Real-Time Novel-View Freehand Ultrasound Imaging via Point-Cloud Rendering and Diffusion-Bridge Completion

Hanrui Shi HANRUISH@UW.EDU
University of Washington, Electrical & Computer Engineering, Seattle, WA 98105, USA

Boris Mailhé BORIS.MAILHE@UII-AI.COM
Zheyuan Zhang ZHEYUAN.ZHANG01@UII-AI.COM
Yikang Liu YIKANG.LIU@UII-AI.COM
Xiao Chen XIAO.CHEN01@UII-AI.COM
Ankush Mukherjee ANKUSH.MUKHERJEE@UII-AI.COM
Terrence Chen TERRENCE.CHEN@UII-AI.COM
Shanhui Sun SHANHUI.SUN@UII-AI.COM
United Imaging Intelligence, Boston, MA 01803, USA

Editors: Under Review for MIDL 2026

Abstract

Freehand ultrasound imaging is limited by sparse sampling and restricted probe coverage, which prevent consistent visualization of unseen planes and oblique orientations. We propose a real-time framework for novel-view ultrasound imaging that combines point-cloud rendering with diffusion-bridge completion. Given a sequence of 2D B-mode images and tracked probe poses, each novel view is first rendered as a partially observed slice from the reconstructed point cloud geometry, then completed by an Image-to-Image Schrödinger Bridge (I^2SB) model to synthesize anatomically coherent textures. The diffusion-bridge formulation accelerates convergence by conditioning on visible regions instead of noise, enabling stochastic yet efficient generation. A latent I^2SB variant further improves computational efficiency for high-resolution ultrasound data. Experiments on an abdominal dataset demonstrate realistic novel-view synthesis with fine structural continuity and real-time inference (<0.2s per view), outperforming standard diffusion inpainting baselines in both speed and visual fidelity. The proposed method provides an efficient generative approach for interactive and view-adapt ultrasound visualization.

Keywords: ultrasound imaging, novel-view synthesis, diffusion bridge, point-cloud reconstruction, real-time imaging

1. Introduction

Ultrasound imaging is widely used in clinical diagnostics and interventional procedures due to its real-time capability, portability and safety. However, freehand acquisition only samples one slice at a time as shown in Figure 1, and the sonographer must navigate to the target region of interest based on that limited information. Multiple navigation assistance methods have been proposed, from providing hints to full automated robotic scanning. To be most useful, such methods must function for any views, not just good quality ones acquired during a typical clinical scan. Therefore, enabling novel-view ultrasound imaging from limited freehand inputs is an important step towards improving the performance and robustness of ultrasoudn navigation systems..

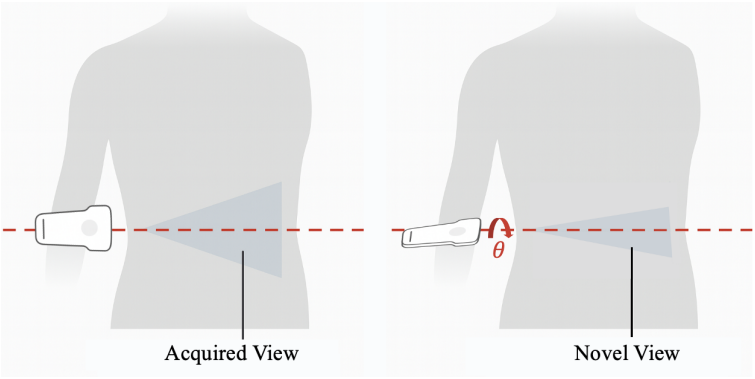


Figure 1: Illustration of an acquired view and an example novel view. The acquired view reflects the probe's true imaging plane, whereas the novel view is synthesized by rotating the probe orientation by an angle θ.

