HDRSNET: LEARNING HEMODYNAMIC REPRESENTATIONS VIA RESIDUAL DIFFUSION FOR TUMOR SEGMENTATION FROM NON-CONTRAST MRI

Anonymous authors
Paper under double-blind review

ABSTRACT

Dynamic contrast-enhanced MRI (DCE-MRI) provides rich hemodynamic information for breast tumor segmentation, but its clinical use is often limited by contrast-agent risks, acquisition cost, and incomplete temporal sampling—sometimes with only a pre-contrast volume available. Existing generate-then-segment methods synthesize missing post-contrast phases before segmentation, but synthesis artifacts and over-smoothed enhancement dynamics can compromise downstream performance. To address this challenge, we propose HDRSNet, a Hemodynamic Diffusion Representation Network that learns task-aligned hemodynamic representations for accurate breast tumor segmentation from a single non-contrast MRI volume. During training, HDRSNet exploits complete DCE-MRI sequences to learn enhancement dynamics as residual variations relative to the pre-contrast baseline. Specifically, a residual diffusion module incorporates recurrent modeling and local decoupling to suppress background-driven noise accumulation and capture tumor-relevant enhancement patterns. The resulting hemodynamic representations are integrated into a segmentation backbone through multi-scale task-aligned conditioning, enabling the model to exploit learned dynamic cues when only a non-contrast volume is available at inference. Experiments on multiple DCE-MRI benchmarks demonstrate that HDRSNet outperforms state-of-the-art 4D models requiring complete temporal sequences and generate-then-segment pipelines. These results highlight task-aligned hemodynamic representation learning as an effective alternative to explicit image synthesis for tumor segmentation under missing contrast-enhanced observations. The source code is available at: https://anonymous.4open.science/r/HDRSNet-6A42.

1 INTRODUCTION

DCE-MRI is important for breast tumor diagnosis and segmentation because it captures temporal enhancement dynamics that provide cues beyond static anatomy (Ottens et al., 2022), as shown in Fig. 1(a). Motivated by this, many DCE-MRI-based tumor segmentation methods have been proposed and achieve strong performance by modeling spatiotemporal information (Chen et al., 2024b; Zhao et al., 2024; Zheng et al., 2025). However, complete DCE-MRI is often hard to obtain due to contrast-agent safety concerns, long scan time, protocol complexity, and cost, so sequences can be sparse or partially missing (sometimes only the pre-contrast MRI volume is available), which degrades models that rely on full temporal cues (Li et al., 2026).

A common way to handle missing phases is to first synthesize post-contrast images and then run a standard multi-phase segmentation model (Fig.1(b)) (Fan et al., 2025). However, such generate-then-segment pipelines can be fragile: small hallucinated textures or high-frequency artifacts in the synthesized volumes may be amplified by temporal fusion, and higher reconstruction fidelity does not always improve segmentation—especially in DCE-MRI where tumor segmentation depends on localized enhancement dynamics rather than global appearance (Xu et al., 2025; Wei et al., 2025; Chen et al., 2023). Recent studies show that intermediate features learned along the diffusion denoising trajectory can encode semantic and structural cues that are useful for segmentation, providing complementary guidance beyond image reconstruction (Zhang et al., 2024).

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.

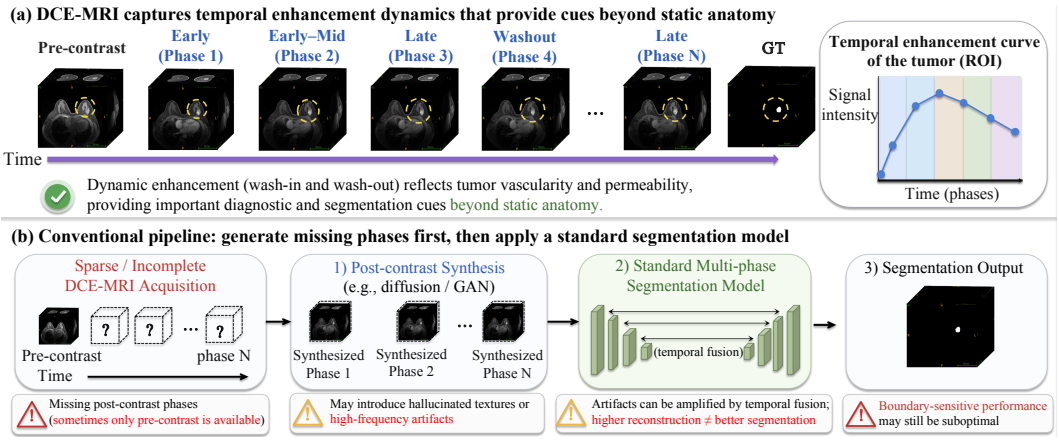

Figure 1: (a) DCE-MRI provides hemodynamic cues beyond static anatomy for tumor segmentation. (b) Generate-then-segment pipelines synthesize missing post-contrast phases, but synthesis artifacts and over-smoothed dynamics can impair downstream segmentation.

Building on this insight, we propose HDRSNet, a diffusion-based framework that learns task-aligned hemodynamic representations for breast tumor segmentation from a single non-contrast MRI volume at inference. Specifically, HDRSNet models temporal hemodynamic evolution as residual dynamics relative to the pre-contrast baseline, and introduces recurrent diffusion with local decoupling and dynamic blending to suppress background-driven artifacts and stabilize hemodynamic representation learning. The learned task-aligned representations are further integrated into a segmentation backbone via a lightweight multi-scale temporal–spatial conditioning module, enabling robust tumor segmentation from incomplete DCE-MRI observations. Experiments on the DUKE and I-SPY2 datasets under severe missing-phase settings demonstrate that HDRSNet consistently outperforms generate-then-segment pipelines and surpasses strong multi-phase models requiring complete temporal sequences.

The main contributions of this paper are summarized as follows:

- We find that diffusion-derived hemodynamic representations are more beneficial for breast tumor segmentation than directly using synthesized post-contrast images, revealing the importance of task-aligned dynamics learning for incomplete DCE-MRI analysis.
- We propose HDRSNet, a diffusion-based framework that learns task-aligned hemodynamic features from incomplete DCE-MRI via recurrent diffusion with local decoupling to facilitate accurate breast tumor segmentation.
- We introduce a dynamic blending strategy and a lightweight multi-scale temporal–spatial conditioning mechanism to stabilize recurrent diffusion learning and effectively inject task-aligned dynamics into segmentation.
- Extensive experiments on multiple DCE-MRI benchmarks show that HDRSNet outperforms both generate-then-segment pipelines and strong 4D segmentation models requiring complete multi-phase sequences, using only a single pre-contrast MRI at inference.

2 Related Work

2.1 Breast Tumor Segmentation in DCE-MRI

Breast tumor segmentation in DCE-MRI has gained increasing attention because DCE-MRI captures lesion-specific enhancement patterns beyond static anatomy (Garrucho et al., 2024). Segmentation approaches are commonly categorized into three types based on the utilization of temporal dynamics. The first approach uses a maximum-intensity projection (MIP) of the DCE-MRI sequence for 3D segmentation, which is efficient but loses phase ordering and subtle dynamics (Huang et al., 2023;

Wu et al., 2024). The second approach integrates both pre-contrast and a single post-contrast MRI to capture phase differences explicitly, aiming to reduce false positives and improve boundary localization while requiring only sparse acquisition (Wang et al., 2021; Zhou et al., 2022). The third approach leverages the whole DCE-MRI sequence for segmentation by modeling 4D spatial and temporal information. (He et al., 2024; Yang & Wang, 2025; Zhou et al., 2024). However, these approaches generally rely on complete or partially post MRI sequences, and their performance degrades noticeably when post phases are missing or unavailable at inference.

2.2 Diffusion Models for Incomplete Medical Imaging

Incomplete medical imaging is common in clinical practice due to limited acquisition time, motion artifacts, or contrast-agent risks, resulting in missing modalities or incomplete temporal observations. Recently, diffusion models have been widely explored for incomplete medical imaging, mainly in two directions: (i) image synthesis/completion (Chen et al., 2024c; Lin et al., 2024; Chen et al., 2024a), and (ii) task-oriented representation learning (Shi et al., 2024; Puglisi et al., 2024; Osuala et al., 2024; Han et al., 2025). Most existing methods follow a generate-then-analyze paradigm, where diffusion models first synthesize missing modalities or temporal phases and then perform downstream segmentation or diagnosis (Rahman et al., 2023; Zbinden et al., 2023). However, such pipelines can be fragile, since hallucinated textures, over-smoothed structures, or phase inconsistencies in synthesized images may degrade downstream dense prediction performance (Wei et al., 2025; Chen et al., 2023). This issue is particularly critical in DCE-MRI, where tumor segmentation depends more on localized hemodynamic enhancement dynamics than global appearance similarity. Meanwhile, recent studies show that intermediate denoising features in diffusion models contain rich semantic and structural information that can serve as effective representations for dense prediction tasks (Zhang et al., 2024; Ramanarayanan et al., 2024). Motivated by this insight, we propose to directly learn task-aligned hemodynamic representation from the diffusion process instead of explicitly synthesizing missing post-contrast phases.

3 Methods

3.1 Problem Setup and Notation

Problem setup. Each DCE-MRI sequence is denoted as $\mathbf{V} = \{V_0, V_1, \cdots, V_T\}$, where V_0 is the pre-contrast volume and $V_i (i \geq 1)$ are post-contrast phases. The tumor mask is Y. During training, we have access to $(\mathbf{V}, Y)$; during inference, only V_0 is available and the goal is to predict $\hat{Y}$.

Overview. As shown in Fig. 2, the proposed HDRSNet consists of three coupled components: a Residual Hemodynamic Distillation (RHD) module, a Multi-scale Task-aligned Conditioning (MTC) module, and a segmentation network. Given V_0, RHD models the temporal evolution across phases and provides multi-phase multi-level hemodynamic representations $\mathbb{F}_H = \{\mathbf{hf}_i^l\}$, where $i \in \{1, \ldots, T\}$ indexes phase and $l \in \{1, \ldots, L\}$ indexes feature level. In parallel, the segmentation network takes V_0 as input and produces multi-level spatial features $\mathbb{F}_S = \{\mathbf{sf}^l\}_{l=1}^L$. At each level l, MTC module injects $\{\mathbf{hf}_i^l\}_{i=1}^T$ into $\mathbf{sf}^l$ to form fused features $\mathbf{ff}^l$, which are then decoded to output the final prediction $\hat{Y}$.

3.2 Residual Hemodynamic Distillation (RHD)

Recurrent conditional diffusion for phase dynamics. RHD models the phase-wise transition in a conditional diffusion manner, where the post-contrast volume at phase $i+1$ is generated conditioned on the preceding phase i. For each $i \in \{1, \ldots, T\}$, the denoising network $\epsilon_\theta(\cdot)$ learns to predict the noise added to the target phase $V_i \in \mathbb{R}^{H \times W \times D}$ under the condition of a preceding volume $\tilde{V}_{i-1} \in \mathbb{R}^{H \times W \times D}$.

