Patient-Adaptive Focused Transmit Beamforming using Cognitive Ultrasound

Wessel L. van Nierop, *Member, IEEE*, Oisín Nolan, *Member, IEEE*,
Tristan S.W. Stevens, *Member, IEEE*, and Ruud J.G. van Sloun, *Member, IEEE*

Abstract— Focused transmit beamforming is the most commonly used acquisition scheme for echocardiograms, but suffers from relatively low frame rates, and in 3D, even lower volume rates. Fast imaging based on unfocused transmits has disadvantages such as motion decorrelation and limited harmonic imaging capabilities. This work introduces a patient-adaptive focused transmit scheme that has the ability to drastically reduce the number of transmits needed to produce a high-quality ultrasound image. The method relies on posterior sampling with a temporal diffusion model to perceive and reconstruct the anatomy based on partial observations, while subsequently taking an action to acquire the most informative transmits. This active perception modality outperforms random and equispaced subsampling on the 2D EchoNet-Dynamic dataset and a 3D Philips dataset, where we actively select focused elevation planes. Furthermore, we show it achieves better performance in terms of generalized contrast-to-noise ratio when compared to the same number of diverging waves transmits on three in-house echocardiograms. Additionally, we can estimate ejection fraction using only 2% of the total transmits and show that the method is robust to outlier patients. Finally, our method can be run in real-time on GPU accelerators from 2023. The code is publicly available at https://tue-bmd.github.io/ulsa/

Index Terms— Beamforming, cognitive ultrasound, diffusion models

I. INTRODUCTION

ULTRASOUND imaging is one of the most used medical imaging modalities. It brings advantages that other modalities such as magnetic resonance imaging (MRI) and computed tomography (CT) do not bring, such as, being affordable, portable, real-time and non-ionizing. These advantages make ultrasound very accessible [1].

For 2D ultrasound, we can acquire images at very high frame rates due to acquisition schemes such as diverging waves, but in more challenging circumstances, such as echocardiograms, scanners typically rely on harmonic imaging, which in turn needs a high-amplitude pressure field generated by focused transmits [2]. However, focused transmits

reduce frame-rate dramatically, which means that, especially for 3D echocardiography, it is hard to obtain high-quality and fast ultrasound scans. This shows there is a need for a reduction of transmit events while keeping the high-quality images for diagnostic accuracy.

In addition to accelerating frame rates, reducing the number of necessary transmit events also reduces certain cost factors associated with the acquisition. One such cost is power usage, which currently bottlenecks imaging modalities that depend on battery power, such as wearable ultrasound patches for continuous monitoring [3], [4]. Another cost factor is the bandwidth required to communicate the acquired data to a server for processing, which is of particular relevance to cloud-based ultrasound [5].

This work aims to reduce the number of acquisitions needed to obtain a high-quality ultrasound image by actively selecting those measurements that are expected to be most informative. This fits into the recently proposed paradigm of cognitive ultrasound, in which an autonomous agent actively designs future transmit events to maximize information-gain [6]. We drastically reduce the number of transmit events per frame and thus increase frame rate as a potential alternative to unfocused transmits, with improved tissue-harmonic generation and reduced motion decorrelation. We achieve this by equipping an imaging agent with a generative model of the ultrasound scene and observations, tracking beliefs about plausible anatomical explanations for the observations it performs. Based on these beliefs, the agent pursues acquisitions that have the highest expected information gain.

This paper presents the following main contributions. (1) We propose a method for reconstructing ultrasound video from sparse acquisitions using a temporal diffusion model that exploits the sequential nature of ultrasound. (2) We propose an active perception algorithm that designs transmits which maximizes information gain in a computationally efficient way. (3) The experimental results show that selecting focused transmits outperforms diverging waves for the same number of transmit events in terms of generalized contrast-to-noise ratio (gCNR).

II. BACKGROUND

A. Focused Ultrasound Imaging

Focused imaging is a technique used to concentrate acoustic energy at specific locations within the body. Focused line

scanning is the most widely used transmit strategy in commercial ultrasound systems, offering enhanced lateral resolution and improved image contrast relative to unfocused transmissions [7]. This strategy allows the generation of high-amplitude pressure fields, which are necessary for the generation of harmonic components used in harmonic imaging [2]. Harmonic imaging has become the gold standard for echocardiograms due to the superior image quality in hard-to-image patients [8]. However, line-by-line acquisition is time-consuming, as each lateral line requires a separate transmit event. As a result, the frame rate in this transmit mode is limited by the number of lines, imaging depth, and the speed of sound.

B. Active Perception

The goal of sensing is to acquire measurements in order to gain information about parameters describing the state of some environment of interest. Often, however, the acquisition process has some constraints – for example, a limited field of view might require that the sensor is steered in order to capture a certain aspect of the environment [9]. Such a constraint implies that the environment will only ever be *partially observed* by each acquisition. Given some prior knowledge about the parameters of the environment, however, the sensor gains the ability to infer properties of the environment without directly observing them. This process of inference on sensory states may be described as *perception*, as distinct from simple measurement [6]. We may then model this perception using the Bayesian framework, where the perceiver infers a Bayesian posterior over the parameters of the environment, with a causal model mapping those parameters to observations serving as the likelihood [10]. The aforementioned goal of sensing may then be formalized in Bayesian terms, where H is the entropy functional, x are the environmental parameters to be estimated, A is the set of sensing actions, and y are the resulting observations [11]:

$$\text{InfoGain}_{\boldsymbol{x}}(A, \boldsymbol{y}) = H[p(\boldsymbol{x})] - H[p(\boldsymbol{x} \mid A, \boldsymbol{y})]. \tag{1}$$

In other words, the information gained by performing a sensing action A is equal to the difference in uncertainty in x before versus after observing the resulting measurements y.