Conventional approaches for deriving unseen ultrasound views rely on interpolation or grid-based resampling, yet these methods struggle to reconstruct realistic content in regions that were never directly sampled, leading to blurred or incomplete appearances(Solberg et al., 2007; Cong et al., 2017; Prager et al., 1999). Kernel-regression techniques provide smoother estimates but remain highly dependent on dense and uniform sampling(Chen et al., 2014; Wen et al., 2013, 2018). More recently, implicit-field and learning-based representations such as NeRF and Gaussian splatting (Wysocki et al., 2024; Dou et al., 2025; Yeung et al., 2021; Eid et al., 2025) have shown potential for novel-view synthesis by modeling ultrasound appearance continuously across viewpoints. However, these methods typically require per-volume optimization and are therefore impractical for real-time deployment. Moreover, because they rely on consistency with acquired rays, they cannot reliably infer structures in regions that were not acoustically sampled. Generative diffusion models offer stronger priors for synthesizing realistic textures(Ho et al., 2020; Lugmayr et al., 2022), but their iterative sampling remains computationally expensive. Diffusion-bridge methods(Chen et al., 2021; Liu et al., 2023) provide a more efficient alternative by conditioning on partially observed inputs, yet have not been explored for real-time ultrasound novel-view synthesis.

To address these limitations, we propose a real-time novel-view ultrasound imaging framework that integrates point-cloud rendering with diffusion-bridge completion. Acquired B-mode pixels are first projected into 3D space to form a dense point cloud, from which novel viewpoints are rendered as partially observed slices using forward splatting. An Image-to-Image Schrödinger Bridge (I^2SB) model is then trained with the fully acquired ultrasound slices to stochastically complete these slices, generating anatomically consistent textures beyond the observed regions. A latent I^2SB variant further is proposed to accelerate processing for high-resolution ultrasound images. We demonstrate the effectiveness of the proposed pipeline and show that the diffusion-bridge formulation achieves superior reconstruction quality and substantially faster inference compared with standard diffusion-based completion. Together, these components enable efficient, view-adaptive ultrasound visual-

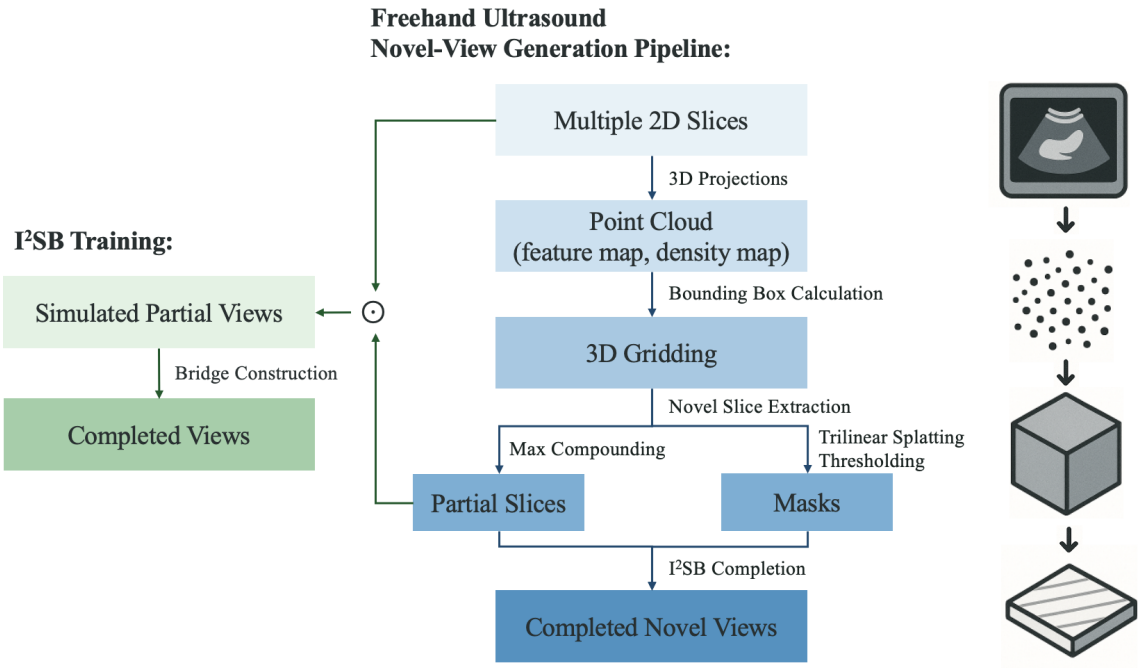


Figure 2: Overview of the Proposed Freehand Ultrasound Novel-View Generation Pipeline. Multiple freehand ultrasound slices are projected into a unified point cloud, gridded into a 3D volume, and re-sliced to produce partial novel views and their masks. These incomplete slices are refined by an I^2SB completion model, which is trained using simulated partial views derived from acquired slices and randomly sampled novel-view masks.

