Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.083677

ARTICLE

Structure-Aware Diffusion Image Outpainting for Echocardiographic Field-of-View Extension

Ruijia He[1], Yixin Hu[2], Jinze Liu[3], Qixin Zhang[4], Duo Peng[5] and Yanyan Chen[5,*]

[1]State Key Laboratory of Primate Biomedical Research, Institute of Primate Translational Medicine, Kunming University of Science and Technology, Kunming, China
[2]College of Electronics and Information Engineering, Sichuan University, Chengdu, China
[3]Central South University of Forestry and Technology, Changsha, China
[4]College of Computing and Data Science, Nanyang Technological University, Singapore, Singapore
[5]School of Computer Science and Technology, Tongji University, Shanghai, China

*Corresponding Author: Yanyan Chen. Email: cyanyan@tongji.edu.cn

Received: 08 April 2026; Accepted: 02 June 2026; Published: 23 July 2026

ABSTRACT: Transthoracic Echocardiography (TTE) often suffers from a limited field of view (FoV), which may obscure peripheral cardiac structures and hinder comprehensive visual assessment. Although FoV extension can be formulated as outpainting, public studies on echocardiographic FoV outpainting remain scarce and are mainly represented by cGAN-based reconstruction methods such as echoGAN. However, in noisy ultrasound images with weak boundaries, the challenge is not only realistic texture synthesis, but also structural continuity across the observed–generated boundary. To address this issue, we propose a structure-aware diffusion framework for echocardiographic FoV outpainting. To the best of our knowledge, this is the first work to introduce diffusion models into this task, moving beyond the existing cGAN-based formulation. Our framework combines a diffusion baseline, a Structural Cue Encoder (SCE) for structure-sensitive conditioning, and Structure-aware Diffusion Learning (SDL) for structural regularization during denoising. Experiments show clear and consistent improvements over cGAN-based reconstruction, producing more coherent structures, smoother transitions, and better FoV extension quality.

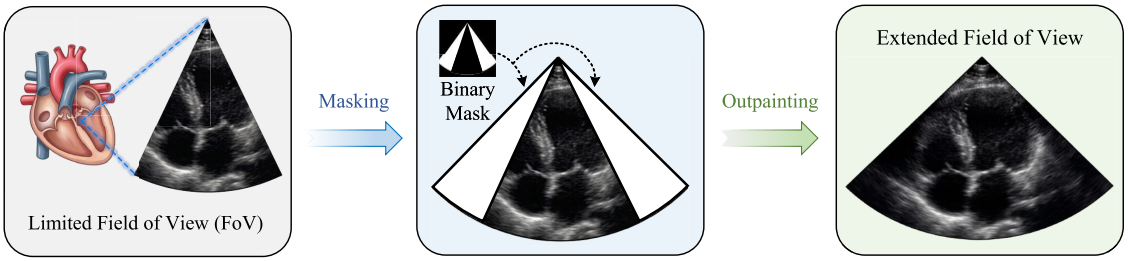
KEYWORDS: Field-of-view extension; diffusion model; image generation; outpainting; structure-aware learning; transthoracic echocardiography

1 Introduction

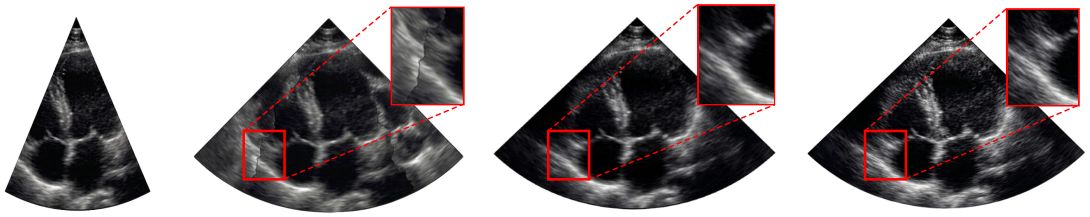
Transthoracic Echocardiography (TTE) is one of the most widely used imaging modalities for cardiac examination because of its non-invasiveness, accessibility, low cost, and real-time capability [1]. It provides essential information for evaluating cardiac anatomy, chamber motion, and functional abnormalities. However, ultrasound acquisition imposes a trade-off between field of view (FoV) and image resolution [2]: a narrower FoV preserves local details but may obscure peripheral structures and reduce contextual information, thereby limiting TTE interpretability.

From the perspective of image analysis, this limitation naturally motivates the problem of echocardiographic FoV extension, that is, recovering missing peripheral content while preserving visible structures. This task can be formulated as image outpainting [3,4], as shown in Fig. 1a. However, echocardiographic FoV outpainting remains underexplored [5]. A representative method, echoGAN [5], uses a cGAN framework [6] to synthesize missing regions and expand the observable area, demonstrating the feasibility of

TTE FoV extension. At the same time, its formulation also reveals a fundamental challenge that remains insufficiently addressed.



(a) Task of Echocardiographic FoV Outpainting



(b) Visual Comparison of Structural Continuity

Figure 1: (**a**) Echocardiographic FoV outpainting task and (**b**) qualitative comparison.

The difficulty of echocardiographic FoV outpainting is not merely to generate realistic textures. In TTE, weak boundaries and fragmented structures often degrade image quality [5]. Thus, the synthesized region must remain structurally compatible with the observed FoV, with smooth transitions, coherent contour continuation, and anatomically plausible extension. A visually smooth result may still be unsatisfactory if it breaks contour continuity. Therefore, for this task, the central issue is not only how to fill the missing area, but more importantly how to preserve structural continuity during FoV completion.

This observation also clarifies an important limitation of existing echocardiographic outpainting research. Current methods mainly emphasize image-level realism [5], which is insufficient for noisy ultrasound images where boundary patterns and contour transitions are critical. Existing approaches show the feasibility of FoV extension, but do not fully model structural continuity during generation, as shown in Fig. 1b. Once this limitation is recognized, a key methodological question naturally arises: how to effectively model structure-sensitive FoV outpainting in echocardiography?

In this work, we argue that diffusion models provide a more suitable foundation for this problem than conventional GAN-based reconstruction. Beyond strong generation, diffusion models align well with echocardiographic outpainting because they synthesize images through iterative denoising [7,8], enabling progressive structural refinement. This makes them suitable for structure-sensitive FoV completion, where the missing region must be refined while remaining consistent with observed structures. Based on this observation, we first establish a diffusion-based baseline for echocardiographic FoV outpainting, in which the incomplete input image is used as a condition and a complete FoV image is progressively generated through iterative denoising, as shown in Fig. 2.

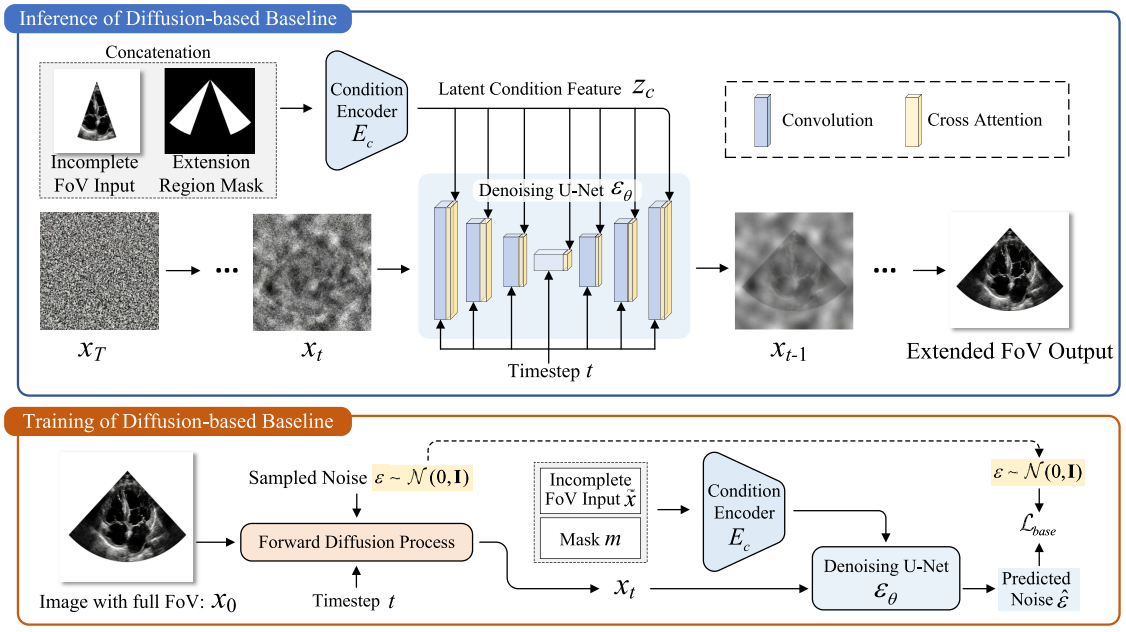


Figure 2: Diffusion-based baseline for echocardiographic FoV outpainting. (**Top**) shows the inference process. (**Bottom**) shows the training process.