During training, the forward noising process is applied to the ground-truth target phase V_i:

$$q\left(z_i^s \mid V_i\right) = \mathcal{N}\left(\sqrt{\bar{\alpha}_s}\, V_i, (1 - \bar{\alpha}_s) I\right), \tag{1}$$

while the reverse process is parameterized as

$$p_\theta\left(z_i^{s-1} \mid z_i^s,\, \tilde{V}_i\right) = \mathcal{N}\left(\mu_\theta(z_i^s, s, \tilde{V}_{i-1}),\, \sigma_s^2 I\right), \tag{2}$$

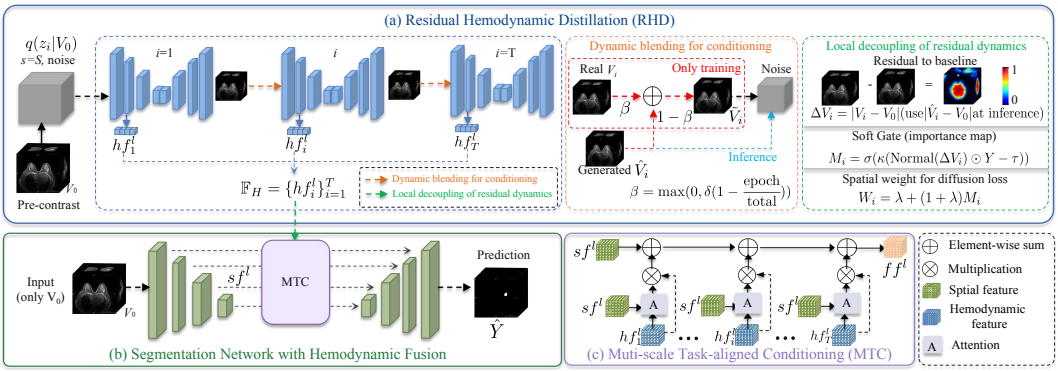

Figure 2: Overview of the proposed framework. (a) RHD learns multi-level hemodynamic representations through recurrent residual diffusion. (b) The segmentation network takes only V_0 as input and predicts the tumor mask with distilled hemodynamic guidance. (c) MTC injects multi-phase hemodynamic features into spatial features through cross-attention fusion.

where $\bar{\alpha}_s$ follows a predefined noise schedule, $s \in 0, \ldots, S$ denotes the diffusion step index. $\tilde{V}_{i-1}$ denotes a blended phase context constructed from either the ground-truth V_{i-1} and the synthesized estimate $\hat{V}_{i-1}$.

At inference time, the next-phase estimate $\hat{V}_i$ is obtained by running the reverse diffusion process starting from Gaussian noise $z_i^S \sim \mathcal{N}(0, I)$.

Dynamic blending to stabilize recurrent conditioning. Because RHD is recurrent across phases, using $\hat{V}_i$ as the sole condition early in training can cause compounding errors. We therefore construct the phase context via a scheduled blend of real and generated previous-phase volumes:

$$\tilde{V}_i = (1-\beta)\hat{V}_i + \beta V_i, \ \beta = \max\Big(0, \delta\big(1 - \frac{\text{epoch}}{\text{total}}\big)\Big), \tag{3}$$

where δ controls the initial weight of real conditioning and total specifies the portion of training over which blending is applied. This implements a curriculum from real-phase conditioning ($\tilde{V}_i \approx V_i$) to fully self-conditioned generation ($\tilde{V}_i \approx \hat{V}_i$). At inference, since only V_0 is available, we set $\tilde{V}_0 = V_0$ and use $\tilde{V}_i = \hat{V}_i$ for $i \geq 1$.

Local decoupling of residual dynamics. For DCE phase modeling, the information that drives segmentation is concentrated in localized enhancement changes rather than global appearance. To encourage RHD to focus its capacity on task-relevant dynamics, we formulate phase evolution as a residual process relative to the pre-contrast baseline V_0 : $\Delta V_i = V_i - V_0$. We then decouple dynamic regions (large residual magnitude) from background-stable regions (near-zero residual) using a soft spatial gate $M_i \in [0, 1]$, which serves as a per-voxel importance map for diffusion learning.

During training, we compute M_i from the residual magnitude and tumor mask Y,

$$M_i = \sigma\Big(\kappa\big(\text{Norm}(|V_i - V_0|) \odot Y - \tau\big)\Big), \tag{4}$$

where $\text{Norm}(\cdot)$ denotes per-volume normalization to $[0, 1]$, $\sigma(\cdot)$ is the sigmoid, $Y \in 0, 1$ is the tumor mask, τ is a soft threshold, and κ controls sharpness.

During inference, since V_i is unavailable, we use the predicted residual magnitude $|\hat{V}_i - V_0|$ to obtain M_i in the same form (computed after each phase is generated).

Instead of using a uniform noise-prediction loss, we apply a single spatial weight map

$$W_i = \lambda + (1-\lambda)\, M_i, \qquad \lambda \in (0, 1), \tag{5}$$

and optimize

$$\mathcal{L}_{\text{LD-diff}} = \mathbb{E}_{i,s,\epsilon}\Big[\big\| \sqrt{W_i} \odot \big(\epsilon - \epsilon_\theta(z_i^s, s, \tilde{V}_{i-1})\big)\big\|_2^2\Big]. \tag{6}$$

This form simultaneously emphasizes dynamic foreground regions while suppressing background-stable responses. Intuitively, voxels with strong enhancement dynamics ($M \approx 1$) receive full weight, while background-stable voxels ($M \approx 0$) are down-weighted by λ, discouraging the model from spending capacity on background residual hallucinations.

We apply the same gate to the diffusion features $\mathbf{hf}_i^l$ extracted from the denoising U-Net to separate dynamics from background:

$$\mathbf{hf}_i^l \leftarrow \mathbf{hf}_i^l \odot \mathrm{Down}(M_i), \tag{7}$$

where $\mathrm{Down}(\cdot)$ resizes M_i to the resolution of level l.

Importantly, the synthesized volumes $\hat{V}_i$ from RHD are not used for segmentation; instead, RHD acts as a dynamics encoder that extracts multi-level denoising features $\mathbb{F}_H = \{\mathbf{hf}_i^l\}(i = 1 \cdots T; l = 1 \cdots L)$ from its 3D encoder, which are fused into the segmentation network.

3.3 Segmentation with Hemodynamic Fusion

The segmentation component is a 3D U-Net that takes only the pre-contrast volume V_0 as input. The encoder produces a pyramid of spatial features $\mathbb{F}_S = \{\mathbf{sf}^l\}_{l=1}^L$, where l indexes the resolution level. To incorporate the distilled hemodynamic representations from Sec. 3.2, we inject the diffusion features $\{\mathbf{hf}_i^l\}_{i=1}^T$ into the segmentation stream at each skip connection via a Multi-scale Task-aligned Conditioning (MTC) mechanism, yielding fused features $\mathbf{ff}^l$ that are passed to the decoder, as illustrated in Fig. 2 (c).

Multi-scale Task-aligned Conditioning (MTC). For each level l, MTC treats $\mathbf{sf}^l$ as the spatial query and each phase-wise dynamics feature $\mathbf{hf}_i^l$ as the key/value source. We first compute a cross-attention map for phase i:

$$\mathbf{A}_i^l = \mathrm{Softmax}\left(\frac{(\mathbf{sf}^l W_q)\,(\mathbf{hf}_i^l W_k)^\top}{\sqrt{c}}\right), \tag{8}$$

where W_q, W_k are learned projections and c is the channel dimension used for scaling.

We then obtain the fused feature at level l by injecting the phase-wise diffusion dynamics into the spatial stream via residual cross-attention aggregation:

$$\mathbf{ff}^l = \mathbf{sf}^l + \frac{1}{T}\sum_{i=1}^{T} \mathbf{A}_i^l(\mathbf{hf}_i^l W_v), \tag{9}$$

where W_v is the value projection and $\mathbf{A}_i^l$ is the cross-attention map computed between $\mathbf{sf}^l(query)$ and $\mathbf{hf}_i^l(key)$ for phase i. This design performs multi-scale, task-aligned conditioning by allowing spatial features from V_0 to selectively aggregate hemodynamic across phases, while the residual connection preserves anatomical structure for stable decoding.

Decoder and prediction. The decoder follows the standard 3D U-Net design, progressively upsamples and combines the fused features $\{\mathbf{ff}^l\}_{l=1}^L$ via skip connections, and outputs the final tumor mask prediction $\hat{Y}$.

3.4 Optimization and Inference

Training objectives. We jointly optimize the residual hemodynamic distillation module (Sec. 3.2) and the segmentation network (Sec 3.3). For each phase transition $i \to i+1$ and diffusion step s, we sample $\epsilon \sim \mathcal{N}(0, I)$ and construct z_{i+1}^s by the forward process in Sec. 3.2. RHD is trained with the (locally decoupled) noise prediction objective

$$\mathcal{L}_{\text{LD-diff}} = \mathbb{E}_{i,s,\epsilon}\Big[\big|\sqrt{W_{i+1}} \odot \big(\epsilon - \epsilon_\theta(z_{i+1}^s, s, \tilde{V}_i)\big)\big|_2^2\Big], \tag{10}$$

where $\tilde{V}_i$ is the blended conditioning volume and $W_{i+1} = \lambda + (1-\lambda)M_{i+1}$ is the spatial weighting derived from the local decoupling gate M_{i+1}. The segmentation network is supervised by a standard Dice–cross-entropy loss,

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{Dice}}(\hat{Y}, Y) + \eta\mathcal{L}_{\text{CE}}(\hat{Y}, Y), \tag{11}$$

with η a fixed scalar. The overall objective is a weighted sum

$$\mathcal{L}_{\text{total}} = \gamma \frac{1}{T} \sum_{i=0}^{T-1} \mathcal{L}_{\text{LD-diff}}^{(i+1)} + (1-\gamma)\mathcal{L}_{\text{seg}}, \tag{12}$$

where $\gamma \in [0, 1]$ balances dynamics distillation and segmentation.

Inference. Given only V_0, we set $\tilde{V}_0 = V_0$ and recursively generate phases $\{\hat{V}_1, \ldots, \hat{V}_T\}$ using the reverse diffusion process. For $i \geq 1$, the conditioning volume is $\tilde{V}_i = \hat{V}_i$. At each phase, we extract multi-level diffusion features $\{\mathbf{hf}_i^l\}_{l=1}^L$ and fuse them with the segmentation encoder features $\{\mathbf{sf}^l\}_{l=1}^L$ through MTC to obtain $\{\mathbf{ff}^l\}_{l=1}^L$, which are decoded to produce the final mask $\hat{Y}$. When local decoupling is enabled at inference, we compute the gate M_i from the predicted residual magnitude $|\hat{V}_i - V_0|$ and apply it consistently with training to stabilize dynamics extraction. To reduce computation, we adopt accelerated sampling (e.g., DDIM) and use a fixed number of reverse steps S per phase in all experiments. **More implementation details are provided in the supplementary material (Implementation Details).**

4 Experiments

4.1 Experimental Setup

Datasets & Metrics. We evaluate on two public breast DCE-MRI benchmarks: DUKE (Saha et al., 2021) (291 sequences, T=4) and I-SPY2 (Li et al., 2020) (970 sequences, T=5), each containing a pre-contrast volume V_0, post-contrast phases $V_1, \ldots, V_T$, and voxel-level tumor annotations. Voxel intensities are clipped at the top/bottom 0.1% and normalized consistently across phases. Both datasets are split into training/validation/testing sets with a ratio of 7:1:2, shared across all methods. We report DSC (%), 95% HD (mm), and Sensitivity (%) for segmentation, and additionally report SSIM for methods involving post-contrast synthesis.

Implementation details. All models are implemented in PyTorch and trained on four NVIDIA RTX-3090Ti GPUs using Adam with a learning rate of 1×10^{-4}. Models are trained for up to 1000 epochs with batch size 1 per GPU and early stopping (patience 50). We use random crops of size $96 \times 96 \times 48$ during training and sliding-window inference with the same window size.