ization and data augmentation from sparse freehand acquisitions. Our main contributions are summarized as follows:

- A real-time synthesis pipeline for novel-view from freehand ultrasound, enabled by coupling geometry-aware point-cloud rendering with a diffusion-bridge completion mechanism that produces anatomically consistent slices at <0.2 s per view.

- A high-resolution latent diffusion-bridge model that performs completion in a compact latent space, preserving fine anatomical details while greatly reducing computation cost and enabling full-resolution real-time inference.

- A comprehensive experimental study on a large abdominal ultrasound dataset, including quantitative benchmarks, uncertainty analysis, and efficiency evaluations, demonstrating substantial gains in both fidelity and speed over diffusion-based inpainting baselines.

2. Methods

2.1. Novel View Synthesis from Point Cloud Rendering

As show in Figure 2, starting from the sequence of images acquired during an ultrasound sweep and the 6 dgree of freedom pose of the ultrasound probe at every time point, novel views are generated by extracting slices from a reconstructed 3D volume.

To reconstruct the volume, the world coordinates of every non-zero pixel of every image are first computed using a calibrated intrinsic matrix $K = \begin{bmatrix} \Delta & c \\ 0 & 1 \end{bmatrix}$ and tracked probe pose $T = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix}$. The intrinsic matrix K converts pixel indices to coordinates in a probe coordinate system, with Δ the pixel spacing of the probe and c the offset between the image corner and the attachment position of the tracker on the probe. The through-plane pixel index is 0 for the 2D ultrasound probes considered in this work. Each pixel contributes a point with spatial position $p_w = TKp$ and an associated ultrasound intensity value $f(p)$. All the points from acquired slices form the point set $\mathcal{P}$.

A common voxel grid for all images is computed as the bounding box of all pixel coordinates and a fixed voxel spacing. Every image is then resampled onto that grid. We implemented this resampling using PyTorch3D [1]: each image is first converted to a point cloud of pixels then voxelized. Voxelization also generates a density mask over the grid that counts the number of points contributing to a voxel, weighted by the trilinear resampling kernel. In order to provide robustness to shading artifacts from ultrasound acquisition, the volumes from each image are combined using max compounding, which assigns each voxel the maximum intensity among its contributing samples:

$$F(i_x, i_y, i_z) = \max_{p_i \in \mathcal{P}_{i_x, i_y, i_z}} f(p_i)$$
$$D(i_x, i_y, i_z) = \max_{p_i \in \mathcal{P}_{i_x, i_y, i_z}} d(p_i)$$

with f the voxel intensities and d the voxel densities. This preserves strong echo boundaries, naturally handles occlusions, and reduces sensitivity to probe pressure or gain variations when multiple frames contribute to the same voxel.

Novel views at arbitrary poses are generated by computing the corresponding pixel world coordinates and interpolating their intensities and densities from the 3D volume. A binary visibility mask is then constructed from the point-cloud density map, where voxels with normalized density below a threshold are labeled as unobserved and marked as missing in the rendered slice. The threshold is determined from the bimodal density histogram computed on the acquired in-plane slices, with the valley between the signal and background modes selected as the decision boundary. The resulting partially observed slice, together with its mask, is subsequently fed into the diffusion-bridge completion model.

1. https://pytorch3d.org/

2.2. I^{2}SB for Novel-View Completion

To address the incomplete spatial coverage present in novel-view slices, we adopt an I^{2}SB formulation for stochastic inpainting. The model operates directly on the partially observed slices produced by point-cloud rendering and aims to synthesize anatomically plausible completions that are consistent with both the visible pixels and the learned distribution of fully acquired ultrasound images.