Nevertheless, simply replacing a cGAN with a diffusion model is still not enough. The conditioning signal should also encode *informative structural cues* to guide contour coherence and boundary continuity. One possible solution is to use an external segmentor or preprocessing model to explicitly extract anatomical structures for the input image, but it makes the condition dependent on another model and may discard details needed for echocardiographic completion.

Motivated by this observation, we choose to learn structural conditions in an implicit manner. Specifically, we introduce a Structural Cue Encoder (SCE) to learn structure-oriented representations from the incomplete image itself, making latent features more sensitive to boundaries and contours while preserving image information. During training, an auxiliary decoder reconstructs pseudo-structural targets generated by fixed operators, using them only as weak supervision rather than inference-time inputs. This design keeps inference simple and distills noisy low-level responses into robust latent cues for diffusion-based FoV outpainting.

Once latent structural cues are extracted, the remaining challenge is how to use them to reconstruct the missing structure. Structure-sensitive conditions alone are insufficient; the denoising process must also recover the unknown region coherently. Since standard diffusion training mainly relies on noise prediction and appearance modeling, it lacks explicit structural guidance. Therefore, structural constraints should be introduced into diffusion learning to better guide the reconstruction of missing regions.

This consideration motivates the second core component of our method, namely the **Structure-aware Diffusion Learning (SDL)** scheme. Instead of only regularizing the final output, SDL imposes progressive structural constraints during denoising, allowing structural cues to guide the refinement process. Specifically, SDL promotes edge consistency, seam continuity, and high-frequency consistency, thereby improving contour coherence, boundary smoothness, and local structural fidelity in FoV outpainting.

Based on the above designs (i.e., diffusion-based baseline, SCE, and SDL), we propose a structure-aware diffusion framework for echocardiographic FoV outpainting. To the best of our knowledge, this is the first

diffusion-based method for this task, moving beyond the existing cGAN-based formulation represented by echoGAN [5]. Our framework aims to make structural continuity a central objective of FoV completion, rather than treating the task as generic image-level reconstruction.

It is worth noting that structure-aware generation has also been investigated before the recent rise of diffusion models. In particular, GAN-based image-to-image translation has provided a broad foundation for learning mappings between different image domains, including paired translation, unpaired translation, edge-to-image synthesis, and modality conversion [9,10]. Beyond generic adversarial translation, subsequent studies have further incorporated structural priors into GAN-based generation. For example, structure-aware adversarial learning has been explored for infrared image translation [11], edge-guided adversarial learning has been used for medical image translation [12], and gradient-guided GAN-based super-resolution has been developed to reduce geometric distortion and preserve image structures [13]. These studies demonstrate that structural information is important for generative modeling and should not be regarded as exclusive to diffusion-based methods. However, most existing structure-aware GAN studies are designed for general image translation, medical modality translation, or image enhancement, rather than echocardiographic FoV outpainting. In this task, the model must recover missing peripheral regions under speckle noise, weak boundaries, and cone-shaped ultrasound acquisition constraints, while maintaining continuity across the observed–generated interface. Therefore, our work focuses on how to formulate structure-aware FoV extension within a diffusion-based framework tailored to echocardiographic images. The main contributions of this work are summarized as follows:

- We introduce a diffusion framework for progressive TTE FoV extension.
- We devise an SCE to learn structural cues from incomplete echocardiographic inputs.
- We develop an SDL to impose structural constraints during diffusion training.
- Experiments show clear superiority over previous GAN-based reconstruction.

2 Related Work

2.1 Echocardiographic Image Generation

Generative modeling has attracted increasing attention in ultrasound imaging, where data synthesis can support image simulation, augmentation, and downstream analysis [14]. Early work was mainly dominated by GANs and conditional variants [14,15], which synthesize realistic ultrasound appearances under different acquisition or anatomical settings. In echocardiography, such models are useful due to limited data, strong variability, and acquisition-dependent artifacts. Recently, echoGAN [5] extended this direction to TTE FoV outpainting by formulating view extension as a cGAN-based reconstruction problem. It uses domain-aware augmentation to simulate cone-shaped ultrasound FoV and outpaint occluded side regions, broadening the observable area while preserving realistic echocardiographic appearance.

Alongside GAN-based approaches, diffusion-based ultrasound generation has begun to emerge [16,17]. Echo from Noise [16] used DDPMs guided by cardiac semantic maps to synthesize ultrasound images for segmentation, showing the value of anatomical guidance. Latent-diffusion-based ultrasound generation [17] has also been explored by fine-tuning large diffusion models on ultrasound datasets. These studies suggest that diffusion models are promising alternatives to adversarial learning in ultrasound generation. However, despite this progress, diffusion models have not yet been introduced into the clinically important task of echocardiographic FoV outpainting.

Difference from prior work. Existing ultrasound generation studies [16,17] mainly focus on image synthesis or augmentation, while echoGAN [5] introduces echocardiographic FoV outpainting as a dedicated task. Following this line, we extend FoV extension from the current cGAN-based formulation to a diffusion-based framework, with stronger emphasis on structural continuity.

2.2 Image Outpainting and FoV Extension

Image outpainting aims to infer plausible content outside the observed image boundary and has been widely studied in natural-image generation [3,4]. In medical imaging, it is more challenging because synthesized regions must be both realistic and anatomically meaningful [18]. This challenge is greater in ultrasound, where weak boundaries, speckle noise, and structural ambiguity make coherent generation difficult [5,19]. In this context, echoGAN is important because it formulates TTE FoV extension as an outpainting task and shows that missing side regions can be reconstructed using cone-shaped ultrasound-specific masking patterns [5].

Beyond echocardiography, related ideas have also appeared in broader medical imaging [20]. Structure- or pathology-preserving outpainting has been explored in modalities such as MRI, where maintaining clinically meaningful content is essential [21]. These studies suggest that medical outpainting is gradually shifting from generic visual completion toward structure-aware and clinically constrained completion [18].

Difference from prior work. Existing medical outpainting literature has shown the feasibility of extending truncated views or recovering missing image regions [22,23]. However, echocardiographic outpainting still mainly focuses on image-level reconstruction [5,24]. Our method differs in that it treats structural continuity as the main modeling target, rather than only recovering visually plausible missing regions.

2.3 Structure-Aware Generative Modeling

Beyond echocardiographic image generation, structure-aware generative modeling has also been explored in several GAN-based tasks. Early image-to-image translation frameworks demonstrated that adversarial learning can be used to learn mappings between different visual domains and can support tasks such as edge-to-image synthesis, paired translation, and unpaired translation [9,10]. Building on this foundation, later studies introduced more explicit structural mechanisms into adversarial generation. For example, StawGAN incorporated structural awareness into infrared image translation to improve the shape and definition of generated objects [11]. In medical imaging, edge-guided GAN-based translation has been explored to preserve anatomical structures across imaging domains [12]. In image super-resolution, structure-preserving GAN methods have used gradient guidance or learned structure extractors to reduce geometric distortion while maintaining perceptually realistic details [13]. These studies indicate structure-aware generation is a broad research direction that predates diffusion models.

Nevertheless, these methods are mainly developed for image translation, modality conversion, or image enhancement, and they do not directly address the specific challenges of echocardiographic FoV outpainting.

Difference from prior work. Compared with these structure-aware GAN-based methods, our work focuses on the echocardiographic FoV extension setting and further exploits the iterative denoising mechanism of diffusion models to impose structural constraints on intermediate clean-image estimates.

2.4 Advanced Diffusion-Based Image Enhancement

Recent diffusion-based image enhancement studies have shown strong potential in image super-resolution, restoration, and high-resolution synthesis. A representative early work is SR3, which formulates image super-resolution as an iterative refinement process and progressively generates high-resolution images conditioned on low-resolution inputs [25]. Beyond standard super-resolution, DDRM further demonstrates that diffusion priors can be used for general image restoration and linear inverse problems, including super-resolution, deblurring, inpainting, and colorization [26]. These studies indicate that diffusion models are effective not only for unconditional generation, but also for recovering degraded or missing visual content through conditional denoising.

More recent studies have explored how to better exploit diffusion priors, multi-scale representations, and efficient sampling strategies in high-resolution restoration. For example, StableSR leverages the generative prior of pre-trained text-to-image diffusion models for real-world image super-resolution, and introduces a time-aware encoder and controllable feature wrapping to balance perceptual quality and fidelity [27]. Multi-scale adversarial diffusion further introduces multi-scale generation guidance to learn feature information at different scales from low-resolution images, thereby enhancing representation capability for super-resolution [28]. These studies show that diffusion priors, multi-scale guidance, and efficient sampling strategies are important directions for preserving fine details and improving the scalability of diffusion-based image enhancement.