For residual hemodynamic distillation, the diffusion module is pretrained for the first 200 epochs using a linear noise schedule from 1×10^{-6} to 1×10^{-2} with 1000 diffusion steps. The denoising network adopts a 3D U-Net architecture with timestep embeddings injected at each scale. For efficient inference, DDIM sampling with a fixed number of reverse steps S is used for all phases. All results are averaged over five runs with different random seeds.

Table 1: Quantitative comparison of the proposed method with SOTA methods on both datasets.

Methods	Input	I-SPY2			DUKE			Params (M)	FLOPs (G)
		DSC (%)	95% HD (mm)	Sen (%)	DSC (%)	95% HD (mm)	Sen (%)		
nnUNet (Garrucho et al., 2024)	3D MIP	69.31±1.03	97.37±3.17	64.32±0.98	76.17±1.45	20.53±2.98	74.32±1.07	30.70	358.60
MedSegDiff-V2 (Wu et al., 2024)	3D MIP	73.81±3.16	88.94±5.41	70.93±3.08	79.21±2.87	15.35±4.19	80.08±2.31	48.00	983.00
Tumor-sen (Wang et al., 2021)	Two 3D	69.97±2.37	97.01±3.58	65.15±2.19	79.16±2.03	16.22±4.07	77.17±1.93	63.70	143.61
ALMN (Zhou et al., 2022)	Two 3D	67.12±3.17	104.18±5.45	62.17±3.03	80.19±2.01	15.19±4.15	78.54±2.85	53.60	2478.80
PLHN (Zhou et al., 2024)	4D	74.51±2.29	89.50±3.61	70.44±2.17	80.54±2.93	14.92±4.17	81.33±2.51	146.20	314.72
TSESNet (He et al., 2024)	4D	76.49±1.32	71.53±4.06	72.48±1.81	83.72±1.54	7.51±3.17	83.97±1.62	103.58	98.66
BBDiffSF (Feng et al., 2026)	4D	70.15±2.28	91.36±2.54	64.93±2.08	82.23±1.92	10.15±1.93	82.54±1.86	240.12	137.76
DEFNet (Yang & Wang, 2025)	4D	74.93±1.47	81.05±2.18	70.07±1.59	82.17±2.37	11.93±2.98	82.15±2.03	207.14	271.58
AKF-Net (Li et al., 2026)	4D	74.38±2.56	85.27±3.14	69.84±2.31	81.45±2.78	12.87±3.21	80.76±2.45	158.56	646.00
HDRSNet	3D	**78.54±1.19**	**67.95±1.87**	**77.41±1.16**	**84.93±1.67**	**6.87±2.17**	**85.61±1.35**	137.93	216.32

4.2 Comparisons with State-of-the-art methods

We compare our method with eight representative breast DCE-MRI tumor segmentation approaches, grouped by their input requirements: **(1) 3D MIP-based methods:** nnUNet and MedSegDiff-V2; **(2) Two-phase 3D methods (Pre- and Post-contrast MRI):** ALMN and Tumor-sen; and **(3) 4D-based methods (whole DCE-MRI sequence):** DEFNet, PLHN, TSESNet, AKF-Net, and BBDiffSF. All methods are evaluated under identical data splits and preprocessing. Quantitative results on I-SPY2 and DUKE are reported in Tab. 1, together with model size and FLOPs. We have the following observations:

(1) Our method achieves state-of-the-art performance on both datasets. This is mainly driven by residual hemodynamic distillation from V_0 and multi-scale task-aligned fusion. On I-SPY2, HDRSNet achieves 78.54% DSC / 67.95 mm HD, outperforming the strongest 4D baseline TSESNet (76.49% / 71.53 mm) and DEFNet (74.93% / 81.05 mm). On DUKE, it reaches 84.93% DSC / 6.87 mm HD, surpassing TSESNet (83.72% / 7.51 mm) and BBDiffSF (82.23% / 10.15 mm), showing that distilled dynamics can replace explicit multi-phase inputs.

(2) 4D models outperform 3D/two-3D baselines by explicitly modeling temporal enhancement. Full-sequence methods leverage contrast uptake trajectories, reducing ambiguity that remains with MIP or sparse phases. For example, on I-SPY2, TSESNet achieves 76.49% DSC, higher than nnUNet (69.31%) and Tumor-sen (69.97%). On DUKE, it reaches 83.72% DSC / 7.51 mm HD, outperforming nnUNet (76.17% / 20.53 mm) and ALMN (80.19% / 15.19 mm).

(3) HDRSNet offers a strong accuracy–efficiency trade-off. Instead of heavy 4D backbones or explicit completion, HDRSNet conditions a 3D U-Net using distilled dynamics, improving accuracy with moderate compute. It uses 216.32G FLOPs, far lower than ALMN (2478.80G) and MedSegDiff-V2 (983.00G), yet is more accurate on I-SPY2 (78.54 vs. 67.12/73.81 DSC) and DUKE (84.93 vs. 80.19/79.21 DSC).

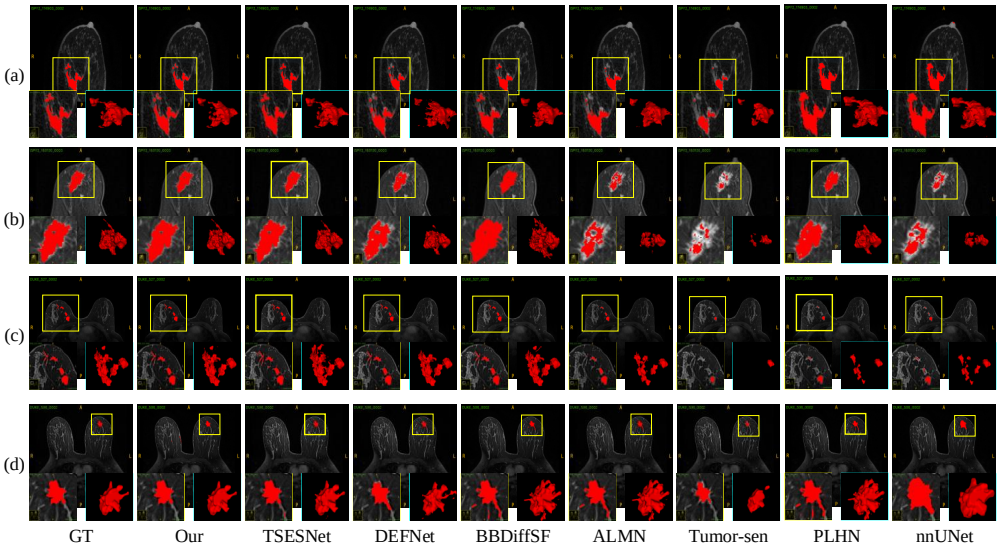

Figure 3: Qualitative results on I-SPY2 (a, b) and DUKE (c, d). Red indicates predicted tumors; GT is ground truth. Insets show zoomed regions (yellow box) and 3D segmentation views.

Qualitative Comparison:

Fig. 3 presents qualitative results on I-SPY2 (a,b) and DUKE (c,d). HDRSNet consistently produces masks closer to the GT, better preserving irregular boundaries and fine structures while reducing boundary leakage and missed lesions compared with existing 3D/4D methods. These results indicate that RHD and MTC provide effective task-relevant temporal cues even when only V_0 is available at inference.

4.3 Comparison with generate-then-segment pipelines

We compare our method with representative generate-then-segment pipelines on the I-SPY2 dataset. Specifically, we first report a fully supervised upper bound using TSESNet trained and tested with the complete DCE-MRI sequence. Then, for generation-based methods, we synthesize the full post-contrast sequence from the pre-contrast scan V_0 using different generative models (TeNCA, DCEtriformer, and SPHERE), and feed the generated sequences into the same TSESNet for segmentation. We evaluate generation quality using mean-SSIM and downstream segmentation using DSC, 95% HD, and Sensitivity (Tab. 2).

Table 2: Comparison of HDRSNet and generate-then-segment pipelines for DCE-MRI generation and tumor segmentation on the I-SPY2.

Methods	mean-SSIM	DSC (%)	95% HD (mm)	Sen (%)
TSESNet	-	76.49	71.53	72.48
+TeNCA (Lang et al., 2025)	0.8074	70.93	97.72	67.93
+DCEtriformer (Ramanarayanan et al., 2025)	0.8472	75.47	81.34	71.92
+SPHERE (Wang et al., 2025)	0.8632	74.17	90.05	70.15
HDRSNet	0.8561	78.54	67.95	77.41

Table 3: Effect of HDRSNet components on I-SPY2. DB: Dynamic Blending. LD: Local Decoupling.

Method	mean-SSIM	DSC (%)	95% HD (mm)	Sen (%)
DDPM (baseline)	0.3032	54.12	163.91	52.71
+DB	0.7810	74.02	79.84	73.28
+LD	0.4245	68.11	106.37	67.54
HDRSNet (DB+LD)	**0.8561**	**78.54**	**67.95**	**77.41**

(1) Higher synthesis quality does not necessarily translate to better segmentation. Although stronger generators yield higher mean-SSIM, the downstream gains are limited or even inconsistent. For example, SPHERE achieves the highest mean-SSIM (0.8632) but attains only 74.17% DSC, which is below the oracle TSESNet using real sequences (76.49%). DCEtriformer similarly reaches 0.8472 mean-SSIM but achieves 75.47% DSC, suggesting that improving global appearance does not guarantee preserving lesion-specific enhancement cues needed for accurate delineation.

(2) Task-aligned dynamics are more effective than directly segmenting synthesized sequences. Although generate-then-segment methods achieve high synthesis quality, artifacts and over-smoothed dynamics in synthesized phases can degrade downstream segmentation. for example, SPHERE achieves the highest mean-SSIM (0.8632) but only 74.17% DSC with 90.05 mm HD. In contrast, HDRSNet distills task-aligned hemodynamic from the diffusion process instead of directly segmenting synthesized images, achieving the best performance on I-SPY2 (78.54% DSC, 67.95 mm HD, 77.41% Sen), surpassing both generate-then-segment pipelines and the full-sequence TSESNet.

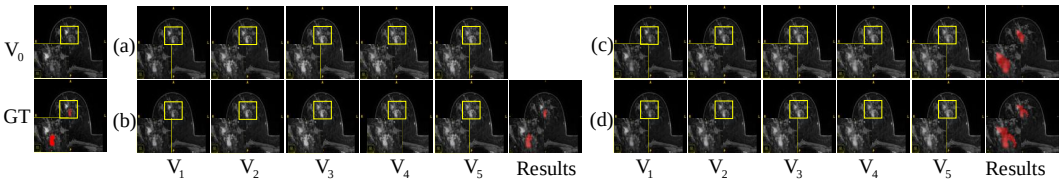

Figure 4: Qualitative comparison of DCE-MRI sequence generation on I-SPY2. Rows show (a) the real DCE-MRI sequence and sequences generated by (b) HDRNet, (c) SPHERE, and (d) DCEtriformer. Insets show enlarged views of the regions indicated by the yellow boxes. GT denotes the ground-truth tumor mask.

Qualitative analysis: Fig. 4 further supports these findings. While SPHERE and DCEtriformer largely maintain global appearance across phases (Fig. 4(c,d)), their zoomed views show over-smoothed or locally distorted lesion patterns, resulting in less accurate masks. In contrast, HDRSNet better preserves lesion-local structures across $V_1 - V_5$ and yields predictions closer to GT (Fig. 4(b)), consistent with its foreground-focused dynamics distillation.

4.4 Ablation Study

4.4.1 Effect of HDRSNet Components

To evaluate the contribution of each component in HDRSNet, we perform ablation studies on I-SPY2 by progressively adding Dynamic Blending (DB) and Local Decoupling (LD) to a phase-conditioned DDPM baseline. All variants share the same settings, and we report SSIM for synthesized post-contrast volumes and DSC/HD95/Sensitivity for downstream segmentation (Tab. 3).