Problem Formulation: Let $\mathbf{x}_0$ denote a simulated slice with missing regions encoded by a visibility mask $\mathbf{m}$, and let $\pi_1(\mathbf{x}_1)$ represent the distribution of fully sampled in-plane ultrasound slices. The completion task is framed as learning a continuous stochastic process that transports the conditional distribution $\pi_0(\mathbf{x}_0|\mathbf{m})$ to $\pi_1(\mathbf{x}_1)$. Specifically, I^{2}SB learns a generative bridge $\{p_t(\mathbf{x}_t)\}_{t=0}^1$ that interpolates between these distributions through a sequence of forward and reverse diffusions. The forward process gradually perturbs the masked slice toward a reference distribution, while the reverse process reconstructs missing structures by conditioning explicitly on the mask, thereby ensuring that synthesis remains compatible with the observed pixels.

Model Design: Following the diffusion-bridge framework, I^{2}SB differs from standard denoising diffusion models in that sampling does not begin from pure Gaussian noise. Instead, the reverse trajectory is initialized from the partially observed input, allowing the model to focus its stochastic updates on the missing regions. During training, visibility masks $\mathbf{m}$ are sampled from out-of-plane slices reconstructed by the rendering pipeline; these masks reflect realistic spatial patterns of missing data that occur during freehand scanning. At each diffusion step t, the network predicts the score function needed to reverse the diffusion process, and is optimized using the Schrödinger bridge objective:

$$\mathcal{L}_{\mathrm{SB}} = \mathbb{E}_{t,\mathbf{x}_t}\left[\|s_\theta(\mathbf{x}_t, t, \mathbf{m}) - \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{m})\|_2^2\right].$$

At inference time, the learned bridge performs stochastic sampling that progressively refines the partially observed slice. Because the model conditions on the visible regions throughout the denoising trajectory, it naturally preserves geometric features already present in the data while synthesizing plausible textures in the unobserved areas.

Latent I^{2}SB: To further accelerate inference for high-resolution slices, we introduce a latent diffusion-bridge variant. In this formulation, the input slice is first mapped into a compact latent representation using a pretrained VAE encoder, after which the I^{2}SB model performs completion entirely within this lower-dimensional space. The decoder then reconstructs the completed slice back to image space. This approach substantially reduces computational load while maintaining structural fidelity, as the latent representation captures the essential anatomical content of the ultrasound images. Operating in the latent domain also enables real-time performance for full-resolution synthesis, making the approach well suited for interactive imaging applications.

3. Experiments and Results

3.1. Experimental Setup

Dataset. Experiments were conducted on a private abdominal ultrasound dataset containing multi-sweep freehand scans from 175 subjects, each acquired under routine clinical

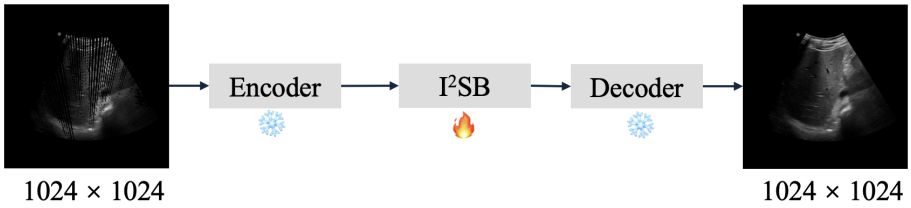


Figure 3: Framework of the latent I^{2}SB model. A pretrained VAE encoder–decoder pair is employed to process the inputs and outputs of I^{2}SB, reducing computational overhead.

protocols. For every sweep, the 6-degree-of-freedom probe poses were recorded using an electromagnetic tracking system, enabling accurate mapping between 2D pixel coordinates and 3D spatial locations. Each B-mode slice provides an in-plane resolution of 0.25×0.25,mm and a raw size of 900×1100 pixels, capturing fine-scale abdominal structures. To balance fidelity and computational cost, the vanilla I^{2}SB model operated on slices downsampled to 1×1,mm resolution, while the latent I^{2}SB variant processed full-resolution inputs through the VAE encoder. In total, approximately 400k acquired slices were included in the training set. For each slice, a synthetic partial observation mask was generated by the novel-view rendering pipeline, emulating the visibility patterns typically encountered in sparse freehand acquisition. The dataset was partitioned into training, validation, and test sets using an 8:1:1 ratio.