Difference from prior work. Although these diffusion-based enhancement methods are relevant to FoV extension, our task is different from conventional image super-resolution or restoration. Most existing methods aim to improve the resolution or quality of already observed image content, whereas echocardiographic FoV outpainting requires the model to synthesize unobserved peripheral anatomical regions outside the limited ultrasound FoV. Moreover, the generated region must remain structurally coherent with the observed region across the mask boundary. Therefore, our work focuses on structure-aware FoV completion and introduces SCE and SDL to enhance boundary continuity, contour coherence, and structural consistency during iterative denoising.

2.5 Neural Networks and Machine Learning in Medical Applications

Neural networks and machine learning have been increasingly used in a wide range of biomedical and clinical analysis tasks, which suggests that learning-based methods can provide useful computational tools for extracting patterns from complex biomedical data and supporting disease-related analysis.

In medical imaging, deep learning methods have also been widely studied for image interpretation, segmentation, classification, and computer-aided diagnosis across different imaging modalities [29]. Recent clinical reviews further suggest that deep learning-based image analysis has the potential to support disease diagnosis and improve medical imaging workflows, while also emphasizing the importance of clinical validation, interpretability, and safe deployment [30]. These studies demonstrate the broader utility of neural networks and machine learning in clinically relevant medical applications.

Difference from prior work. Different from the above studies that mainly focus on general biomedical analysis, disease diagnosis, medical image interpretation, or clinical decision support, this work investigates a specific echocardiographic image generation problem, namely field-of-view extension for TTE images. Our goal is not to directly perform disease prediction or diagnostic classification, but to improve the visual completeness and structural continuity of echocardiographic images through structure-aware diffusion-based FoV outpainting.

3 Method

3.1 Problem Formulation

Following the task setting in echoGAN [5], we construct echocardiographic FoV outpainting from paired complete and incomplete images. During training, the complete image is used as the supervision target, denoted by $x \in \mathbb{R}^{H\times W}$. A binary mask $m \in \{0,1\}^{H\times W}$ indicates the observed region, where $m_{ij} = 1$ retains pixel (i, j) and $m_{ij} = 0$ removes it. The incomplete input is defined as

$$\tilde{x} = m \odot x, \tag{1}$$

where $\odot$ denotes element-wise multiplication. In this way, x serves as the reconstruction target, while $\tilde{x}$ is the input to the FoV outpainting model.

The goal of echocardiographic FoV outpainting is to recover a complete image $\hat{x}$ from the incomplete observation $\tilde{x}$:

$$\hat{x} = \mathcal{F}(\tilde{x}, m), \tag{2}$$

where $\mathcal{F}$ denotes the outpainting model. The reconstruction should preserve the observed region and plausibly complete the missing region.

In this paper, our framework consists of two main parts: a *condition encoder* for extracting conditioning information from the incomplete image and a *denoising U-Net* for conditional diffusion generation. We first describe the **basic diffusion baseline**, composed of the *condition encoder* and *denoising U-Net*, and then introduce two dedicated designs: **SCE**, an improved *condition encoder*, and **SDL**, an improved learning scheme for the *denoising U-Net*.

3.2 Diffusion-Based Baseline

We first establish a diffusion-based baseline for echocardiographic FoV outpainting, which serves as the foundation of the proposed framework.

3.2.1 Input and Output

As shown in Fig. 2, the baseline takes the incomplete image $\tilde{x}$ and mask m as inputs. During training, it predicts the noise $\hat{\epsilon}$, while at inference it progressively reconstructs a complete FoV image $\hat{x}_0$ from Gaussian noise conditioned on the input representation z_c.

3.2.2 Network Formulation

The baseline consists of a condition encoder E_c and a denoising U-Net ϵ_θ. Given $\tilde{x}$ and m, the condition encoder extracts a latent condition feature

$$z_c = E_c(\tilde{x}, m), \tag{3}$$

which is injected into the denoising U-Net for conditional generation.

Let x_0 denote the clean complete FoV image sampled from the data distribution $\mathcal{D}(x)$. Following the standard diffusion formulation [7], a noisy sample at timestep $t \in \{1, \ldots, T\}$ is generated by

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\ \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{4}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1 - \beta_s)$ and $\{\beta_t\}_{t=1}^{T}$ is a predefined noise schedule.

As shown in Fig. 2 (bottom), given the noisy sample x_t, the denoising U-Net predicts the injected noise conditioned on the incomplete image representation:

$$\hat{\epsilon} = \epsilon_\theta\ (x_t, z_c, t). \tag{5}$$

This follows the standard diffusion formulation [7], where accurate noise prediction helps recover the underlying clean image for complete FoV generation.

3.2.3 Training Objective

Based on the above formulations, the baseline is trained with the standard noise prediction loss:

$$\mathcal{L}_{\text{base}} = \mathbb{E}_{x_0,\epsilon,t}\left[\| \epsilon - \epsilon_\theta\left(x_t, z_c, t\right)\|_2^2\right]. \tag{6}$$

This baseline is not the final form of our framework, but serves as the foundation for later designs. It can be viewed as a pre-training stage that provides an initial conditional diffusion model, upon which the SCE replacement and SDL optimization are built.

3.3 Structural Cue Encoder (SCE)

The baseline condition encoder only provides general conditional information for diffusion generation. However, echocardiographic FoV outpainting requires more informative structural cues. Therefore, we replace the baseline condition encoder E_c with a Structural Cue Encoder (SCE), denoted by E_s.

3.3.1 Structural Encoding

Given the incomplete input $\tilde{x}$ and mask m, SCE extracts a structure-sensitive representation $z_s = E_s(\tilde{x}, m)$ to replace the generic condition z_c. The denoising U-Net is then reformulated as $\hat{\epsilon} = \epsilon_\theta\left(x_t, z_s, t\right)$.

SCE is a lightweight multi-branch convolutional encoder for structural representation learning. As shown in Fig. 3, it combines standard convolutions for appearance encoding, directional convolutions with asymmetric kernels [31] (e.g., 1×3 and 3×1) for orientation-sensitive boundaries, and dilated convolutions [32] for longer-range contour dependencies. The fused multi-scale and multi-directional features form a compact structural embedding for diffusion-based FoV outpainting.

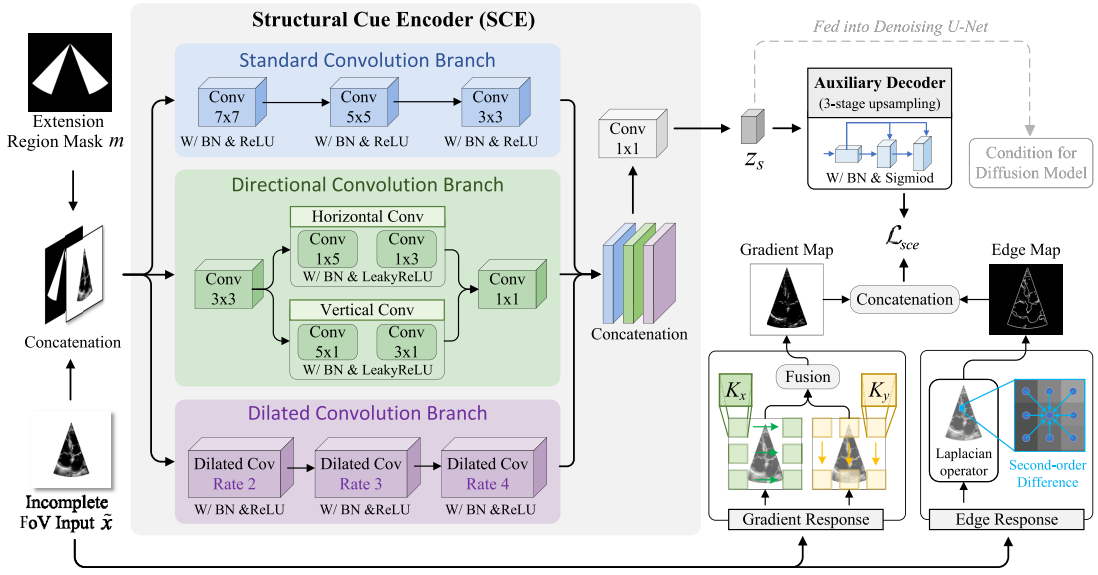


Figure 3: Structural cue encoder (SCE). (**Left**) shows the architecture of the proposed SCE, while (**right**) shows its representation learning process.

3.3.2 Representation Learning

To encourage z_s to capture structural information, we attach an auxiliary decoder D_s during training and supervise the encoder–decoder branch with a pseudo-structural target derived from $\tilde{x}$. This target combines gradient- and edge-based responses.