(1) Dynamic blending is the primary driver of phase-consistent synthesis. Adding DB largely resolves recurrent drift by gradually transitioning from ground-truth conditioned training to self-conditioned generation, boosting mean-SSIM from 0.3032 to 0.7810 and improving DSC from 54.12% to 74.02%. This indicates that stabilizing the conditional trajectory across phases directly reduces error accumulation that otherwise contaminates later dynamics cues.

(2) Local decoupling mainly improves task-relevant fidelity near lesions. LD alone yields a smaller SSIM gain (0.3032→0.4245) but noticeably improves boundary-related metrics (DSC 68.11%, HD 106.37 mm) compared with DDPM. This supports our motivation that suppressing background residual energy reallocates diffusion capacity to foreground-relevant changes, which is more aligned with segmentation than global reconstruction.

(3) DB and LD are complementary and achieve the best segmentation together. Combining both components achieves the strongest overall results (0.8561 mean-SSIM, 78.54% DSC, 67.95 mm HD, 77.41% Sen), outperforming either single-component variant. This suggests that RHD benefits from (i) globally stable recurrent conditioning (DB) and (ii) localized, task-aligned dynamics emphasis (LD), producing dynamics features that are both coherent across phases.

4.4.2 Different FEATURE combinations in MTC Module

We ablate the MTC fusion on I-SPY2 by selectively injecting diffusion dynamics features $\mathbf{hf}^l$ from different depths (l=1, . . . , 4) into the segmentation backbone, while keeping all other settings fixed. Each configuration is evaluated with DSC, 95% HD, and Sensitivity (Tab. 4);

Table 4: Ablation study of MTC at different feature levels on I-SPY2. The best results among single-level fusion settings are highlighted in green, while the best overall results are highlighted in red.

$\mathbf{hf}^1$	$\mathbf{hf}^2$	$\mathbf{hf}^3$	$\mathbf{hf}^4$	DSC (%)	95% HD (mm)	Sen (%)
✓				65.74	137.56	64.18
	✓			69.81	105.32	68.12
		✓		71.53	85.43	70.85
			✓	75.53	76.12	75.21
✓	✓			72.19	87.31	71.12
		✓	✓	76.53	71.82	75.93
✓	✓	✓	✓	78.54	67.95	77.41

(1) Deeper dynamics features contribute more when used alone. As shown in Tab. 4, single-layer injection benefits from deeper features: $\mathbf{hf}^1$ achieves 65.74% DSC / 137.56 mm HD, while $\mathbf{hf}^4$ improves to 75.53% DSC / 76.12 mm HD. This indicates that deeper diffusion features capture more discriminative, tumor-relevant dynamics, leading to clearer boundary delineation.

(2) Multi-scale injection is complementary and yields the best results. Fusing multiple levels consistently outperforms single-layer settings, showing that dynamics cues at different scales are complementary. Notably, deeper pairs work better: $\mathbf{hf}^3 + \mathbf{hf}^4$ reaches 76.53% DSC vs. 72.19% for $\mathbf{hf}^1 + \mathbf{hf}^2$. Using all four scales gives the best result (78.54% DSC, 67.95 mm HD, 77.41% Sen).

More detailed analyses and discussions are provided in the supplementary material.

5 Conclusion

In this study, we proposed a novel solution for breast tumor segmentation under incomplete DCE-MRI acquisition, addressing critical clinical challenges arising from contrast-agent toxicity, cost, and missing temporal phases. By distilling task-aligned hemodynamic representations from a single pre-contrast MRI, our framework enables accurate tumor segmentation without relying on explicit post-contrast MRI synthesis. This design mitigates the error accumulation and artifact propagation often seen in generation-then-segmentation methods, and achieves performance that is competitive with SOTA 4D segmentation models that rely on complete DCE-MRI sequences. By reducing dependence on full DCE-MRI protocols, the proposed approach has the potential to lower imaging costs, shorten examination time, and improve patient safety, thereby facilitating more efficient and accessible clinical deployment of DCE-MRI–based diagnosis.

REFERENCES

Jiaqi Chen, Jiachen Lu, Xiatian Zhu, and Li Zhang. Generative semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7111–7120, 2023.

Qi Chen, Xiaoxi Chen, Haorui Song, Zhiwei Xiong, Alan Yuille, Chen Wei, and Zongwei Zhou. Towards generalizable tumor synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11147–11158, 2024a.

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, 2024b.

Tao Chen, Chenhui Wang, Zhihao Chen, Yiming Lei, and Hongming Shan. Hidiff: Hybrid diffusion framework for medical image segmentation. *IEEE Transactions on Medical Imaging*, 2024c.

Xueying Fan, Liming Xu, Bochuan Zheng, Xianhua Zeng, Weisheng Li, and Zhiwei Huang. Pre-to post-contrast medical image synthesis with outline-guide accelerate diffusion model. *Neural Networks*, pp. 107851, 2025.

Feiyan Feng, Tianyu Liu, Fulin Zheng, Yanshen Sun, and Hong Wang. Efficient breast cancer segmentation via brownian bridge diffusion with semantic fusion strategy. *Pattern Recognition*, pp. 112993, 2026.

Lidia Garrucho, Claire-Anne Reidel, Kaisar Kushibar, Smriti Joshi, Richard Osuala, Apostolia Tsirikoglou, Maciej Bobowicz, Javier del Riego, Alessandro Catanese, Katarzyna Gwoździewicz, et al. Mama-mia: A large-scale multi-center breast cancer dce-mri benchmark dataset with expert segmentations. *arXiv preprint arXiv:2406.13844*, 2024.

Andi Han, Wei Huang, Yuan Cao, and Difan Zou. On the feature learning in diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 51189–51256, 2025.

Jiezhou He, Xue Zhao, Zhiming Luo, Songzhi Su, Shaozi Li, and Guojun Zhang. Tsesnet: Temporal-spatial enhanced breast tumor segmentation in dce-mri using feature perception and separability. In *IJCAI*, pp. 803–811, 2024.

Rian Huang, Zeyan Xu, Yu Xie, Hong Wu, Zixian Li, Yanfen Cui, Yingwen Huo, Chu Han, Xiaotang Yang, Zaiyi Liu, et al. Joint-phase attention network for breast cancer segmentation in DCE-MRI. *Expert Systems with Applications*, 224:119962, 2023.

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.

Wen Li, David C Newitt, Jessica Gibbs, Lisa J Wilmes, Ella F Jones, Vignesh A Arasu, Fredrik Strand, Natsuko Onishi, Alex Anh-Tu Nguyen, John Kornak, et al. Predicting breast cancer response to neoadjuvant treatment using multi-feature mri: results from the i-spy 2 trial. *NPJ breast cancer*, 6 (1):63, 2020.

Xiangyu Li, Yifan Liu, Manxi Xu, Hongwei Yu, Jixin Luan, Aocai Yang, Bing Liu, Amir Shmuel, Yuan Zhen, Gongning Luo, et al. Anatomic-kinetic fusion network for pituitary gland and microadenoma segmentation in multi-phase dce-mri. *Information Fusion*, 125:103457, 2026.

Tianyu Lin, Zhiguang Chen, Zhonghao Yan, Weijiang Yu, and Fudan Zheng. Stable diffusion segmentation for biomedical images with single-step reverse process. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 656–666. Springer, 2024.

Richard Osuala, Daniel M Lang, Preeti Verma, Smriti Joshi, Apostolia Tsirikoglou, Grzegorz Skorupko, Kaisar Kushibar, Lidia Garrucho, Walter HL Pinaya, Oliver Diaz, et al. Towards learning contrast kinetics with multi-condition latent diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 713–723. Springer, 2024.

Tim Ottens, Sebastiano Barbieri, Matthew R Orton, Remy Klaassen, Hanneke WM van Laarhoven, Hans Crezee, Aart J Nederveen, Xiantong Zhen, and Oliver J Gurney-Champion. Deep learning dce-mri parameter estimation: Application in pancreatic cancer. *Medical image analysis*, 80: 102512, 2022.

Lemuel Puglisi, Daniel C Alexander, and Daniele Ravì. Enhancing spatiotemporal disease progression models via latent diffusion and prior knowledge. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 173–183. Springer, 2024.

Aimon Rahman, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, and Vishal M Patel. Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11536–11546, 2023.

Sriprabha Ramanarayanan, Arunima Sarkar, Matcha Naga Gayathri, Keerthi Ram, Mohanasankar Sivaprakasam, et al. Dce-diff: Diffusion model for synthesis of early and late dynamic contrast-enhanced mr images from non-contrast multimodal inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5174–5183, 2024.

Sriprabha Ramanarayanan, Kishore Kumar, Keerthi Ram, Harsh Agarwal, Ramesh Venkatesan, Mohanasankar Sivaprakasam, et al. Dcetriformer: A hybrid attention transformer for dce-mri synthesis in prostate imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1062–1071, 2025.

Ashirbani Saha, Michael R Harowicz, Lars J Grimm, Jingxi Weng, EH Cain, CE Kim, SV Ghate, R Walsh, and Maciej A Mazurowski. Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations [data set]. *The Cancer Imaging Archive*, 2021.

Yongyi Shi, Wenjun Xia, Ge Wang, and Xuanqin Mou. Blind ct image quality assessment using ddpm-derived content and transformer-based evaluator. *IEEE Transactions on Medical Imaging*, 43(10):3559–3569, 2024.

Dayang Wang, Srivathsa Pasumarthi Venkata, Ajit Shankaranarayanan, and Greg Zaharchuk. Staged and physics-grounded learning framework with hyperintensity prior for pre-contrast mri synthesis. In *Forty-second International Conference on Machine Learning*, 2025.

Shuai Wang, Kun Sun, Li Wang, Liangqiong Qu, Fuhua Yan, Qian Wang, and Dinggang Shen. Breast tumor segmentation in dce-mri with tumor sensitive synthesis. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Jinbao Wei, Yuhang Chen, Zhijie Wang, Gang Yang, Shimin Tao, Jian Gao, Aiping Liu, and Xun Chen. Rethinking diffusion bridge model with dual alignments for medical image synthesis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1052–1061, 2025.

Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6030–6038, 2024.

Xiang´ an Wu, Haoyang Su, Yiwei Hua, Yali Xu, Lilong Wang, Xiaosong Wang, Shitian Wang, Bao Jin, Xiao Liu, Xueshuai Wan, et al. A multi-phase ct dataset for automated differential diagnosis of liver tumors. *Scientific Data*, 13(1):31, 2025.

Jiahua Xu, Dawei Zhou, Lei Hu, Jianfeng Guo, Feng Yang, Zaiyi Liu, Nannan Wang, and Xinbo Gao. Motion artifact removal in pixel-frequency domain via alternate masks and diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8878–8886, 2025.

Xuelian Yang and Yuanjun Wang. Defnet: A dual-encoder feature fusion network for collaborative multi-task segmentation and predictive modeling in breast dce-mri. *Expert Systems with Applications*, pp. 130289, 2025.

Lukas Zbinden, Lars Doorenbos, Theodoros Pissas, Adrian Thomas Huber, Raphael Sznitman, and Pablo Márquez-Neila. Stochastic segmentation with conditional categorical diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1119–1129, 2023.

Jiadong Zhang, Zhiming Cui, Zhenwei Shi, Yingjia Jiang, Zhiliang Zhang, Xiaoting Dai, Zhenlu Yang, Yuning Gu, Lei Zhou, Chu Han, et al. A robust and efficient ai assistant for breast tumor segmentation from dce-mri via a spatial-temporal framework. *Patterns*, 4(9), 2023.