Implementation. Point-cloud rendering and voxelization were implemented using the PyTorch3D framework, which provides efficient GPU-accelerated routines for forward splatting, density accumulation, and trilinear interpolation. The completion model follows the I^{2}SB formulation, with a U-Net backbone used to estimate diffusion scores in both pixel and latent domains. For the latent I^{2}SB variant, we employ the widely used `sd-vae-ft-mse` encoder–decoder to map images to a compact latent space and reconstruct completed slices, which provides stable color fidelity and minimal reconstruction bias.

Model training was conducted on a single NVIDIA V100 GPU (16,GB memory) for approximately 48 hours. We used a batch size of 4 and the Adam optimizer with a learning rate of 1×10^{-4}, along with standard data normalization and random mask perturbations to improve robustness. Inference time scales linearly with the number of reverse diffusion steps, requiring roughly $0.03 \times r$ seconds per synthesized view. With as few as $r = 5$ steps, the model achieves visually stable and anatomically coherent inpainting, enabling near real-time performance suitable for interactive applications.

Evaluation. To evaluate the proposed completion framework, we compare I^{2}SB against RePaint, a representative diffusion-based inpainting baseline. For supervised in-plane completion, where paired ground-truth slices are available, we measure reconstruction fidelity using SSIM and PSNR. For novel-view completion, in which no explicit ground truth exists, we generate 4k synthetic views and assess perceptual consistency with the acquired slices using LPIPS and FID. In addition, Monte Carlo sampling with repeated stochastic inference is employed to estimate epistemic uncertainty and characterize the stability of the generated completions under different noise realizations.

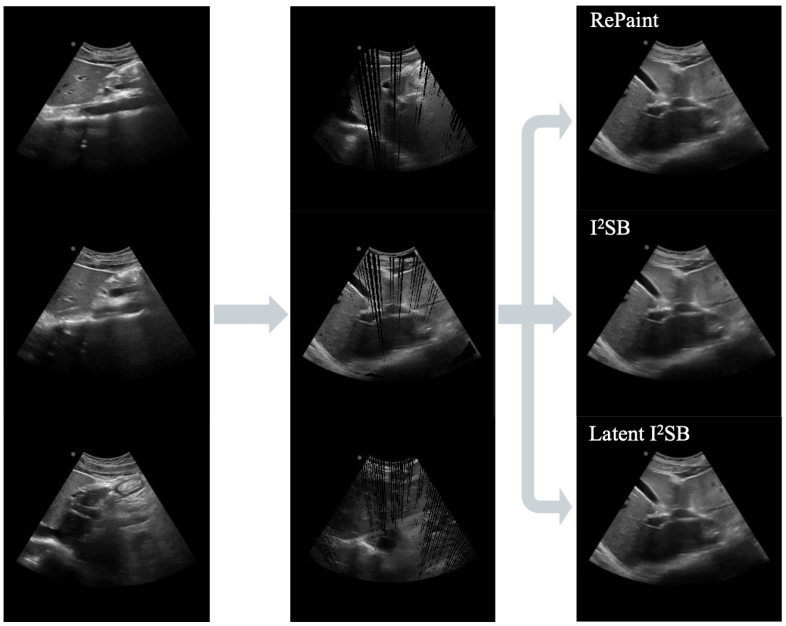


Figure 4: Novel-view ultrasound synthesis results. (a) Acquired views from sparse freehand scanning. (b) Random partial views rendered from the point-cloud representation, showing missing regions prior to completion. (c) Synthesized views produced by RePaint, I^2SB, and latent I^2SB.

3.2. Results

We present synthesized novel-view results obtained with different completion methods and analyze their performance across resolution settings, perceptual quality, and computational efficiency. RePaint and the vanilla I^2SB model are evaluated in a downsampled configuration to fit the available computational budget, whereas the latent I^2SB variant is assessed at full spatial resolution by operating directly within a compressed latent space. In addition to comparing the final synthesized images, we also examine how key hyperparameters influence the trade-off between reconstruction fidelity and runtime, which is crucial for real-time ultrasound applications.