For gradient response, inspired by [33], we use fixed horizontal and vertical kernels to capture boundary variations. The derivatives are computed as $G_x = K_x * \tilde{x}$ and $G_y = K_y * \tilde{x}$, where $*$ denotes convolution and K_x, K_y are fixed directional operators. The gradient magnitude is defined as

$$G_{\text{mag}}(\tilde{x}) = \sqrt{(K_x * \tilde{x})^2 + (K_y * \tilde{x})^2 + \varepsilon}, \tag{7}$$

where ε is a small constant for numerical stability. We then compute a masked normalized gradient map

$$\bar{G}(\tilde{x}) = \frac{m \odot G_{\text{mag}}(\tilde{x})}{\max(G_{\text{mag}}(\tilde{x})) + \varepsilon}, \tag{8}$$

which suppresses invalid responses outside the observed FoV and normalizes the gradient scale, yielding a more robust pseudo-structural target.

For edge response, we use a Laplacian operator to highlight contour-sensitive second-order variations. For each pixel (i, j), the response is computed as

$$L(\tilde{x})_{i,j} = \tilde{x}_{i-1,j} + \tilde{x}_{i+1,j} + \tilde{x}_{i,j-1} + \tilde{x}_{i,j+1} - 4\tilde{x}_{i,j}, \tag{9}$$

which is then normalized as

$$\bar{E}(\tilde{x}) = \sigma\left(\frac{m \odot L(\tilde{x})}{\tau}\right), \tag{10}$$

where $\sigma(\cdot)$ is the sigmoid function and τ is a temperature parameter. The final pseudo-structural target is obtained by channel-wise concatenation: $s^* = \text{Concat}\left(\bar{G}(\tilde{x}),\ \bar{E}(\tilde{x})\right)$.

3.3.3 Training Objective

The decoder reconstructs this target from the latent representation, i.e., $\hat{s} = D_s(z_s)$. We define the structural reconstruction loss as

$$\mathcal{L}_{\text{sce}} = \|\hat{s} - s^*\|_1 + \left(\|\nabla_x \hat{s} - \nabla_x s^*\|_1 + \|\nabla_y \hat{s} - \nabla_y s^*\|_1\right), \tag{11}$$

where ∇_x and ∇_y denote horizontal and vertical finite-difference operators.

This auxiliary supervision encourages the encoder to learn structure-sensitive latent features that capture boundary and edge variations, providing more informative conditions for FoV outpainting.

We note that the gradient- and Laplacian-based pseudo-structural targets are not the only possible choices for SCE supervision. Frequency-domain and multi-scale representations, such as wavelet-based high-frequency sub-bands, can also be used to provide structural cues that are potentially robust to ultrasound speckle noise. Previous studies [34,35] have shown that wavelet-based and multi-scale methods are effective for ultrasound despeckling and medical image denoising because they can separate noise-dominated components from meaningful structural information while preserving edges and diagnostic details. To examine this possibility, we further implemented a wavelet-based SCE variant and compared it with the proposed design in the ablation study.

3.4 Structure-Aware Diffusion Learning (SDL)

Although SCE provides more informative structural conditions for diffusion-based FoV outpainting, this alone is still insufficient. Echocardiographic FoV completion also requires the missing region to be reconstructed in a structure-preserving manner under these cues. To this end, we further develop a Structure-aware Diffusion Learning (SDL) scheme to enhance the denoising U-Net itself.

3.4.1 Intermediate Clean-Image Estimation

In standard diffusion training [7], the denoising U-Net predicts the injected noise rather than directly generating an image-domain output. Given x_t and z_s, it outputs $\hat{\epsilon} = \epsilon_\theta\,(x_t, z_s, t)$. Since our structural supervision is defined in the image domain, we first recover an intermediate clean-image estimate from the current diffusion state.

Following the standard diffusion formulation, at an arbitrary timestep t, the corresponding clean-image estimate can be reconstructed as

$$\hat{x}_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}}. \tag{12}$$

The reconstructed $\hat{x}_0$ serves as an image-domain proxy of the current denoising result at timestep t, enabling structural constraints to be imposed on intermediate denoising results [36].

3.4.2 Progressive Structural Constraints

Based on the intermediate clean-image estimate $\hat{x}_0$, we impose step-wise structural constraints to encourage coherent FoV completion. SDL regularizes $\hat{x}_0$ from three complementary aspects: edge consistency, seam continuity, and high-frequency consistency.

(1) Edge consistency. We enforce consistency between the edge activations of $\hat{x}_0$ and x_0. Using the same Laplacian-style operator as in SCE, the edge consistency term is

$$\mathcal{L}_{\text{edge}} = \left\| \bar{E}(\hat{x}_0) - \bar{E}(x_0) \right\|_1, \tag{13}$$

where $\bar{E}(\cdot)$ denotes the normalized edge response. This term encourages the model to recover contour-sensitive transitions in the missing region.

(2) Seam continuity. FoV outpainting requires smooth transition across the observed–generated boundary. Let $B(m)$ denote the mask boundary, and let Ω_b be a narrow band centered on $B(m)$. The seam continuity term is

$$\mathcal{L}_{\text{seam}} = \frac{1}{|\Omega_b|} \sum_{(i,j)\in\Omega_b} \left(\left| \nabla_x \hat{x}_{0,ij} - \nabla_x x_{0,ij} \right| + \left| \nabla_y \hat{x}_{0,ij} - \nabla_y x_{0,ij} \right| \right), \tag{14}$$

where ∇_x and ∇_y are horizontal and vertical finite-difference operators. This term promotes structural continuity near the observed–generated interface.

(3) High-frequency consistency. Preserving fine local details is also important for echocardiographic FoV outpainting. Let $\mathcal{H}(\cdot)$ denote a high-pass filtering operator. The high-frequency constraint is

$$\mathcal{L}_{\text{hf}} = \left\| \mathcal{H}(\hat{x}_0) - \mathcal{H}(x_0) \right\|_1. \tag{15}$$

This term helps maintain local sharpness and reduces over-smoothed completion.

3.4.3 Training Objective

By combining the above constraints, the SDL loss is formulated as

$$\mathcal{L}_{\text{sdl}} = \mathcal{L}_{\text{edge}} + \mathcal{L}_{\text{seam}} + \mathcal{L}_{\text{hf}}. \tag{16}$$

Thus, SDL injects structural bias into diffusion training through intermediate clean-image estimates, enabling more structure-preserving reconstruction of the missing region.

3.5 Training and Inference

Training. The framework is trained in three stages. First, the baseline encoder E_c and denoising U-Net ϵ_θ are optimized with $\mathcal{L}_{\text{base}}$ (Eq. (6)). Second, SCE E_s is trained with $\mathcal{L}_{\text{sce}}$ (Eq. (11)) and then replaces E_c. Third, with E_s fixed, the denoising U-Net is further optimized using SDL with $\mathcal{L}_{\text{sdl}}$ (Eq. (16)).

Inference. During inference, only SCE and the denoising U-Net are used. Given $\tilde{x}$ and m, SCE produces $z_s = E_s(\tilde{x}, m)$, and the denoising U-Net performs reverse diffusion conditioned on z_s to generate the complete FoV image.

4 Experiments

4.1 Datasets

We follow the setting of echoGAN [5] to conduct experiments on two public datasets.

CAMUS [37]. CAMUS is a public 2D echocardiography benchmark with apical two-chamber (2CH) and four-chamber (4CH) images from 500 patients. Its standardized views and clear structures make it suitable for evaluating structural coherence in FoV outpainting.

EchoNet-Dynamic [38]. EchoNet-Dynamic contains 10,030 apical four-chamber (A4C) videos with clinical labels and expert tracings. Compared with CAMUS, it is larger and more variable, providing a realistic benchmark for evaluating model robustness.

4.2 Implementation Details

Following echoGAN, incomplete FoV inputs were generated by masking peripheral side regions while preserving the central observed region. All experiments were implemented in PyTorch on a single NVIDIA 3090 GPU. We used $T = 1000$ diffusion steps with a linear noise schedule from 10^{-4} to 2×10^{-2}. The baseline was trained with AdamW using a learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, and batch size of 8. For SCE, Sobel kernels were used as K_x and K_y, with $\varepsilon = 10^{-6}$ and $\tau = 0.1$. For SDL, the seam band width was set to 5 pixels, and the denoising U-Net was further optimized with AdamW at a learning rate of 5×10^{-5}.

4.3 Evaluation Metrics

Following echoGAN [5], we use **FID** [39] as the main metric and report **LPIPS** [40], **PSNR** [41], and **SSIM** [42] for image-level comparison. To assess structural quality, we further report **FSIM** [43], **GMSD** [44], and **PFOM** [41] to measure feature similarity, gradient distortion, and edge preservation, respectively.

4.4 Quantitative Results

4.4.1 Typical Evaluation

Table 1 reports quantitative results under four FoV outpainting settings, namely *Cut 30*, *Cut 46*, *Cut 60*, and *Cut 80*, on CAMUS and EchoNet-Dynamic.

Table 1: Quantitative comparison under different outpainting sizes. Best results are in bold.