Zheyuan Zhang, Lanhong Yao, Bin Wang, Debesh Jha, Gorkem Durak, Elif Keles, Alpay Medetalibeyoglu, and Ulas Bagci. Diffboost: Enhancing medical image segmentation via text-guided diffusion model. *IEEE Transactions on Medical Imaging*, 2024.

Zhihe Zhao, Siyao Du, Zeyan Xu, Zhi Yin, Xiaomei Huang, Xin Huang, Chinting Wong, Yanting Liang, Jing Shen, Jianlin Wu, et al. Swinhr: Hemodynamic-powered hierarchical vision transformer for breast tumor segmentation. *Computers in Biology and Medicine*, pp. 107939, 2024.

Lei Zheng, Yuzhong Zhang, Jiadong Zhang, Tao Zhou, Kun Sun, Lei Zhou, and Dinggang Shen. Hcrt: Hybrid network with correlation-aware region transformer for breast tumor segmentation in dce-mri. *Pattern Recognition*, pp. 112934, 2025.

Lei Zhou, Shuai Wang, Kun Sun, Tao Zhou, Fuhua Yan, and Dinggang Shen. Three-dimensional affinity learning based multi-branch ensemble network for breast tumor segmentation in mri. *Pattern Recognition*, 129:108723, 2022.

Lei Zhou, Yuzhong Zhang, Jiadong Zhang, Xuejun Qian, Chen Gong, Kun Sun, Zhongxiang Ding, Xing Wang, Zhenhui Li, Zaiyi Liu, et al. Prototype learning guided hybrid network for breast tumor segmentation in dce-mri. *IEEE Transactions on Medical Imaging*, 2024.

A Experiments

A.1 Statistical significance analysis

We further conduct paired t-tests on both I-SPY2 and DUKE to verify the statistical significance of the observed gains. As shown in Tab. 5, HDRSNet achieves the best mean performance on I-SPY2 (DSC 78.54%±1.19, 95% HD 67.95±1.87 mm, Sen 77.41%±1.16) and consistently outperforms representative baselines, with all comparisons yielding highly significant differences (e.g., vs. TSESNet: $p<10^{-11}$ across DSC/HD/Sen). On DUKE, HDRSNet also ranks first (DSC $84.93\% \pm 1.67$, 95% HD 6.87 ± 2.17 mm, Sen $85.61\% \pm 1.35$), and the improvements over strong 4D models (TSESNet) and diffusion-based baselines (BBDiffSF/DEFNet) remain statistically significant ($p<10^{-7}$ for all metrics). Overall, these results indicate that our performance gains are stable across runs and consistent across the two datasets rather than arising from random variation.

Table 5: Statistical significance analysis of performance improvements by t-test on the I-SPY2 and DUKE datasets from five runs.

I-SPY2					**DUKE**				
Methods		DSC (%)	95% HD (mm)	Sen. (%)	Methods		DSC (%)	95% HD (mm)	Sen. (%)
TSESNet	$\mu \pm \sigma$	76.49 ± 1.32	71.53 ± 4.06	72.48 ± 1.81	TSESNet	$\mu \pm \sigma$	83.72 ± 1.54	7.51 ± 3.17	83.97 ± 1.62
	CI	[74.85, 78.13]	[66.49, 76.57]	[70.23, 74.73]		CI	[82.37, 85.07]	[5.56, 9.46]	[82.54, 85.40]
	p-value	3.3×10^{-12}	1.3×10^{-11}	1.5×10^{-13}		p-value	2.6×10^{-9}	4.1×10^{-8}	1.9×10^{-7}
BBDiffSF	$\mu \pm \sigma$	70.15 ± 2.28	91.36 ± 2.54	64.93 ± 2.08	BBDiffSF	$\mu \pm \sigma$	82.23 ± 1.92	10.15 ± 1.93	82.54 ± 1.86
	CI	[67.32, 72.98]	[88.21, 94.51]	[62.35, 67.51]		CI	[80.84, 83.62]	[8.78, 11.52]	[81.19, 83.89]
	p-value	3.3×10^{-14}	4.3×10^{-9}	1.7×10^{-15}		p-value	6.4×10^{-12}	2.7×10^{-15}	9.3×10^{-11}
DEFNet	$\mu \pm \sigma$	74.93 ± 1.47	81.05 ± 2.18	70.07 ± 1.59	DEFNet	$\mu \pm \sigma$	82.17 ± 2.37	11.93 ± 2.98	82.15 ± 2.03
	CI	[73.11, 76.75]	[78.34, 83.76]	[68.10, 72.04]		CI	[80.68, 83.66]	[10.06, 13.80]	[80.83, 83.47]
	p-value	3.0×10^{-13}	8.6×10^{-8}	5.4×10^{-11}		p-value	8.1×10^{-8}	3.6×10^{-14}	2.4×10^{-9}
HDRSNet	$\mu \pm \sigma$	78.54 ± 1.19	67.95 ± 1.87	77.41 ± 1.16	HDRSNet	$\mu \pm \sigma$	84.93 ± 1.67	6.87 ± 2.17	85.61 ± 1.35
	CI	**[77.06, 80.02]**	**[65.63, 70.27]**	**[75.97, 78.85]**		CI	**[83.69, 86.17]**	**[5.50, 8.24]**	**[84.57, 86.65]**

A.2 Sensitivity to DDIM Sampling Steps

We analyze the sensitivity of HDRSNet to the number of DDIM sampling steps on I-SPY2 while keeping all other inference settings unchanged. As shown in Tab. 6, segmentation performance improves with more sampling steps but quickly saturates. Increasing the number of steps from 5 to 25 improves DSC from 75.68% to 78.54%, whereas further increasing it to 50 and 100 yields only marginal gains of 0.04% and 0.07%, respectively.

Table 6: Sensitivity to the number of DDIM sampling steps on I-SPY2.

DDIM Steps	DSC (%) ↑	HD95 (mm) ↓	Sensitivity (%) ↑	Time (s/case) ↓
5	75.68 ± 1.35	75.84 ± 2.43	73.96 ± 1.42	0.48
10	77.31 ± 1.24	71.06 ± 2.05	75.82 ± 1.28	0.91
25	**78.54 ± 1.19**	**67.95 ± 1.87**	**77.41 ± 1.16**	**2.23**
50	78.58 ± 1.18	67.88 ± 1.85	77.45 ± 1.15	4.43
100	78.61 ± 1.17	67.81 ± 1.84	77.48 ± 1.14	8.81

Meanwhile, the inference time increases nearly linearly with the number of sampling steps, from 2.23 s/case at 25 steps to 8.81 s/case at 100 steps. We therefore use 25 DDIM steps in all experiments, which provides a favorable trade-off between segmentation accuracy and computational efficiency.

A.3 Robustness under Recursive Generation

We evaluate robustness to recursive generation failures on I-SPY2 by injecting Gaussian noise (σ=0.02) into the generated volume at a randomly selected phase during inference, and then continuing the recurrent generation without correction. We report the final segmentation performance to quantify how perturbations at different phases affect the downstream prediction (Tab. 7).

Table 7: Robustness evaluation of RHD under recursive generation on the I-SPY2 dataset. Noise indicates Gaussian perturbation ($\sigma = 0.02$) injected at a given phase during inference.

Setting	DSC (%)	95% HD (mm)	Sen (%)
No Perturbation	78.54	67.95	77.41
Noise at V_1	74.56	76.45	72.34
Noise at V_2	75.89	75.12	73.45
Noise at V_3	77.62	72.23	76.51
Noise at V_4	78.14	69.56	77.19

(1) HDRSNet remains stable under recursive perturbations, indicating controlled phase-to-phase transitions. Overall, segmentation degrades mildly even when intermediate generations are corrupted, suggesting that the learned dynamics transitions are stable and the segmentation relies on aggregated multi-phase features rather than a single fragile phase. Compared with the clean setting (78.54 DSC, 67.95 mm HD), the worst perturbation case still achieves 74.56 DSC with 76.45 mm HD, i.e., within 3.98 DSC and 8.50 mm HD of the baseline. This robustness supports that our recurrent diffusion dynamics are not overly sensitive to local synthesis noise.

(2) Early-phase perturbations are most harmful, consistent with error propagation in recurrent conditioning. Injecting noise at the earliest phase V_1 causes the largest drop (78.54 $\rightarrow$ 74.56 DSC; 67.95 $\rightarrow$ 76.45 mm HD), because subsequent phases are conditioned on earlier generated volumes, making early errors more likely to propagate. In contrast, when perturbations are injected later, the degradation becomes smaller (e.g., noise at V_3: 77.62 DSC, 72.23 mm HD), indicating that later-phase corruption has limited opportunity to cascade through the remaining recurrence.

(3) Later-phase perturbations have negligible impact due to multi-scale dynamics fusion and residual conditioning. When noise is injected at V_4, performance remains close to the clean baseline (78.14 DSC vs. 78.54; 69.56 mm HD vs. 67.95), with sensitivity essentially unchanged (77.19 vs. 77.41). This suggests that MTC-style multi-scale injection aggregates dynamics cues across phases and resolutions, and the residual fusion pathway prevents a single corrupted phase from dominating the final prediction. Together, these results demonstrate that HDRSNet is resilient to realistic generation imperfections during recursive inference.

A.4 Feature–Kinetics Correspondence Analysis

To examine whether the learned residual hemodynamic representations capture lesion-specific enhancement dynamics rather than generic diffusion semantics, we conduct a feature–kinetics correspondence analysis. For each test case, the ground-truth tumor enhancement curve is compared with

Table 8: Feature–kinetics correspondence analysis comparing HDRSNet with generic diffusion features and representative ablation variants.

Method	Pearson ↑	Wash-in MAE ↓	Peak-phase Acc. ↑	Tumor/BG Response ↑	DSC (%) ↑
Generic DDPM features	0.412	0.226	43.8	1.42	54.12
w/o Residual Modeling + Local Decoupling	0.574	0.169	59.6	1.87	74.02
RHD (Shuffled Phase Order)	0.301	0.241	32.7	2.11	71.18
HDRSNet (Full)	**0.781**	**0.089**	**82.4**	**3.26**	**78.54**

the phase-wise RHD feature response using Pearson correlation, wash-in MAE, peak-phase accuracy, and tumor-to-background response ratio.

We compare HDRSNet with generic DDPM features, a variant without residual modeling and local decoupling, and an RHD variant with shuffled phase order. The latter disrupts temporal dynamics to assess whether RHD preserves ordered enhancement evolution.

As shown in Tab. 8, the full HDRSNet exhibits the strongest correspondence with the ground-truth enhancement kinetics, achieving a Pearson correlation of 0.781, a wash-in MAE of 0.089, and a peak-phase accuracy of 82.4%. It also yields the highest tumor-to-background response ratio (3.26), indicating stronger lesion-specific activation.

In contrast, generic DDPM features show substantially weaker kinetic correspondence (Pearson 0.412 and peak-phase accuracy 43.8%) and result in only 54.12% DSC. Removing residual modeling and local decoupling also reduces both kinetic correspondence and segmentation performance. More importantly, shuffling the temporal phase order decreases the Pearson correlation from 0.781 to 0.301 and peak-phase accuracy from 82.4% to 32.7%, despite retaining the same phase observations. This pronounced degradation suggests that the learned RHD representation depends on the ordered enhancement evolution rather than merely aggregating phase-wise appearance features.

Together with the recursive perturbation analysis in Appendix A.3, these results provide complementary evidence that HDRSNet learns task-oriented representations associated with lesion enhancement dynamics, rather than relying primarily on image-generation fidelity or generic diffusion features.