3.3. Results

Qualitative Results. Figure 4 presents representative synthesized novel-view slices produced by the different completion models. The partially observed slices generated through point-cloud forward splatting exhibit complex and irregular missing regions, directly reflecting the heterogeneous spatial coverage typical of freehand scanning. These incomplete views therefore constitute a challenging test bed for evaluating each model's ability to infer anatomically plausible structures without direct observations.

Under a fixed computation budget of 0.15,s per view, RePaint produces globally coherent textures but tends to oversmooth the fine-scale details characteristic of abdominal

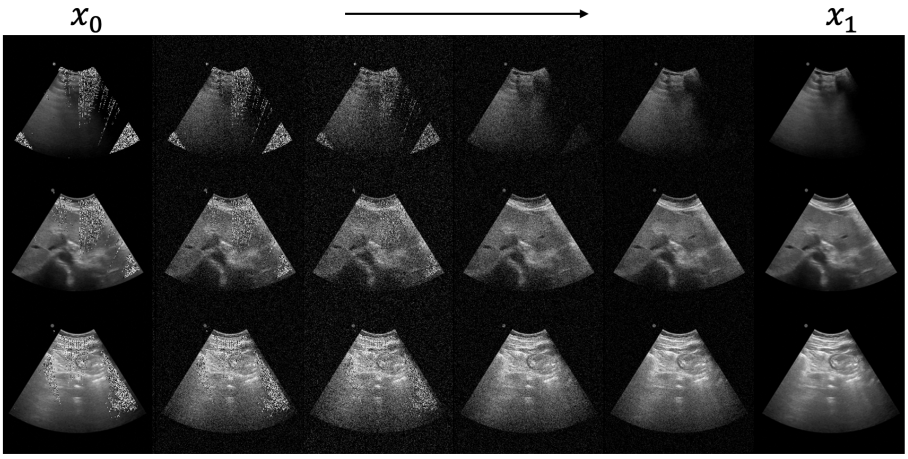


Figure 5: I^{2}SB Trajectory for Novel View Completion. The stochastic bridge learned to transition from the the masked novel view x_0 to the complete view x_1. Early states exhibit noise and uncertainty in missing regions, while later states progressively recover anatomical structure and realistic speckle patterns.

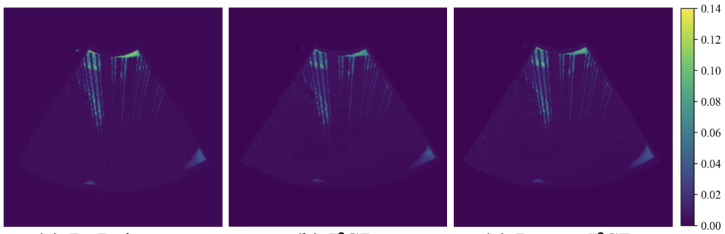
ultrasound, leading to blurred organ boundaries and diminished visibility of subtle anatomical structures. In contrast, the pixel-space I^{2}SB model reconstructs sharper interfaces and more faithfully reproduces speckle statistics, yielding completions that are visually closer to acquired slices. The latent I^{2}SB variant achieves comparably detailed reconstructions while operating in a compressed latent domain, demonstrating that high-level geometric and anatomical cues can be effectively preserved even under reduced computational overhead. These results highlight the advantage of diffusion-bridge conditioning, which provides stronger and more stable guidance from observed pixels into missing regions, thereby promoting anatomically continuous and spatially coherent reconstructions. Figure 5 further illustrates this behavior by showing the bridge trajectory, where intermediate states progressively recover structure and converge toward a realistic completed slice.

Quantitative Comparison. Table 1 reports quantitative performance across SSIM, PSNR, LPIPS, and FID. The vanilla I^{2}SB model achieves the highest fidelity across all metrics, indicating that the bridge formulation enables more accurate restoration of both global structure and fine-grained texture. RePaint shows moderate improvements over the uncompleted slices but remains substantially limited in perceptual metrics, consistent with its tendency to over-smooth textures. The latent I^{2}SB variant, despite operating at full resolution, attains competitive performance while using far fewer computational resources. This suggests that latent-space modeling is well suited for ultrasound completion tasks, where structural information dominates and high-frequency content can be effectively reconstructed through decoding.