Method	Metric	CAMUS				EchoNet-Dynamic			
		Cut 30	Cut 46	Cut 60	Cut 80	Cut 30	Cut 46	Cut 60	Cut 80
echoGAN [5]	FID↓	122.08	143.83	145.74	149.63	60.24	74.33	89.21	110.42
	LPIPS↓	0.053	0.09	0.13	0.12	0.16	0.16	0.16	0.22
	PSNR↑	23.91	21.46	19.77	19.95	20.15	20.14	20.13	17.89
	SSIM↓	0.28	0.31	0.47	0.46	0.48	0.48	0.48	0.59
CoMoDGAN [24]	FID↓	109.59	158.68	208.59	194.03	75.94	88.67	106.77	132.81
	LPIPS↓	0.1642	0.21	0.27	0.26	0.21	0.25	0.29	0.35
	PSNR↑	11.59	9.31	7.54	7.79	7.04	7.03	7.03	6.99
	SSIM↓	0.57	0.70	0.80	0.80	0.78	0.82	0.85	0.90
TFill [19]	FID↓	99.80	163.05	240.32	273.60	62.49	77.94	78.28	112.28
	LPIPS↓	0.26	0.37	0.46	0.51	0.14	0.21	0.258	0.33
	PSNR↑	15.57	12.90	11.58	10.24	22.34	19.84	17.85	15.74
	SSIM↓	0.34	0.48	0.60	0.63	0.19	0.29	0.39	0.53
Ours	FID↓	**92.41**	**104.67**	**112.58**	**118.91**	**55.73**	**66.85**	**71.48**	**96.35**
	LPIPS↓	**0.041**	**0.071**	**0.097**	**0.095**	**0.121**	**0.145**	**0.148**	**0.187**
	PSNR↑	**25.12**	**22.84**	**20.96**	**20.74**	**23.06**	**20.93**	**20.76**	**18.84**
	SSIM↓	**0.24**	**0.27**	**0.40**	**0.38**	**0.16**	**0.24**	**0.31**	**0.44**

Overall, our method achieves the best performance across both datasets and all cut settings. Compared with echoGAN [5], the average FID is reduced from 140.32 to 107.14 on CAMUS and from 83.55 to 72.60 on EchoNet-Dynamic, corresponding to relative improvements of **23.6%** and **13.1%**. Averaged over all settings, FID is reduced by about **18.4%**, demonstrating the advantage of the proposed diffusion-based formulation.

LPIPS and PSNR show similar improvements. Compared with echoGAN, our method reduces LPIPS by **22.6%** on CAMUS and **14.1%** on EchoNet-Dynamic, while improving PSNR by **1.15** and **1.32 dB**, respectively. These gains are more evident under larger cut settings, indicating robustness when peripheral FoV completion becomes more challenging.

As shown in Fig. 4 (left), our method is also more stable than CoMoDGAN [24] and TFill [19] as outpainting difficulty increases. Although TFill is competitive in easier settings, especially on EchoNet-Dynamic Cut 30, its performance drops faster under larger cuts. CoMoDGAN shows a similar trend, particularly on CAMUS. In contrast, our method consistently maintains the best overall results, supporting the robustness and generality of the proposed framework.

4.4.2 Structure-Aware Evaluation

Although the results in Table 1 show clear advantages in image-level metrics, these metrics mainly reflect visual quality and perceptual fidelity. To further assess structural quality, we additionally report FSIM, GMSD, and PFOM, as detailed in Section 4.3.

As shown in Table 2, our method achieves the best performance across all three structural metrics on both datasets and all cut settings. On CAMUS, compared with echoGAN [5], the average FSIM improves from 0.772 to 0.834, GMSD decreases from 0.105 to 0.078, and PFOM increases from 0.712 to 0.799, corresponding to gains of **8.0%**, **25.7%**, and **12.1%**, respectively. Similar improvements are observed on

EchoNet-Dynamic, where FSIM increases from 0.803 to 0.859, GMSD decreases from 0.097 to 0.078, and PFOM improves from 0.747 to 0.825.

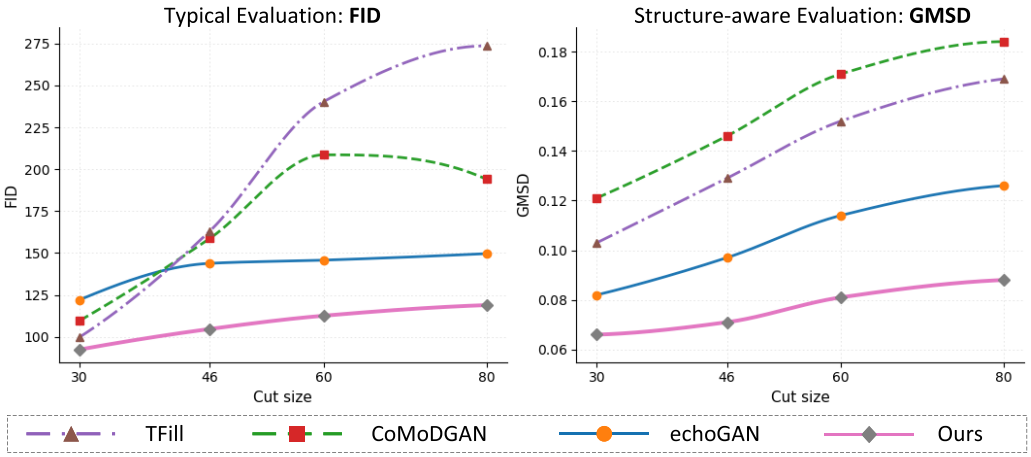


Figure 4: Performance variation curves on CAMUS under different outpainting (cut) sizes. (**Left**) reports FID, a typical image-level evaluation metric, where lower values indicate better visual fidelity and distributional similarity. (**Right**) reports GMSD, a structure-aware metric, where lower values indicate better gradient consistency and structural preservation. Overall, our method achieves the lowest FID and GMSD across different cut sizes, demonstrating superior visual quality and structural robustness under increasing outpainting difficulty.

Table 2: Structure-aware evaluation under different outpainting sizes. Best results are in bold.

Method	Metric	CAMUS				EchoNet-Dynamic			
		Cut 30	Cut 46	Cut 60	Cut 80	Cut 30	Cut 46	Cut 60	Cut 80
echoGAN [5]	FSIM↑	0.836	0.791	0.742	0.718	0.861	0.824	0.786	0.742
	GMSD↓	0.082	0.097	0.114	0.126	0.074	0.089	0.104	0.121
	PFOM↑	0.781	0.732	0.684	0.651	0.814	0.768	0.721	0.683
CoMoDGAN [24]	FSIM↑	0.701	0.642	0.588	0.561	0.722	0.681	0.639	0.601
	GMSD↓	0.121	0.146	0.171	0.184	0.116	0.133	0.151	0.169
	PFOM↑	0.612	0.554	0.497	0.469	0.641	0.598	0.556	0.517
TFill [19]	FSIM↑	0.763	0.701	0.628	0.596	0.846	0.792	0.751	0.703
	GMSD↓	0.103	0.129	0.152	0.169	0.081	0.098	0.117	0.136
	PFOM↑	0.703	0.641	0.578	0.541	0.793	0.741	0.696	0.652
Ours	FSIM↑	**0.887**	**0.852**	**0.812**	**0.784**	**0.913**	**0.879**	**0.842**	**0.801**
	GMSD↓	**0.066**	**0.071**	**0.081**	**0.093**	**0.059**	**0.071**	**0.084**	**0.098**
	PFOM↑	**0.862**	**0.817**	**0.774**	**0.741**	**0.889**	**0.846**	**0.803**	**0.761**

These results support our main claim that the proposed method not only generates plausible outpainting results, but also better preserves structural patterns, boundary transitions, and contour continuity. Moreover, as shown in Fig. 4 (right), our method remains more stable as the cut size increases from Cut 30 to Cut 80. Its average FSIM drop is **11.9%**, smaller than echoGAN [5] (14.0%), TFill [19] (19.4%), and CoMoDGAN [24] (18.4%), with similar trends observed for GMSD and PFOM.

4.4.3 View-Wise Evaluation

In the original quantitative comparison, CAMUS was evaluated as a combined dataset containing both apical two-chamber (2CH) and apical four-chamber (4CH) images. To further examine whether the proposed method performs consistently across different TTE views, we conduct a view-wise analysis by separately evaluating CAMUS-2CH and CAMUS-4CH under the representative Cut 60 setting.

As shown in Table 3, the proposed method consistently outperforms echoGAN [5] on both CAMUS-2CH and CAMUS-4CH. On CAMUS-2CH, our method reduces FID from 147.82 to 114.03 and LPIPS from 0.134 to 0.099, while improving PSNR from 19.61 to 20.81. In terms of structure-aware metrics, our method also improves FSIM from 0.736 to 0.807, reduces GMSD from 0.115 to 0.083, and improves PFOM from 0.678 to 0.767. Similar improvements can also be observed on CAMUS-4CH. These results indicate that the proposed structure-aware diffusion framework can provide consistent improvements across different apical TTE views in CAMUS, not only in image-level quality but also in structural continuity and boundary preservation.