A.5 Different combinations in MTC Module

We ablate MTC on I-SPY2 by selectively injecting diffusion dynamics features $\mathbf{hf}^l$ from different depths (l=1, . . . , 4) into the segmentation backbone, while keeping all other settings fixed. Each configuration is evaluated by DSC, 95% HD, and Sensitivity (Tab. 9).

Table 9: Ablation study for MTC at difference layer on the I-SPY2 dataset. Single-layer fusion best results are highlighted in green. The best overall results are highlighted in red.

$\mathbf{hf}^1$	$\mathbf{hf}^2$	$\mathbf{hf}^3$	$\mathbf{hf}^4$	DSC (%)	95% HD (mm)	Sen (%)
✓				65.74	137.56	64.18
	✓			69.81	105.32	68.12
		✓		71.53	85.43	70.85
			✓	75.53	76.12	75.21
✓	✓			72.19	87.31	71.12
✓		✓		73.31	86.82	73.19
✓			✓	74.29	80.37	74.52
	✓	✓		73.91	86.42	73.75
	✓		✓	75.82	73.82	75.49
		✓	✓	76.53	71.82	75.93
✓	✓	✓		75.94	74.63	75.87
✓	✓		✓	76.32	72.31	76.28
	✓	✓	✓	76.93	71.12	76.39
✓	✓	✓	✓	78.54	67.95	77.41

(1) Deeper dynamics features are more effective when used alone. With single-layer injection, performance improves monotonically with feature depth. Using $\mathbf{hf}^1$ yields only 65.74% DSC and 137.56 mm HD, while injecting the deepest feature $\mathbf{hf}^4$ achieves 75.53% DSC and 76.12 mm HD. This indicates that deeper diffusion layers encode more semantic and tumor-aligned hemodynamic, whereas shallow features are dominated by local textures and noise.

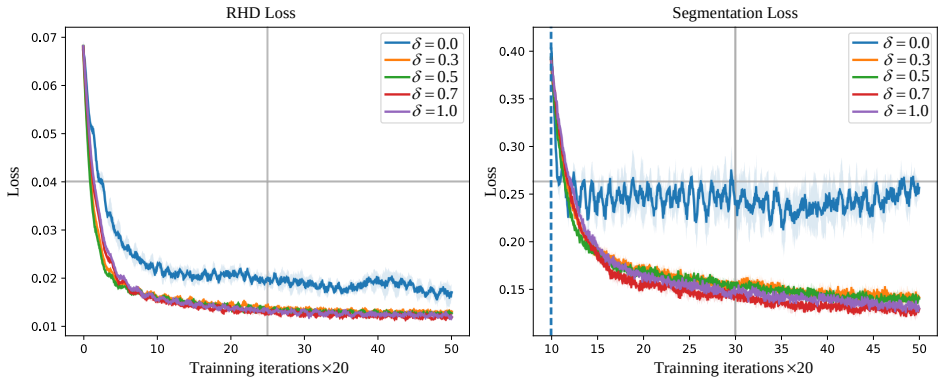

Figure 5: Training loss curves under different dynamic blending weights δ on I-SPY2. Left: RHD diffusion loss during training. Right: segmentation loss. The dashed vertical line indicates the transition point where segmentation training is introduced.

(2) Multi-scale dynamics are complementary and improve boundary accuracy. Combining multiple depths consistently outperforms any single-layer setting, demonstrating that dynamics at different resolutions provide complementary cues. For example, $\mathbf{hf}^3+\mathbf{hf}^4$ improves performance to 76.53% DSC and 71.82 mm HD, compared with 75.53% / 76.12 mm for $\mathbf{hf}^4$ alone. These gains are more pronounced in HD, suggesting that MTC mainly helps refine tumor boundaries by aggregating coarse localization and fine-grained dynamics.

(3) Full multi-scale injection achieves the best overall performance. Injecting all four dynamics features yields the best results (78.54% DSC, 67.95 mm HD, 77.41% Sen), outperforming all partial combinations. This confirms that MTC effectively integrates coarse-to-fine hemodynamic cues, enabling robust and boundary-aware tumor delineation from pre-contrast input alone.

A.6 Effect of Dynamic Blending Weight δ

We study the impact of the dynamic blending weight δ in Eq. (3) while keeping the network architecture, training schedule, and all other hyper-parameters fixed. We report mean-SSIM of the generated phase and the downstream segmentation metrics (DSC, 95% HD, Sen) on I-SPY2 (Tab. 10), and visualize the training losses of RHD and the segmentation network in Fig. 5.

Table 10: Effect of dynamic blending weight δ on I-SPY2. Results are reported as mean $\pm$ std over five runs.

δ	mean-SSIM ↑	DSC (%) ↑	95% HD (mm) ↓	Sen (%) ↑
0.0	0.5245	60.11±2.05	136.37±6.72	58.54±2.14
0.3	0.7918	75.62±1.41	76.84±3.95	74.18±1.58
0.5	0.8354	78.12±1.28	69.21±3.26	76.83±1.36
0.7	**0.8561**	**78.54±1.19**	**67.95±1.87**	**77.41±1.16**
1.0	0.8389	78.08±1.34	69.02±3.11	76.95±1.29

(1) Nonzero blending is necessary for stable recurrent conditioning and accurate segmentation. When δ=0, the model relies solely on generated conditioning $\hat{V}_i$ from the beginning, which amplifies compounding errors across phases and leads to unstable optimization. This is reflected by the sharp drop in mean-SSIM (0.5245) and poor segmentation (60.11% DSC, 136.37 mm HD, 58.54% Sen). In contrast, introducing real-to-generated blending markedly improves both synthesis fidelity and segmentation reliability; for example, δ=0.3 boosts mean-SSIM to 0.7918 and improves DSC to 75.62% while reducing HD to 76.84 mm, indicating that early real-phase conditioning provides a cleaner temporal context for learning dynamics and boundaries.

(2) Moderate blending achieves the best trade-off, while overly strong blending yields diminishing returns. The best performance is obtained at δ=0.7, which achieves the highest mean-SSIM

(0.8561) and the strongest segmentation (78.54% DSC, 67.95 mm HD, 77.41% Sen). Compared with δ=0.5, δ=0.7 provides a consistent gain (+0.42% DSC and -1.26 mm HD), suggesting that a stronger early reliance on real conditioning better regularizes the recurrent diffusion trajectory. However, increasing δ to 1.0 does not further improve performance: mean-SSIM slightly decreases (0.8389) and DSC/HD marginally degrade (78.08% DSC, 69.02 mm HD), suggesting that overly strong real-phase conditioning reduces exposure to generated conditioning during training and slightly weakens robustness under fully self-conditioned inference.

Loss visualization. Fig. 5 is consistent with these findings. With δ=0, the RHD loss exhibits noticeable oscillation after the transition to joint training, and the segmentation loss fails to steadily decrease, indicating error accumulation and unstable coupling between generation and segmentation. For $\delta \in \{0.3, 0.5, 0.7, 1.0\}$, both losses decrease smoothly and remain stable after joint optimization starts, with δ=0.7 converging to the lowest final loss level, matching its best quantitative performance in Tab. 10.

A.7 Effect of loss-weight η and γ

We investigate the impact of loss-weight coupling on model performance by jointly varying η (Eq.(11)) and γ (Eq.(12)), where η controls the relative strength of the cross-entropy term in $\mathcal{L}_{\text{seg}}$ and γ balances diffusion dynamics distillation against segmentation supervision in $\mathcal{L}_{\text{total}}$. The results on both I-SPY2 and DUKE are reported in Tab. 11.

Table 11: Joint ablation of η and γ on I-SPY2 and DUKE. Each cell reports mean $\pm$ std over five runs: DSC (%)↑, 95% HD (mm)↓, and Sen (%)↑. Best results on each dataset are highlighted in bold.

I-SPY2						DUKE					
Setting		η=0.3	η=0.7	η=1.0	η=1.5	Setting		η=0.3	η=0.7	η=1.0	η=1.5
γ=0.3	DSC	74.31±1.5	76.04±1.4	75.62±1.6	74.77±1.7	γ=0.3	DSC	81.63±1.6	83.17±1.4	82.74±1.5	81.96±1.7
	HD	79.63±2.4	74.86±2.0	76.17±2.2	78.39±2.6		HD	10.73±2.1	9.14±1.9	9.67±2.0	10.36±2.2
	Sen	72.58±1.8	74.79±1.6	74.16±1.7	72.94±1.9		Sen	82.35±1.7	83.92±1.5	83.48±1.6	82.69±1.8
γ=0.5	DSC	76.63±1.3	**78.54±1.2**	77.74±1.4	76.79±1.5	γ=0.5	DSC	83.76±1.3	**84.93±1.2**	84.45±1.3	83.82±1.4
	HD	72.97±2.0	**67.95±1.8**	69.79±1.9	71.86±2.1		HD	8.13±1.6	**6.87±1.5**	7.34±1.5	7.96±1.6
	Sen	74.96±1.5	**77.41±1.3**	76.68±1.4	75.27±1.6		Sen	84.08±1.4	**85.61±1.3**	85.07±1.4	84.29±1.5
γ=0.7	DSC	75.46±1.7	76.98±1.5	76.27±1.6	75.05±1.8	γ=0.7	DSC	82.86±1.7	84.14±1.5	83.58±1.6	82.63±1.8
	HD	75.48±2.3	71.28±2.1	72.66±2.2	75.94±2.4		HD	8.95±1.9	7.62±1.7	8.17±1.8	8.74±1.9
	Sen	73.56±1.9	75.69±1.7	74.88±1.8	73.27±2.0		Sen	83.28±1.8	84.64±1.6	84.03±1.7	83.19±1.9

(1) A balanced objective yields the best accuracy. Across both datasets, the best performance is achieved at η=0.7, γ=0.5 (I-SPY2: DSC 78.54%, HD 67.95 mm, Sen 77.41%; DUKE: DSC 84.93%, HD 6.87 mm, Sen 85.61%), indicating that effective segmentation requires both (i) sufficient hemodynamic supervision from RHD and (ii) a well-calibrated CE term to complement Dice. When γ is reduced to 0.3, the model under-emphasizes dynamics distillation and the DSC drops to 76.04% on I-SPY2 and 83.17% on DUKE even at η=0.7 (HD increases to 74.86 mm and 9.14 mm, respectively). Conversely, increasing γ to 0.7 shifts optimization toward diffusion, leading to weaker segmentation (best at η=0.7: I-SPY2 DSC 76.98%, HD 71.28 mm; DUKE DSC 84.14%, HD 7.62 mm), showing that over-weighting distillation can dilute discriminative supervision.

(2) Moderate CE weighting is consistently preferable across γ. Across all γ, η=0.7 gives the strongest or near-strongest results on both datasets, while too small/large η degrades performance, reflecting the trade-off between region overlap (Dice) and voxel-wise calibration (CE). for example, at γ=0.5, moving from η=0.7 to η=0.3 reduces DSC from 78.54% to 76.63% on I-SPY2 and from 84.93% to 83.76% on DUKE, and lowers Sen from 77.41% to 74.96% and from 85.61% to 84.08%, respectively, suggesting insufficient CE regularization harms sensitivity. Increasing η to 1.5 also hurts (I-SPY2: DSC 76.79%, HD 71.86 mm; DUKE: DSC 83.82%, HD 7.96 mm), likely because overly strong CE over-penalizes hard voxels and weakens boundary-aware overlap learning.

Visualization. Fig. 6 further confirms this trend on both I-SPY2 and DUKE: DSC peaks around (η, γ)=$(0.7, 0.5)$ and decreases when either the segmentation loss is over-weighted (large η) or the diffusion objective dominates (large γ). The consistency of the peak region across datasets suggests that stable pre-contrast-only segmentation benefits most from a balanced coupling between task-aligned dynamics distillation and discriminative supervision.