Uncertainty Analysis. To assess stochastic stability, we perform Monte Carlo sampling with 1000 realizations for each method and compute pixelwise variance maps, shown in Figure 6. RePaint exhibits elevated variance, particularly along tissue boundaries, shad-

Table 1: Quantitative comparison of completion methods. The best performance is marked with an asterisk (*).

Method	SSIM ↑	PSNR ↑	LPIPS ↓	FID ↓
w/o Completion	0.6052	11.38	0.5578	155.80
RePaint	0.8998	29.15	0.0688	64.57
I^{2}SB	0.9798*	39.15*	0.0172*	2.113*
latent I^{2}SB	0.9456	37.25	0.0312	32.19



(a) RePaint (b) I^{2}SB (c) Latent I^{2}SB

Figure 6: Variance map from Monte Carlo sampling. I^{2}SB models exhibit higher sample consistency than RePaint.

owed regions, and occluded zones. This indicates that its completion behavior is more sensitive to sampling noise and less constrained by the observed input.

Both pixel-space and latent I^{2}SB models produce significantly lower and more localized variance, reflecting tighter adherence to the observed pixels and more stable conditional sampling trajectories. Notably, the latent I^{2}SB variant preserves the uncertainty profile of the full-resolution model despite performing denoising in a compressed space, underscoring the robustness of the bridge formulation under dimensionality reduction. This stability is especially valuable for clinical or real-time settings where consistent predictions across repeated inference runs are essential.

Step Reduction for Real Time Completion. Figure 7 illustrates the relationship between sampling steps and reconstruction quality. For I^{2}SB, SSIM increases rapidly as the number of replaced diffusion steps increases, with performance saturating around five steps, which yields an average inference time of 0.15s per image, enabling real-time completion. This plateau suggests that the bridge formulation effectively minimizes the number of required sampling iterations, making real-time deployment feasible.

By contrast, RePaint requires substantially more steps and repeated resampling to approach similar performance, leading to runtimes that exceed several seconds per slice. Although the method eventually stabilizes, its computational overhead makes it unsuitable for interactive applications. To ensure equitable comparisons across models, we adopt a matched inference-time configuration with five diffusion steps and a resampling rate of 1 for RePaint in all quantitative experiments.

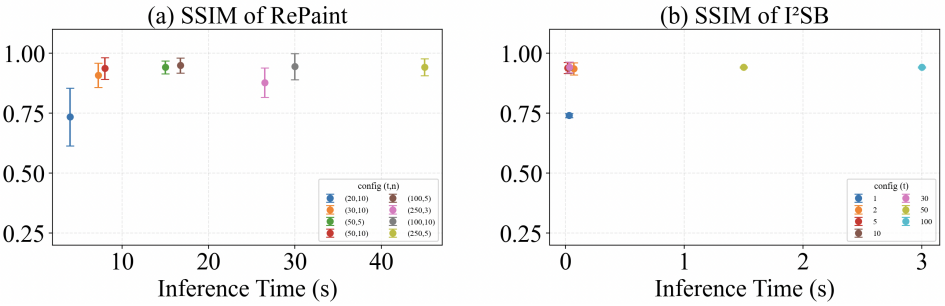


Figure 7: Comparison of SSIM versus inference time between RePaint (a) and I^{2}SB (b). The proposed I^{2}SB achieves comparable SSIM with substantially reduced inference time.

4. Discussion and Conclusion

Our results demonstrate the effectiveness of novel-view ultrasound synthesis under sparse freehand acquisitions. By combining point-cloud rendering with diffusion-bridge completion, the proposed framework preserves geometric consistency while probabilistically filling in missing local details. This hybrid formulation produces structurally coherent and anatomically aligned views, indicating that the pipeline can robustly recover high-quality novel perspectives even from limited and irregular freehand sampling.