Table 3: View-wise evaluation on CAMUS under the Cut 60 setting. 2CH and 4CH denote apical two-chamber and apical four-chamber views, respectively.

View	Method	FID ↓	LPIPS ↓	PSNR ↑	SSIM ↓	FSIM ↑	GMSD ↓	PFOM ↑
CAMUS-2CH	echoGAN [5]	147.82	0.134	19.61	0.48	0.736	0.115	0.678
	Ours	114.03	0.099	20.81	0.41	0.807	0.083	0.767
CAMUS-4CH	echoGAN [5]	143.66	0.126	19.93	0.46	0.748	0.112	0.689
	Ours	111.12	0.095	21.10	0.39	0.817	0.078	0.780

4.5 Qualitative Results

To further verify the advantage of the proposed method in preserving structural continuity, we provide qualitative comparisons on representative cases. As shown in Fig. 5, we compare TFill [19], CoMoDGAN [24], echoGAN [5], our method, and the ground truth.

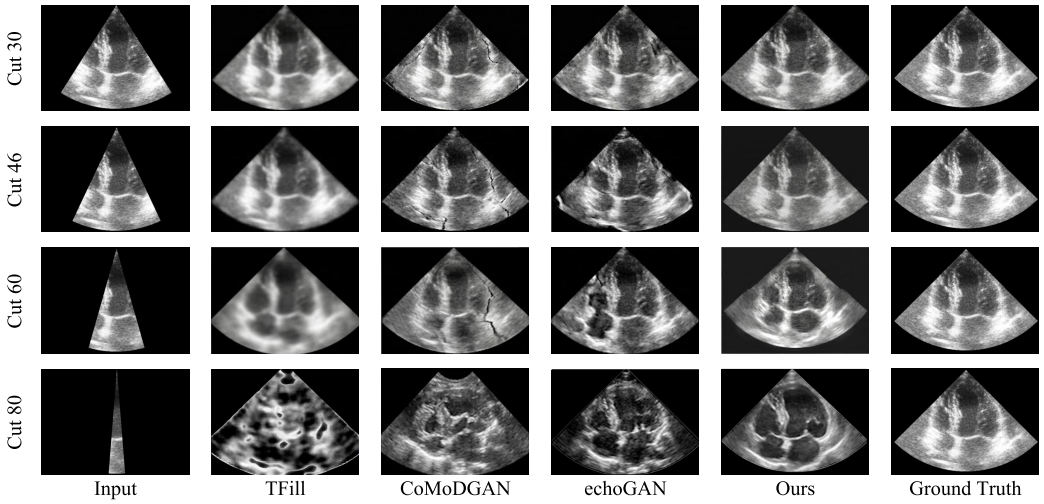


Figure 5: Qualitative comparison under increasing missing severity.

Overall, previous methods still show clear limitations in echocardiographic FoV outpainting. TFill [19] can recover plausible coarse structures, but often produces over-smoothed regions with ambiguous details. GAN-based methods [5,24] frequently show discontinuities near the observed–generated boundary, such as crack-like artifacts, contour distortion, and unsmooth seams. These results suggest that image-level reconstruction or long-range dependency modeling alone is insufficient for fine structural consistency.

By contrast, the proposed method yields smoother structural transitions and more coherent contour continuation from the observed region to the restored region. Compared with the ground truth, the generated peripheral structures by our method are visually more consistent.

4.6 Ablation Study

4.6.1 Ablation on Main Components

We first evaluate the effectiveness of the main components by progressively adding SCE and SDL on top of the diffusion-based baseline. The results are reported in Table 4. Introducing SCE consistently improves both conventional and structure-aware metrics, indicating that structure-sensitive conditions provide more informative guidance for FoV completion. Adding SDL further improves performance, showing the benefit of imposing structural regularization during denoising. The full model achieves the best results, confirming that SCE and SDL are complementary.

Table 4: Ablation study on the main components under the Cut 60 setting on EchoNet-Dynamic.

Method	FID ↓	LPIPS ↓	PSNR ↑	SSIM ↓	FSIM ↑	GMSD ↓	PFOM ↑
Baseline	82.94	0.171	19.88	0.37	0.791	0.101	0.748
Baseline + SCE	77.63	0.160	20.21	0.34	0.816	0.093	0.776
Baseline + SDL	75.84	0.156	20.36	0.33	0.822	0.090	0.783
Full model	71.48	0.148	20.76	0.31	0.842	0.084	0.803

Compared with the baseline, the full model reduces FID by **13.8%** and LPIPS by **13.5%**, improves PSNR by **0.88 dB**, and reduces SSIM by **16.2%**. It also improves FSIM and PFOM by **6.4%** and **7.4%**, while reducing GMSD by **16.8%**. These results show that SCE+SDL improves both image-level quality and structural continuity.

4.6.2 Ablation on Designs in SCE

We further analyze SCE by removing each branch from the complete multi-branch encoder. As shown in Table 5, removing any branch degrades performance, indicating that the standard, directional, and dilated branches provide complementary cues. The directional branch contributes most to structure-aware metrics, especially FSIM and PFOM, highlighting the importance of orientation-sensitive boundary modeling. The complete SCE achieves the best overall performance.

4.6.3 Ablation on Designs in SDL

Finally, we study the contribution of SDL constraints, including edge consistency, seam continuity, and high-frequency consistency. As shown in Table 6, removing any constraint degrades performance, confirming that all three terms are useful. Removing seam continuity causes the largest drop in most metrics, consistent with our motivation that smooth boundary transition is central to echocardiographic FoV outpainting. The full SDL design achieves the best results across all metrics.

Table 5: Ablation study on different branches of SCE under the Cut 60 setting on EchoNet-Dynamic.

SCE Variant	FID ↓	LPIPS ↓	PSNR ↑	SSIM ↓	FSIM ↑	GMSD ↓	PFOM ↑
w/o standard branch	75.12	0.156	20.31	0.33	0.824	0.091	0.784
w/o directional branch	76.48	0.159	20.18	0.35	0.818	0.094	0.777
w/o dilated branch	74.86	0.154	20.35	0.33	0.827	0.090	0.786
Full SCE	71.48	0.148	20.76	0.31	0.842	0.084	0.803

Table 6: Ablation study on different structure-aware constraints in SDL under the Cut 60 setting on EchoNet-Dynamic.

SDL Variant	FID ↓	LPIPS ↓	PSNR ↑	SSIM ↓	FSIM ↑	GMSD ↓	PFOM ↑
w/o $\mathcal{L}_{\text{edge}}$	74.33	0.153	20.41	0.33	0.829	0.089	0.791
w/o $\mathcal{L}_{\text{seam}}$	75.06	0.155	20.27	0.34	0.823	0.091	0.782
w/o $\mathcal{L}_{\text{hf}}$	73.81	0.151	20.48	0.32	0.832	0.088	0.794
Full SDL	71.48	0.148	20.76	0.31	0.842	0.084	0.803

4.6.4 Ablation on Alternative Structural Representations

To address the concern that the proposed SCE mainly relies on spatial-domain structural operators, we further examine an alternative frequency-/multi-scale-based structural representation. Specifically, we implement a wavelet-based variant, termed Wavelet-SCE. In the original SCE, the pseudo-structural target is constructed from gradient and Laplacian responses. In contrast, Wavelet-SCE replaces these spatial-domain pseudo-targets with wavelet-domain structural targets. Motivated by the wavelet-based ultrasound denoising analysis in [34], we apply discrete wavelet decomposition to the incomplete echocardiographic input and use the resulting high-frequency sub-bands as structural supervision. Furthermore, inspired by the multi-scale wavelet strategy in [35], we aggregate high-frequency sub-bands from multiple decomposition levels to form a multi-scale wavelet structural target for training the SCE. During inference, the auxiliary decoder and all pseudo-structural targets are discarded, and only the trained encoder is retained.

As shown in Table 7, Wavelet-SCE does not consistently outperform the original SCE. On CAMUS under the Cut 60 setting, Wavelet-SCE improves LPIPS from 0.097 to 0.096, FSIM from 0.812 to 0.815, and PFOM from 0.774 to 0.777. However, FID slightly increases from 112.58 to 113.06, and GMSD becomes worse from 0.081 to 0.083. On EchoNet-Dynamic, Wavelet-SCE slightly reduces FID from 71.48 to 70.92 and GMSD from 0.084 to 0.083, but LPIPS increases from 0.148 to 0.150, FSIM decreases from 0.842 to 0.840, and PFOM decreases from 0.803 to 0.801. These mixed results indicate that wavelet-based supervision can provide useful frequency-domain structural cues in some cases, but the improvement is limited and not stable across datasets and evaluation metrics.

Table 7: Exploratory comparison between the original SCE and the wavelet-based SCE variant under Cut 60 setting.