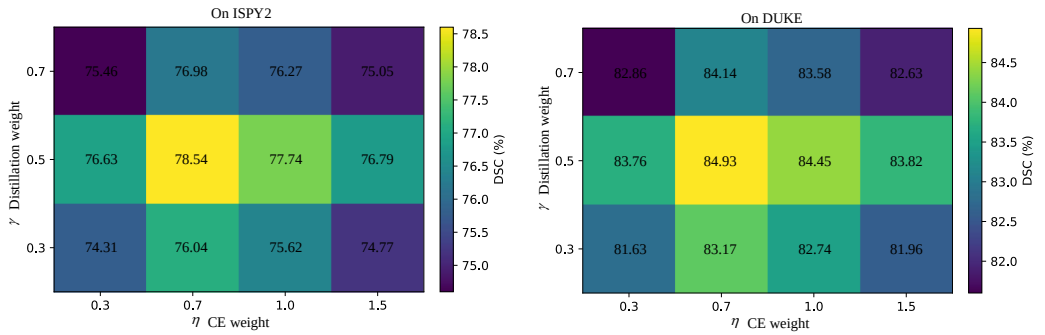

Figure 6: Heatmap of DSC (%) under varying η and γ on I-SPY2 and DUKE.

A.8 Comparison under Missing-Phase Settings

We study robustness to missing phases by evaluating representative methods under different DCE sampling settings on I-SPY2 (T=5) and DUKE (T=4). Specifically, MedSegDiff (Wu et al., 2024) and nnUNet (Garrucho et al., 2024) are tested with V_0 or 3D MIP, ALMN (Zhou et al., 2022) uses $[V_0, V_i]$ with a single post-contrast phase, and DEFNet (Yang & Wang, 2025) and TSESNet (He et al., 2024) progressively take $[V_0 \sim V_k]$ for increasing k. In contrast, HDRSNet uses only V_0 for all cases. Results are summarized in Tab. 12.

Table 12: Comparative Analysis of Segmentation Performance with Varying DCE-MRI Sampling Points on the I-SPY2 and DUKE Datasets. T represents the total number of DCE-MRI post-contrast time points available, which is $T = 5$ for the I-SPY2 and $T = 4$ for the DUKE. Every method's best results are highlighted in green. All method's best results are highlighted in red.

Methods	Input	I-SPY2			DUKE		
		DSC (%)	95% HD (mm)	Sen (%)	DSC (%)	95% HD (mm)	Sen (%)
MedSegDiff	V_0	25.84	289.32	23.41	34.71	189.32	31.19
	MIP 3D	73.81	88.94	70.93	79.21	15.35	80.08
nnUNet	V_0	32.97	268.15	30.53	38.71	156.74	37.19
	3D MIP	69.31	97.37	64.32	76.17	20.53	74.32
ALMN	$[V_0, V_1]$	53.41	182.72	50.16	66.37	30.85	66.19
	$[V_0, V_2]$	59.72	163.42	57.43	72.93	23.89	71.48
	$[V_0, V_3]$	64.47	130.32	59.81	80.19	15.19	78.54
	$[V_0, V_4]$	67.12	104.18	62.17	80.04	16.37	77.34
	$[V_0, V_5]$	65.82	125.62	60.59	/	/	/
TSESNet	$[V_0 \sim V_1]$	54.37	121.30	52.09	59.17	38.14	58.46
	$[V_0 \sim V_2]$	68.17	94.55	65.16	70.15	26.13	69.81
	$[V_0 \sim V_3]$	77.13	69.39	73.16	80.41	15.63	80.11
	$[V_0 \sim V_4]$	76.19	73.17	72.54	83.72	7.514	83.97
	$[V_0 \sim V_5]$	76.49	71.53	72.48	/	/	/
DEFNet	$[V_0 \sim V_1]$	47.57	171.32	44.39	61.49	37.81	61.98
	$[V_0 \sim V_2]$	60.31	156.43	57.54	70.18	25.96	70.31
	$[V_0 \sim V_3]$	69.32	93.76	65.81	79.12	16.38	79.41
	$[V_0 \sim V_4]$	75.17	79.32	72.01	82.17	11.93	82.15
	$[V_0 \sim V_5]$	74.93	81.05	70.07	/	/	/
HDRSNet	V_0	78.54	67.95	77.41	84.93	6.87	85.61

(1) Performance improves as more temporal evidence is available, confirming the value of explicit temporal modeling. for sequence-based models, adding post-contrast phases consistently boosts accuracy because longer trajectories better capture enhancement evolution. On I-SPY2, TSESNet increases from 54.37% DSC with $[V_0 \sim V_1]$ to 77.13% with $[V_0 \sim V_3]$ (HD drops from 121.30 to 69.39 mm), while DEFNet rises from 47.57% to 75.17% when extending to $[V_0 \sim V_4]$. A similar trend appears on DUKE, where TSESNet improves from 59.17% to 83.72% as the input

expands to $[V_0 \sim V_4]$, validating that explicit temporal modeling provides boundary-disambiguating cues beyond static pre-contrast appearance.

(2) Limited-input strategies (single phase or MIP) provide partial gains but remain constrained by incomplete temporal coverage. Using only V_0 leads to severe degradation (e.g., MedSegDiff: 25.84%/34.71% DSC on I-SPY2/DUKE; nnUNet: 32.97%/38.71%), since enhancement dynamics are entirely missing. Compressing information into 3D MIP alleviates this by aggregating salient responses (MedSegDiff: 73.81% DSC on I-SPY2; nnUNet: 69.31%), but it discards phase ordering and can blur subtle temporal changes, leaving performance behind strong temporal models (e.g., DUKE: nnUNet MIP 76.17% vs. TSESNet 83.72%). Two-phase ALMN improves when the chosen V_i contains stronger late enhancement (I-SPY2 best 67.12% at $[V_0, V_4]$; DUKE best 80.19% at $[V_0, V_3]$), yet remains limited because a single post-contrast snapshot cannot enforce multi-phase consistency.

(3) HDRSNet is consistently best under the most extreme missingness by distilling dynamics from V_0. HDRSNet avoids dependence on any post-contrast observations by distilling task-aligned dynamics via diffusion and injecting them for segmentation, which mitigates the brittleness of partial sampling. Consequently, it achieves the best results even in the strict V_0-only setting: 78.54% DSC and 67.95 mm HD on I-SPY2, surpassing the best full-sequence baselines TSESNet (77.13%, 69.39 mm with $[V_0 \sim V_3]$) and DEFNet (75.17%, 79.32 mm with $[V_0 \sim V_4]$). On DUKE, it further reaches 84.93% DSC with 6.87 mm HD, outperforming TSESNet with $[V_0 \sim V_4]$ (83.72%, 7.514 mm), indicating that recovering dynamics as diffusion features can be more reliable than relying on incomplete phase observations.

A.9 Generalization Analysis

We further evaluate HDRSNet from two complementary perspectives: *cross-center generalization* within breast DCE-MRI and *cross-modality/organ generalization* to multiphase contrast-enhanced CT. For cross-center evaluation, we use BTS (Zhang et al., 2023), which contains 100 breast DCE-MRI cases from seven medical centers. We evaluate both in-domain training/testing on BTS and cross-center transfer from DUKE to BTS. To examine whether the proposed residual hemodynamic representation is specific to breast DCE-MRI, we further evaluate HDRSNet on MCT-LTDiag (Wu et al., 2025), a multiphase contrast-enhanced CT dataset for liver tumor segmentation. All results are averaged over five runs with different random seeds.

Table 13: Quantitative comparison on the BTS dataset. Results are reported as mean ± standard deviation over five runs.

Methods	DSC (%) ↑	95% HD (mm) ↓	Sen (%) ↑
TSESNet (He et al., 2024)	78.95 ± 1.13	6.14 ± 2.31	79.13 ± 1.09
BBDiffSF (Feng et al., 2026)	78.14 ± 1.57	7.32 ± 2.18	77.96 ± 1.49
DEFNet (Yang & Wang, 2025)	79.38 ± 1.06	5.92 ± 0.69	79.64 ± 1.18
HDRSNet (Ours)	**82.12 ± 0.98**	**4.41 ± 0.63**	**82.86 ± 1.05**

Table 14: Cross-center generalization from DUKE to BTS. Results are reported as mean ± standard deviation over five runs.

Methods	DSC (%) ↑	95% HD (mm) ↓	Sen (%) ↑
TSESNet (He et al., 2024)	64.38 ± 3.67	20.20 ± 6.34	64.17 ± 3.21
BBDiffSF (Feng et al., 2026)	60.27 ± 4.89	25.34 ± 8.12	59.83 ± 2.57
DEFNet (Yang & Wang, 2025)	54.93 ± 6.98	30.12 ± 7.87	55.21 ± 6.14
HDRSNet (Ours)	**77.86 ± 1.12**	**7.58 ± 0.78**	**75.94 ± 1.32**

(1) HDRSNet achieves strong in-domain performance on multi-center breast DCE-MRI. As shown in Tab. 13, HDRSNet achieves 82.12% DSC, 4.41 mm HD, and 82.86% sensitivity on BTS. Compared with the strongest baseline DEFNet (79.38% DSC and 5.92 mm HD), HDRSNet improves DSC by 2.74% while reducing HD by 1.51 mm, demonstrating more accurate tumor delineation.

(2) HDRSNet remains robust under cross-center domain shift. When trained on DUKE and directly evaluated on BTS (Tab. 14), HDRSNet achieves 77.86% DSC and 7.58 mm HD, substantially

Table 15: Generalization to multiphase contrast-enhanced CT liver tumor segmentation on MCT-LTDiag.

Methods	DSC (%) ↑	HD95 (mm) ↓	Sen (%) ↑
DEFNet (Yang & Wang, 2025)	67.82 ± 1.73	18.46 ± 2.31	66.95 ± 1.89
BBDiffSF (Feng et al., 2026)	68.41 ± 1.56	17.72 ± 2.08	67.83 ± 1.67
HDRSNet (Ours)	**71.03 ± 1.42**	**15.61 ± 1.84**	**70.52 ± 1.51**

outperforming TSESNet (64.38%, 20.20 mm) and BBDiffSF (60.27%, 25.34 mm). This suggests that the distilled residual dynamics are less sensitive to center-dependent variations in acquisition protocols and contrast appearance.

(3) The residual hemodynamic representation generalizes beyond breast DCE-MRI. As shown in Tab. 15, HDRSNet achieves 71.03% DSC, 15.61 mm HD95, and 70.52% sensitivity on multiphase CE-CT liver tumor segmentation, outperforming BBDiffSF (68.41% DSC, 17.72 mm HD95) and DEFNet (67.82% DSC, 18.46 mm HD95). The consistent improvement across different imaging modalities and anatomical targets indicates that HDRSNet captures task-relevant contrast dynamics rather than relying solely on breast-specific DCE-MRI characteristics.

B Discussion

B.1 On Modeling Hemodynamic as Residual Evolution

Let $\{V_i\}_{i=0}^{T}$ denote a DCE-MRI sequence, where V_0 is the pre-contrast volume. Instead of modeling the absolute phase intensity, we parameterize the phase evolution by a residual process

$$\Delta V_i \triangleq V_i - V_0, \quad i = 1, \ldots, T, \tag{13}$$

so that the diffusion model focuses on learning dynamics conditioned on a fixed anatomical reference V_0. In practice, tumor delineation is primarily driven by localized enhancement changes, i.e., sparse regions where $|\Delta V_i|$ is large, rather than global appearance variations. Thus, the residual formulation biases model capacity toward hemodynamic variation and away from reconstructing background appearance.