Beyond validating the pipeline, our experiments highlight the advantages of diffusion-bridge–based completion. I^{2}SB achieves the highest quantitative performance and lower uncertainty compared with traditional diffusion inpainting, indicating that the bridge constraint produces more stable and anatomically consistent completions. Its latent-space variant extends these benefits to full-resolution inference with substantially reduced computational cost. Although the encoder–decoder structure introduces minor and unavoidable information loss, the overall completion quality remains strong. Importantly, the step-reduction analysis shows that I^{2}SB reaches performance saturation with as few as five diffusion steps enabling real-time completion, which is critical for interactive ultrasound applications such as probe guidance or rapid 3D volumetric reconstruction. Despite these strengths, there remains a limitation. Diffusion based completion may generate diverse yet equally plausible outcomes when large spatial regions are missing, reflecting inherent ambiguity rather than model error. While this behavior is expected, it suggests that completion quality depends on having at least moderate geometric support from the point-cloud rendering.

Besides, implementing 3D operations in a differentiable platform also enablesthe use of reprojection error minimization to correct potential tracking errors and improve the the consistency of acquired data. This will be the subject of future works.

In summary, the proposed pipeline achieves high-quality, real-time novel-view ultrasound synthesis by integrating point-cloud rendering with efficient diffusion-bridge completion. The framework offers a principled way to enrich training data for downstream tasks such as segmentation, registration, and quality assessment, and also supports interactive, view-adaptive ultrasound visualization.

References

Xiankang Chen, Tiexiang Wen, Xingmin Li, Wenjian Qin, Donglai Lan, Weizhou Pan, and Jia Gu. Reconstruction of freehand 3d ultrasound based on kernel regression. *Biomedical engineering online*, 13(1):124, 2014.

Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. Stochastic control liaisons: Richard sinkhorn meets gaspard monge on a schrodinger bridge. *Siam Review*, 63(2): 249–313, 2021.

Weijian Cong, Jian Yang, Danni Ai, Hong Song, Gang Chen, Xiaohui Liang, Ping Liang, and Yongtian Wang. Global patch matching (gpm) for freehand 3d ultrasound reconstruction. *Biomedical engineering online*, 16(1):124, 2017.

Yimeng Dou, Yin Li, and Tomy Varghese. Geometry aware neural radiance fields for freehand ultrasound reconstruction. *Authorea Preprints*, 2025.

Mark C Eid, Ana IL Namburete, and João F Henriques. Ultragauss: Ultrafast gaussian reconstruction of 3d ultrasound volumes. *arXiv preprint arXiv:2505.05643*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A Theodorou, Weili Nie, and Anima Anandkumar. I 2 sb: Image-to-image schr\" odinger bridge. *arXiv preprint arXiv:2302.05872*, 2023.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022.

Richard W Prager, Andrew Gee, and Laurence Berman. Stradx: real-time acquisition and visualization of freehand three-dimensional ultrasound. *Medical image analysis*, 3(2): 129–140, 1999.

Ole Vegard Solberg, Frank Lindseth, Hans Torp, Richard E Blake, and Toril A Nagelhus Hernes. Freehand 3d ultrasound reconstruction algorithms—a review. *Ultrasound in medicine & biology*, 33(7):991–1009, 2007.

Tiexiang Wen, Qingsong Zhu, Wenjian Qin, Ling Li, Fan Yang, Yaoqin Xie, and Jia Gu. An accurate and effective fmm-based approach for freehand 3d ultrasound reconstruction. *Biomedical Signal Processing and Control*, 8(6):645–656, 2013.

Tiexiang Wen, Feng Yang, Jia Gu, Shifu Chen, Lei Wang, and Yaoqin Xie. An adaptive kernel regression method for 3d ultrasound reconstruction using speckle prior and parallel gpu implementation. *Neurocomputing*, 275:208–223, 2018.

Magdalena Wysocki, Mohammad Farid Azampour, Christine Eilers, Benjamin Busam, Mehrdad Salehi, and Nassir Navab. Ultra-nerf: Neural radiance fields for ultrasound imaging. In *Medical Imaging with Deep Learning*, pages 382–401. PMLR, 2024.

Pak-Hei Yeung, Linde Hesse, Moska Aliasi, Monique Haak, Weidi Xie, Ana IL Namburete, INTERGROWTH 21st Consortium, et al. Implicitvol: Sensorless 3d ultrasound reconstruction with deep implicit representation. *arXiv preprint arXiv:2109.12108*, 2021.