Dataset	Method	FID ↓	LPIPS ↓	FSIM ↑	GMSD ↓	PFOM ↑	Train Time	GPU Mem.
CAMUS	Original SCE	112.58	0.097	0.812	0.081	0.774	1.00×	1.00×
CAMUS	Wavelet-SCE	113.06	0.096	0.815	0.083	0.777	1.24×	1.16×
EchoNet-Dynamic	Original SCE	71.48	0.148	0.842	0.084	0.803	1.00×	1.00×
EchoNet-Dynamic	Wavelet-SCE	70.92	0.150	0.840	0.083	0.801	1.26×	1.18×

One possible reason is that the high-frequency sub-bands of ultrasound images contain both anatomical boundary information and speckle noise. Although wavelet decomposition can separate frequency components, the high-frequency wavelet responses may still amplify noise-dominated patterns when the echocardiographic boundaries are weak or fragmented. As a result, the wavelet-based pseudo-structural target may provide less stable supervision than the direct gradient- and Laplacian-based targets in some settings. This explains why Wavelet-SCE slightly improves certain metrics but degrades others.

Meanwhile, Wavelet-SCE introduces extra computational overhead due to multi-level wavelet decomposition, multiple high-frequency sub-band targets, and more complex multi-scale structural supervision. Specifically, the training time increases to approximately 1.24–1.26×, and GPU memory consumption increases to approximately 1.16–1.18× compared with the original SCE. Therefore, considering the trade-off between performance stability and computational complexity, we retain the original gradient- and Laplacian-based SCE as the main design. This choice provides a better balance among structural effectiveness, computational efficiency, and implementation simplicity.

4.6.5 Inference Efficiency and Quality–Speed Trade-Off

Since diffusion models generate images through iterative denoising, they are generally slower than GAN-based models that produce outputs in a single forward pass. This computational gap is important for potential real-time clinical applications. To analyze this issue, we further evaluate the trade-off between reconstruction quality and inference speed under the representative Cut 60 setting. All inference times are measured as the average time per image on a single NVIDIA RTX 3090 GPU.

Inspired by DDIM-based accelerated sampling [8], we evaluate several reduced-step inference variants of the trained diffusion model. Specifically, we evaluate the proposed method using 1000, 100, 50, and 25 denoising steps, denoted as Ours-1000, Ours-100, Ours-50, and Ours-25, respectively. As shown in Table 8, Ours-1000 achieves the best reconstruction quality but requires the longest inference time. When the number of steps is reduced to 100 or 50, the inference time is substantially decreased, while most of the reconstruction quality is preserved. For example, on CAMUS, reducing the sampling steps from 1000 to 50 decreases the inference time from 6.82 to 0.49 s per image, while FID only increases from 112.58 to 120.39 and remains clearly better than echoGAN [5]. A similar trend is observed on EchoNet-Dynamic.

Table 8: Inference efficiency and quality–speed trade-off under the Cut 60 setting. Runtime is measured as the average inference time per image on a single NVIDIA RTX 3090 GPU.

Dataset	Method	Infer. Time ↓	FID ↓	LPIPS ↓	PSNR ↑	SSIM ↓
CAMUS	echoGAN [5]	0.021 s	145.74	0.130	19.77	0.47
	Ours-1000	6.82 s	112.58	0.097	20.96	0.40
	Ours-100	0.86 s	116.24	0.100	20.71	0.41
	Ours-50	0.49 s	120.39	0.105	20.42	0.42
	Ours-25	0.30 s	129.72	0.116	20.01	0.44
EchoNet-Dynamic	echoGAN [5]	0.020 s	89.21	0.160	20.13	0.48
	Ours-1000	6.76 s	71.48	0.148	20.76	0.31
	Ours-100	0.84 s	73.62	0.151	20.54	0.33
	Ours-50	0.48 s	76.95	0.155	20.27	0.36
	Ours-25	0.29 s	82.38	0.162	19.89	0.41

These results indicate that the proposed method provides a flexible quality–speed trade-off. The full-step version is suitable for offline analysis or quality-prioritized clinical review, where structural fidelity is more important than inference speed. In contrast, reduced-step sampling can be used when faster inference is required. Nevertheless, we acknowledge that diffusion-based FoV outpainting is still slower than one-step GAN inference. Therefore, real-time clinical deployment will require further acceleration, such as more efficient samplers, diffusion distillation, and latent-space acceleration. This will be an important focus of our future work.

4.6.6 Sensitivity Analysis on Noise Schedule

The noise schedule determines how the clean image is gradually corrupted during the forward diffusion process and may affect the denoising difficulty at different timesteps. In the main experiments, we adopt the standard linear schedule following DDPM [7]. To examine whether the proposed method is sensitive to this choice, we further compare five different schedules under the representative Cut 60 setting, including linear [7], cosine [45], quadratic [46], sigmoid [46], and exponential [46] schedules. The linear schedule [7] increases the noise level approximately uniformly across timesteps, while the cosine schedule [45] follows the improved DDPM formulation and controls the cumulative signal level using a cosine-shaped decay. The quadratic, sigmoid, and exponential schedules [46] are included as representative non-linear alternatives with different noise-growth patterns.

As shown in Table 9, the choice of noise schedule leads to minor variations in performance. On CAMUS, the linear schedule achieves the best results in three out of four metrics, including FID, LPIPS, and PSNR, while the quadratic schedule obtains a slightly better SSIM. On EchoNet-Dynamic, the cosine schedule achieves the best FID and the sigmoid schedule achieves the best PSNR, whereas the linear schedule obtains the best LPIPS and SSIM and remains the second-best choice in both FID and PSNR. These results suggest that different schedules may slightly favor different metrics, but the linear schedule provides the most balanced overall performance across datasets and evaluation criteria.

Table 9: Sensitivity analysis on different noise schedules under the Cut 60 setting. The linear schedule is used in the main experiments. Best and second-best results are marked in bold and underlined, respectively.

Dataset	Noise Schedule	FID ↓	LPIPS ↓	PSNR ↑	SSIM ↓
CAMUS	Linear [7]	**112.58**	**0.097**	**20.96**	<u>0.40</u>
	Cosine [45]	<u>112.61</u>	<u>0.099</u>	<u>20.84</u>	0.41
	Quadratic [46]	114.21	0.100	20.76	**0.39**
	Sigmoid [46]	116.35	0.104	20.55	0.42
	Exponential [46]	118.67	0.107	20.38	0.41
EchoNet-Dynamic	Linear [7]	<u>71.48</u>	**0.148**	<u>20.76</u>	**0.31**
	Cosine [45]	**70.96**	<u>0.150</u>	20.68	<u>0.32</u>
	Quadratic [46]	72.18	0.152	20.61	0.36
	Sigmoid [46]	73.45	0.155	**20.78**	0.35
	Exponential [46]	75.02	0.154	20.60	0.37

Based on these results, we use the linear schedule in the main experiments because it ranks either first or second for all evaluated metrics on both CAMUS and EchoNet-Dynamic. Although some alternative schedules slightly improve individual metrics, none of them consistently outperforms the linear schedule across datasets and evaluation criteria. More importantly, under all tested schedules, the proposed framework

remains clearly better than previous methods. For example, even with the exponential schedule, which gives relatively weaker results among the tested schedules, our method still achieves an FID of 118.67 on CAMUS and 75.02 on EchoNet-Dynamic, which are still clearly better than echoGAN's FID values of 145.74 and 89.21, respectively. This indicates that the advantage of the proposed structure-aware diffusion framework does not strongly depend on a particular schedule choice.

5 Conclusion

In this paper, we proposed a structure-aware diffusion framework for echocardiographic field-of-view (FoV) outpainting. It consists of a diffusion baseline, a Structural Cue Encoder (SCE) for structure-sensitive conditioning, and Structure-aware Diffusion Learning (SDL) for structural denoising. Experiments on CAMUS and EchoNet-Dynamic show consistent improvements over GAN- and transformer-based methods.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Ruijia He and Yanyan Chen; methodology, Ruijia He, Yixin Hu and Duo Peng; software, Ruijia He; validation, Ruijia He, Qixin Zhang, Duo Peng and Yanyan Chen; formal analysis, Ruijia He, Yixin Hu, Jinze Liu and Qixin Zhang; investigation, Ruijia He, Yixin Hu and Duo Peng; resources, Yanyan Chen; data curation, Ruijia He, Jinze Liu and Qixin Zhang; writing—original draft preparation, Ruijia He; writing—review and editing, Ruijia He, Yixin Hu, Jinze Liu, Qixin Zhang, Duo Peng and Yanyan Chen; visualization, Ruijia He and Jinze Liu; supervision, Yanyan Chen; project administration, Yanyan Chen. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: All datasets employed in this work are publicly available.