However, ΔV_i is only a surrogate for the underlying contrast kinetics. Due to temporal misalignment, scanner-dependent intensity scaling, and patient-specific perfusion heterogeneity, the residual can be decomposed as

$$\Delta V_i = \Delta V_i^{\text{lesion}} + \Delta V_i^{\text{nuis}}, \tag{14}$$

where ΔV_i^{nuis} aggregates non-lesion fluctuations that may not be predictive for segmentation. This motivates learning with an explicit spatial separation between dynamic regions and background-stable regions, rather than treating all residual responses as equally informative.

B.2 On Local Decoupling and Its Assumptions

We introduce a soft gate $M_i \in [0,1]^{H\times W\times D}$ to emphasize dynamic (foreground-relevant) voxels and suppress background-stable regions in diffusion learning. Given the residual magnitude w.r.t. the pre-contrast baseline,

$$R_i = |V_i - V_0|, \tag{15}$$

we construct the gate by a normalized-and-thresholded mapping

$$M_i = \sigma\Big(\kappa\big(\operatorname{Norm}(R_i) - \tau\big)\Big), \tag{16}$$

and use it to form a per-voxel weight map

$$W_i = \lambda + (1-\lambda)M_i, \qquad \lambda \in (0,1). \tag{17}$$

The diffusion objective is then reweighted as

$$\mathcal{L}_{\text{LD-diff}} = \mathbb{E}_{i,s,\epsilon}\left[\left\|\sqrt{W_i}\odot\big(\epsilon - \epsilon_\theta(z_i^s, s, \tilde{V}_{i-1})\big)\right\|_2^2\right], \tag{18}$$

which yields a controlled emphasis ratio: voxels with $M_i \approx 1$ receive unit weight, while background-stable voxels with $M_i \approx 0$ are down-weighted by λ. This reduces the gradient contribution of near-zero residual regions that are often dominated by noise, protocol drift, or scanner-specific intensity fluctuations, thereby discouraging the denoiser from spending capacity on background residual fitting and reducing the chance that small hallucinated textures propagate through recurrent generation.

A key assumption is that R_i is a reliable proxy for lesion-relevant dynamics. In practice, non-tumor tissues (e.g., vessels and glandular regions) may also exhibit enhancement, so M_i can over-activate outside the lesion. To bias the gate toward tumor-aligned dynamics during training, one can incorporate segmentation priors Y (binary mask) by modulating the gate, e.g.,

$$M_i \leftarrow \mathrm{clip}\big(M_i \odot (\alpha + (1-\alpha)Y),\, 0,\, 1\big), \qquad \alpha \in (0,1), \tag{19}$$

which enforces a higher minimum weight inside Y while preserving the residual-driven structure outside the mask. This introduces dependence on annotation quality, but in our experiments the decoupling still consistently improves robustness under domain shift, suggesting that suppressing background-driven residual learning is beneficial even when M_i is imperfect.

B.3 On Recurrent Conditioning Stability and Dynamic Blending

RHD models phase progression with a recurrent conditioning chain. At phase $i{+}1$, the denoiser is conditioned on a previous-phase context C_i, and in fully recurrent training this context is the generated volume $\hat{V}_i$:

$$\hat{V}_{i+1} \sim p_\theta(\cdot \mid \hat{V}_i, V_0), \qquad i = 0, \ldots, T-1. \tag{20}$$

In early training, $\hat{V}_i$ is inaccurate and induces a distribution shift in the conditioning signal, which can be viewed as an error-propagation recursion

$$e_{i+1} = \hat{V}_{i+1} - V_{i+1} \;\approx\; \mathcal{F}_\theta(\hat{V}_i) - \mathcal{F}_\theta(V_i), \tag{21}$$

so conditioning solely on $\hat{V}_i$ amplifies compounding errors along the chain and destabilizes diffusion learning, which further harms downstream segmentation.

To mitigate this issue, we adopt a scheduled blend between real and generated previous-phase volumes:

$$\tilde{V}_i = (1-\beta)\hat{V}_i + \beta V_i, \qquad \beta = \max\Big(0, \delta\Big(1 - \tfrac{\text{epoch}}{\text{total}}\Big)\Big), \tag{22}$$

and train the recurrent model by conditioning on $\tilde{V}_i$ instead of $\hat{V}_i$. This schedule implements a smooth transition from real-phase conditioning ($\beta \approx \delta$ at early epochs) to self-conditioning ($\beta \to 0$), analogous to a curriculum that first optimizes under reliable contexts and then gradually exposes the model to its own predictions.

Empirically, a moderate initial blending weight (intermediate δ) yields the best final performance: too small δ provides insufficient stabilization and leads to poor convergence, whereas too large δ over-relies on V_i and slows adaptation to the fully recurrent inference regime, consistent with the observed trend in the δ ablation.

B.4 Why Diffusion-Feature Conditioning Can Outperform Full-Sequence Inputs

A key design choice in HDRSNet is to not feed synthesized volumes $\{\hat{V}_i\}$ (or raw multi-phase intensities) into the segmentation network. Although 4D methods access all DCE phases, directly aggregating phase intensities can be sensitive to temporal misalignment, protocol variability, and patient-specific enhancement heterogeneity, which may introduce spurious dynamics and destabilize fusion-based decoding. Instead, HDRSNet treats diffusion as a dynamics encoder: it extracts multi-level intermediate denoising representations $\{\mathbf{hf}_i^l\}$ from the diffusion U-Net and injects them into the segmentation backbone via multi-scale task-aligned conditioning (MTC). Since these features are produced along the denoising trajectory, they act as a denoising-and-selection mechanism over temporal information, and local decoupling further suppresses background-driven variations. As a result, HDRSNet provides more stable and tumor-relevant temporal cues than either using synthesized images (which can contain subtle drift/artifacts amplified in dense prediction) or naively fusing full sequences, helping explain why it can surpass full-sequence 4D baselines even when only V_0 is available at inference.

C Implementation Details

This section provides detailed implementation and training configurations of HDRSNet to facilitate reproducibility. We describe the joint optimization of the diffusion dynamics distillation module and the segmentation network, including the training schedule, conditioning strategy, multi-scale feature extraction, and inference procedure. A unified pseudocode is provided to clearly illustrate the end-to-end workflow of HDRSNet.

Algorithm 1: HDRSNet Training and Inference (RHD + MTC-conditioned Segmentation)

Require: DCE volumes $\{V_0, \dots, V_T\}$, tumor mask Y, diffusion steps S, epochs E, blending horizon E_{blend}, LD weight λ, loss weights (η, γ)

Ensure: Predicted mask $\hat{Y}$ given only V_0 at inference

Initialize diffusion denoiser ϵ_θ and segmentation network g_ϕ
Precompute diffusion schedule $\{\alpha_s, \bar{\alpha}_s\}_{s=1}^{S}$
Training:
for epoch $e = 1$ to E **do**
  Set blending weight $\beta_e \leftarrow \max\left(0, \delta\left(1 - \frac{e}{E_{\text{blend}}}\right)\right)$
  Sample a training case $(\{V_0, \dots, V_T\}, Y)$ and a crop
  Set $\hat{V}_0 \leftarrow V_0$
  for transition $i = 1$ to T **do**
    Construct context $\tilde{V}_{i-1} \leftarrow (1 - \beta_e)\hat{V}_{i-1} + \beta_e V_{i-1}$
    Sample $s \sim \mathcal{U}\{1, \dots, S\}$, $\epsilon \sim \mathcal{N}(0, I)$
    forward diffuse: $z_i^s \leftarrow \sqrt{\bar{\alpha}_s} V_i + \sqrt{1 - \bar{\alpha}_s}\epsilon$
    Local gate: $M_i = \sigma\Big(\kappa\big(\operatorname{Norm}(|V_i - V_0|) \odot Y - \tau\big)\Big)$
    Weight map: $W_i \leftarrow \lambda + (1 - \lambda)M_i$
    Predict noise: $\hat{\epsilon} \leftarrow \epsilon_\theta(z_i^s, s, \tilde{V}_{i-1})$
    Accumulate diffusion loss: $\mathcal{L}_{\text{LD-diff}}^{(i)} \leftarrow \left\|\sqrt{W_i} \odot (\epsilon - \hat{\epsilon})\right\|_2^2$
    Extract multi-level dynamics features $\{\mathbf{hf}_i^l\}_{l=1}^{L}$ from RHD
    Update $\hat{V}_i$ using DDPM update with $(z_i^s, \hat{\epsilon})$
  end for
  Encode V_0: $\{\mathbf{sf}^l\}_{l=1}^{L} \leftarrow g_\phi^{\text{enc}}(V_0)$
  for scale $l = 1$ to L **do**
    MTC fusion: $\mathbf{ff}^l \leftarrow \mathbf{sf}^l + \frac{1}{T}\sum_{i=1}^{T} \mathbf{A}_i^l(\mathbf{hf}_i^l)$
  end for
  Decode: $\hat{Y} \leftarrow g_\phi^{\text{dec}}(\{\mathbf{ff}^l\}_{l=1}^{L})$
  Seg loss: $\mathcal{L}_{\text{seg}} \leftarrow \mathcal{L}_{\text{Dice}}(\hat{Y}, Y) + \eta\,\mathcal{L}_{\text{CE}}(\hat{Y}, Y)$
  Total loss: $\mathcal{L}_{\text{total}} \leftarrow \gamma\frac{1}{T}\sum_{i=0}^{T-1}\mathcal{L}_{\text{LD-diff}}^{(i+1)} + (1 - \gamma)\mathcal{L}_{\text{seg}}$
  Update (θ, ϕ) by minimizing $\mathcal{L}_{\text{total}}$
end for
Inference (given only V_0):
Set $\hat{V}_0 \leftarrow V_0$
for transition $i = 1$ to T **do**
  Initialize $z_i^S \sim \mathcal{N}(0, I)$
  for $s = S$ down to 1 **do**
    Set $\tilde{V}_{i-1} \leftarrow \hat{V}_{i-1}$
    $\hat{\epsilon} \leftarrow \epsilon_\theta(z_i^s, s, \tilde{V}_{i-1})$ and record $\{\mathbf{hf}_i^l\}$
    Update $z_i^{s-1} \leftarrow \text{DDIM}(z_i^s, \hat{\epsilon})$
  end for
  Set $\hat{V}_i \leftarrow z_i^0$
end for
$\{\mathbf{sf}^l\} \leftarrow g_\phi^{\text{enc}}(V_0)$; fuse with $\{\mathbf{hf}_i^l\}$ via MTC; decode $\hat{Y}$
Output: $\hat{Y}$

D Limitations and Future Directions

First, although HDRSNet reduces reliance on full sequences at inference, recurrent diffusion still introduces computational overhead compared to purely discriminative V_0-only models; future work may explore distilling diffusion dynamics into a lightweight student to accelerate deployment. Second, while local decoupling improves robustness, the gate design could be further refined to better separate

tumor enhancement from other enhancing tissues (e.g., vessel-aware suppression or anatomy-adaptive thresholds). Third, we currently focus on phase-wise dynamics representation learning; integrating explicit pharmacokinetic priors or uncertainty estimation may further improve reliability under severe protocol shift or extreme missingness. Finally, broader validation on more centers and prospective clinical cohorts will be important to quantify generalization beyond the current benchmarks.

AI USE STATEMENT

Generative AI tools were used solely to aid manuscript writing, including language polishing and improving clarity. All AI-assisted text was reviewed and verified by the authors. The authors take full responsibility for the final content of the manuscript.