displacing, or suppressing lesions.

In conclusion, MIRAGE synthesizes peak post-contrast breast MRI from a single pre-contrast slice using tumor-informed supervision available only during training. It provides a stronger overall balance of patient-specific fidelity, tumor-region radiomic agreement, and downstream segmentation than the evaluated generative baselines, while conceding selected appearance-oriented metrics. More broadly, the findings support task-aware medical image synthesis in which pretrained downstream models guide outputs towards preserving clinically relevant information, rather than merely plausible appearance.

References

1. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6228–6237 (2018)
2. Chadebec, C., Tasar, O., Sreetharan, S., Aubin, B.: LBM: Latent bridge matching for fast image-to-image translation. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 29086–29098 (2025)
3. Chung, M., Calabrese, E., Mongan, J., Ray, K.M., Hayward, J.H., Kelil, T., Sieberg, R., Hylton, N., Joe, B.N., Lee, A.Y.: Deep learning to simulate contrast-enhanced breast mri of invasive breast cancer. Radiology **306**(3), e213199 (2022)
4. Cohen, J.P., Luck, M., Honari, S.: Distribution matching losses can hallucinate features in medical image translation. In: International conference on medical image computing and computer-assisted intervention. pp. 529–536. Springer (2018)
5. Dekker, H.M., Stroomberg, G.J., Van der Molen, A.J., Prokop, M.: Review of strategies to reduce the contamination of the water environment by gadolinium-based contrast agents. Insights into Imaging **15**(1), 62 (2024)
6. Garrucho, L., Kushibar, K., Reidel, C.A., Joshi, S., Osuala, R., Tsirikoglou, A., Bobowicz, M., Del Riego, J., Catanese, A., Gwoździewicz, K., et al.: A large-scale multicenter breast cancer dce-mri benchmark dataset with expert segmentations. Scientific data **12**(1), 453 (2025)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020)
8. Ibarra, S., del Riego, J., Catanese, A., Cuba, J., Cardona, J., Leon, N., Infante, J., Lekadir, K., Diaz, O., Osuala, R.: 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)
9. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
10. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 1125–1134 (2017)
11. Kim, E., Cho, H.H., Kwon, J., Oh, Y.T., Ko, E.S., Park, H.: Tumor-attentive segmentation-guided gan for synthesizing breast contrast-enhanced mri without contrast agents. IEEE journal of translational engineering in health and medicine **11**, 32–43 (2022)
12. Konz, N., Osuala, R., Verma, P., Chen, Y., Gu, H., Dong, H., Chen, Y., Marshall, A., Garrucho, L., Kushibar, K., et al.: Fréchet radiomic distance (frd): A versatile metric for comparing medical imaging datasets. Medical Image Analysis p. 103943 (2026)
13. Li, W., Newitt, D.C., Gibbs, J., Wilmes, L.J., et al.: I-spy 2 breast dynamic contrast enhanced mri (i-spy2 trial) (2022). https://doi.org/10.7937/TCIA.D8Z0-9T85, https://www.cancerimagingarchive.net/collection/ispy2/
14. Mann, R.M., Cho, N., Moy, L.: Breast mri: state of the art. Radiology **292**(3), 520–536 (2019)
15. Müller-Franzes, G., Huck, L., Tayebi Arasteh, S., Khader, F., Han, T., Schulz, V., Dethlefsen, E., Kather, J.N., Nebelung, S., Nolte, T., et al.: Using machine learning to reduce the need for contrast agents in breast mri through synthetic images. Radiology **307**(3), e222211 (2023)

16. Newitt, D., Hylton, N.: Single site breast DCE-MRI data and segmentations from patients undergoing neoadjuvant chemotherapy (2016). https://doi.org/10.7937/K9/TCIA.2016.QHsyhJKy, https://doi.org/10.7937/K9/TCIA.2016.QHsyhJKy, [Data set]
17. Newitt, D., Hylton, N., on behalf of the I-SPY 1 Network and ACRIN 6657 Trial Team: Multi-center breast DCE-MRI data and segmentations from patients in the I-SPY 1/ACRIN 6657 trials (2016). https://doi.org/10.7937/K9/TCIA.2016.HdHpgJLK, https://doi.org/10.7937/K9/TCIA.2016.HdHpgJLK
18. Osuala, R., Joshi, S., van Dijk, J., Han, L., Cosaka, M.L., Mysler, D., Garrucho, L., Lekadir, K., Balocco, S., Diaz, O.: The mama-synth challenge: Synthesizing virtual contrast-enhancement in breast mri (Apr 2026). https://doi.org/10.5281/zenodo.19852228, https://doi.org/10.5281/zenodo.19852228
19. Osuala, R., Joshi, S., Tsirikoglou, A., Garrucho, L., Pinaya, W.H., Lang, D.M., Schnabel, J.A., Diaz, O., Lekadir, K.: Simulating dynamic tumor contrast enhancement in breast MRI using conditional generative adversarial networks. Journal of Medical Imaging **12**(S2), S22014 (2025)
20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Proc. Med. Image Comput. Comput. Assist. Interv. (MICCAI). pp. 234–241 (2015)
21. Saha, A., Harowicz, M.R., Grimm, L.J., Weng, J., Cain, E.H., Kim, C.E., Ghate, S.V., Walsh, R., Mazurowski, M.A.: Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations (2021). https://doi.org/10.7937/TCIA.e3sv-re93, https://doi.org/10.7937/TCIA.e3sv-re93, [Data set]
22. Starekova, J., Pirasteh, A., Reeder, S.B.: Update on gadolinium-based contrast agent safety, from the ajr special series on contrast media. American Journal of Roentgenology **223**(3), e2330036 (2024)
23. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
24. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 586–595 (2018)
25. Zhang, T., Han, L., D'Angelo, A., Wang, X., Gao, Y., Lu, C., Teuwen, J., Beets-Tan, R., Tan, T., Mann, R.: 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 (2023)