Ethics Approval: Not applicable. This study used only publicly available, de-identified datasets and did not involve any new collection of human participant data.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Subramaniam K, Subramanian H, Knight J, Mandell D, McHugh SM. An approach to standard perioperative transthoracic echocardiography practice for anesthesiologists—perioperative transthoracic echocardiography protocols. J Cardiothorac Vasc Anesth. 2022;36(2):367–86. doi:10.1053/j.jvca.2021.08.100.
2. Ng A, Swanevelder J. Resolution in ultrasound imaging. Contin Educ Anaesth Crit Care Pain. 2011;11(5):186–92. doi:10.1093/bjaceaccp/mkr030.
3. Cheng YC, Lin CH, Lee HY, Ren J, Tulyakov S, Yang MH. InOut: diverse image outpainting via GAN inversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 19–24; New Orleans, LA, USA. p. 11431–40.
4. Li J, Chen C, Xiong Z. Contextual outpainting with object-level contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 19–24; New Orleans, LA, USA. p. 11451–60.
5. Gazda M, Gazda J, Kadoury S, Kanasz R, Drotar P. echoGAN: extending the field of view in transthoracic echocardiography through conditional GAN-based outpainting. Comput Methods Programs Biomed. 2025;269(2):108869. doi:10.1016/j.cmpb.2025.108869.
6. Mirza M, Osindero S. Conditional generative adversarial nets. arXiv:1411.1784. 2014.

7. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Adv Neural Inf Process Syst. 2020;33:6840–51. doi: 10.48550/arxiv.2006.11239.
8. Song J, Meng C, Ermon S. Denoising diffusion implicit models. arXiv:2010.02502. 2020.
9. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI, USA. p. 1125–34.
10. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision; 2017 Oct 22–29; Venice, Italy. p. 2223–32.
11. Sigillo L, Grassucci E, Comminiello D, StawGAN. Structural-aware generative adversarial networks for infrared image translation. In: Proceedings of the 2023 IEEE International Symposium on Circuits and Systems (ISCAS); 2023 May 21–25; Monterey, CA, USA. p. 1–5.
12. Amirkolaee HA, Amirkolaee HA. Medical image translation using an edge-guided generative adversarial network with global-to-local feature fusion. J Biomed Res. 2022;36(6):409. doi:10.7555/jbr.36.20220037.
13. Ma C, Rao Y, Cheng Y, Chen C, Lu J, Zhou J. Structure-preserving super resolution with gradient guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 7769–78.
14. Mendez M, Sundararaman S, Probyn L, Tyrrell PN. Approaches and limitations of machine learning for synthetic ultrasound generation: a scoping review. J Ultrasound Med. 2023;42(12):2695–706. doi:10.1002/jum.16332.
15. Escobar M, Castillo A, Romero A, Arbeláez P. UltraGAN: ultrasound enhancement through adversarial generation. In: International Workshop on Simulation and Synthesis in Medical Imaging. Berlin/Heidelberg, Germany: Springer; 2020. p. 120–30.
16. Stojanovski D, Hermida U, Lamata P, Beqiri A, Gomez A. Echo from noise: synthetic ultrasound image generation using diffusion models for real image segmentation. In: International Workshop on Advances in Simplifying Medical Ultrasound. Berlin/Heidelberg, Germany: Springer; 2023. p. 34–43.
17. Freiche B, El-Khoury A, Nasiri-Sarvi A, Hosseini MS, Garcia D, Basarab A, et al. Ultrasound image generation using latent diffusion models. In: Medical Imaging 2025: Ultrasonic Imaging and Tomography. Vol. 13412. Billingham, WA, USA: SPIE; 2025. p. 287–92.
18. Wang Q, Chen Y, Zhang N, Gu Y. Medical image inpainting with edge and structure priors. Measurement. 2021;185(10):110027. doi:10.1016/j.measurement.2021.110027.
19. Zheng C, Cham TJ, Cai J, Phung D. Bridging global context interactions for high-fidelity image completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 19–24; New Orleans, LA, USA. p. 11512–22.
20. Tan F, Addala AV, Nunes BAA, Zhu X, Soni R. POWDR: pathology-preserving outpainting with wavelet diffusion for 3D MRI. arXiv:2601.09044. 2026.
21. Liu Z, Pan Y, Xia Y. Structure-preserve expansion for medical image registration with minimal overlap. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention; 2025 Sep 23–27; Daejeon, Republic of Korea. Berlin/Heidelberg, Germany: Springer; 2025. p. 542–51.
22. Armanious K, Kumar V, Abdulatif S, Hepp T, Gatidis S, Yang B. ipA-MedGAN: inpainting of arbitrary regions in medical imaging. In: Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP); 2020 Oct 25–28; Abu Dhabi, United Arab Emirates. p. 3005–9.
23. Xu K, Li T, Khan MS, Gao R, Antic SL, Huo Y, et al. Body composition assessment with limited field-of-view computed tomography: a semantic image extension perspective. Med Image Anal. 2023;88(1):102852. doi:10.1016/j.media.2023.102852.
24. Zhao S, Cui J, Sheng Y, Dong Y, Liang X, Chang EI, et al. Large scale image completion via co-modulated generative adversarial networks. arXiv:2103.10428. 2021.
25. Saharia C, Ho J, Chan W, Salimans T, Fleet DJ, Norouzi M. Image super-resolution via iterative refinement. IEEE Trans Pattern Anal Mach Intell. 2022;45(4):4713–26. doi:10.1109/tpami.2022.3204461.

26. Kawar B, Elad M, Ermon S, Song J. Denoising diffusion restoration models. Adv Neural Inf Process Syst. 2022;35:23593–606. doi:10.52202/068431-1714.
27. Wang J, Yue Z, Zhou S, Chan KC, Loy CC. Exploiting diffusion prior for real-world image super-resolution. Int J Comput Vis. 2024;132(12):5929–49. doi:10.1007/s11263-024-02168-7.
28. Shi Y, Zhang X, Jia Y, Zhao J. Multi-scale adversarial diffusion network for image super-resolution. Sci Rep. 2025;15(1):11690. doi:10.1038/s41598-025-96185-2.
29. Wang J, Wang S, Zhang Y. Deep learning on medical image analysis. CAAI Trans Intell Technol. 2025;10(1):1–35. doi:10.1049/cit2.12356.
30. Wang L, Zhang S, Xu N, He Q, Zhu Y, Chang Z, et al. Role of artificial intelligence in medical image analysis. Chin Med J. 2025;138(22):2879–94. doi:10.1097/cm9.0000000000003824.
31. Ding X, Guo Y, Ding G, Han J. ACNet: strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 1911–20.
32. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. arXiv:1511.07122. 2015.
33. Sobel I, Feldman G. A 3 × 3 isotropic gradient operator for image processing. Talk Stanf Artif Proj. 1968;1968:271–2.
34. Vilimek D, Kubicek J, Golian M, Jaros R, Kahankova R, Hanzlikova P, et al. Comparative analysis of wavelet transform filtering systems for noise reduction in ultrasound images. PLoS One. 2022;17(7):e0270745. doi:10.1371/journal.pone.0270745.
35. Wang L, Yang Z, Pu YF, Yin H, Ren X. An efficient multi-scale wavelet approach for dehazing and denoising ultrasound images using fractional-order filtering. Fractal Fract. 2024;8(9):549. doi:10.3390/fractalfract8090549.
36. Song Y, Sohl-Dickstein J, Kingma DP, Kumar A, Ermon S, Poole B. Score-based generative modeling through stochastic differential equations. arXiv:2011.13456. 2020.
37. Leclerc S, Smistad E, Pedrosa J, Østvik A, Cervenansky F, Espinosa F, et al. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. IEEE Trans Med Imaging. 2019;38(9):2198–210. doi:10.1109/tmi.2019.2900516.
38. Ouyang D, He B, Ghorbani A, Yuan N, Ebinger J, Langlotz CP, et al. Video-based AI for beat-to-beat assessment of cardiac function. Nature. 2020;580(7802):252–6. doi:10.1038/s41586-020-2145-8.
39. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Adv Neural Inf Process Syst. 2017;30(1):6629–40. doi:10.18034/ajase.v8i1.9.
40. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–22; Salt Lake City, UT, USA. p. 586–95.
41. Jähne B. Digital image processing. Berlin/Heidelberg, Germany: Springer; 2005.
42. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process. 2004;13(4):600–12.
43. Zhang L, Zhang L, Mou X, Zhang D. FSIM: a feature similarity index for image quality assessment. IEEE Trans Image Process. 2011;20(8):2378–86.
44. Xue W, Zhang L, Mou X, Bovik AC. Gradient magnitude similarity deviation: a highly efficient perceptual image quality index. IEEE Trans Image Process. 2013;23(2):684–95.
45. Nichol AQ, Dhariwal P. Improved denoising diffusion probabilistic models. In: Proceedings of the 38th International Conference on Machine Learning; 2021 Jul 18–24; Virtual. p. 8162–71.
46. Martínez Guerrero E, Sun G. A comparative study of noise schedules in denoising diffusion probabilistic models. Comput Sist. 2025;29(4):1955. doi:10.13053/cys-29-4-5643.