The perception becomes *active* when the sequence of sensing actions is optimized to maximize the expected information gain, considering all the possible measurements that may result from a given sensing action [11]:

$$\begin{aligned} \boldsymbol{a}^* &= \arg\max_A \; \mathbb{E}_{p(\boldsymbol{y}|A)}[\text{InfoGain}_{\boldsymbol{x}}(A, \boldsymbol{y})] \\ &= \arg\max_A I(\boldsymbol{x}; \boldsymbol{y} \mid A). \end{aligned} \tag{2}$$

Active perception is often performed *greedily*, and *iteratively*, first selecting the optimal sensing action according to (2), performing inference on x given the new observations y, and repeating, setting the posterior at step t to the prior at step $t+1$. This process of iteratively alternating between perception and action is referred to as a *perception-action loop*. For an extensive description of active perception in the context of ultrasound imaging, we refer the reader to [6].

C. Posterior Sampling with Diffusion Models

As mentioned in Section II-B, the ability to infer Bayesian posterior distributions given partial observations is essential to perception. Given the high-dimensional nature of ultrasound video, we employ an approximate Bayesian method, performing posterior sampling with a Diffusion model (DM). DMs are a powerful class of deep generative models capable of performing prior and posterior sampling of high-dimensional signals, such as images and videos [12]–[14]. They operate by learning to reverse a corruption process wherein a sample $\boldsymbol{x}_0 \in \mathbb{R}^N$ from the target distribution is 'diffused' towards a Gaussian noise sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. This forward corruption process is modeled as follows:

$$\boldsymbol{x}_\tau = \alpha_\tau \boldsymbol{x}_0 + \sigma_\tau \boldsymbol{\epsilon}, \tag{3}$$

where α_τ and σ_τ are called the *signal* and *noise rates* at step τ, respectively, collectively forming the *diffusion schedule*. This creates a chain of samples $[\boldsymbol{x}_0, ..., \boldsymbol{x}_\tau, ..., \boldsymbol{x}_\mathcal{T}]$ interpolating between $\boldsymbol{x}_0$ and $\boldsymbol{x}_\mathcal{T} = \boldsymbol{\epsilon}$. DMs then reverse this process iteratively, first predicting an estimate of the clean signal $\hat{\boldsymbol{x}}_0$ from some $\boldsymbol{x}_\tau$ using a denoising neural network, and then re-noising that estimate to a lower noise-level $\tau - 1$ using the forward process [15]. This process of denoising and re-noising is repeated, refining $\hat{\boldsymbol{x}}_0$ as $\tau \to 0$, and approaching a new random sample from the true data distribution $p(\boldsymbol{x})$. More formally, with an estimate of the noise $\hat{\boldsymbol{\epsilon}}$ predicted by the denoiser, $\hat{\boldsymbol{x}}_0$ can be computed by reversing the forward process as follows:

$$\hat{\boldsymbol{x}}_0 = \frac{1}{\alpha_\tau}(\boldsymbol{x}_\tau - \sigma_\tau \hat{\boldsymbol{\epsilon}}). \tag{4}$$

Tweedie's formula [16] relates this quantity to the *score function* of the marginal probability distribution over noisy samples $p_\tau(\boldsymbol{x}_\tau)$, indicating that denoising is equivalent to taking a gradient step towards a region of higher probability density in the target distribution, in the case where $\hat{\boldsymbol{\epsilon}}$ is produced by the minimum mean squared error denoiser:

$$\hat{\boldsymbol{x}}_0 \approx \mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_\tau] = \frac{1}{\alpha_\tau}(\boldsymbol{x}_\tau + \sigma_\tau^2 \nabla_{\boldsymbol{x}_\tau} \log p_\tau(\boldsymbol{x}_\tau)). \tag{5}$$

This notion of taking a step towards a region of higher prior probability density is referred to as the *prior step*. Of particular interest in this application is Bayesian posterior sampling, wherein the model generates high-quality samples conditioned on measurements $\boldsymbol{y} \in \mathbb{R}^M$ obtained according to some known measurement model $p(\boldsymbol{y} \mid \boldsymbol{x})$. The Diffusion posterior sampling (DPS) algorithm [17] solves this problem by formulating a posterior score function:

$$\underbrace{\nabla_{\boldsymbol{x}_\tau} \log p_\tau(\boldsymbol{x}_\tau | \boldsymbol{y})}_{\text{posterior}} = \underbrace{\nabla_{\boldsymbol{x}_\tau} \log p_\tau(\boldsymbol{x})}_{\text{prior}} + \underbrace{\nabla_{\boldsymbol{x}_\tau} \log p_\tau(\boldsymbol{y} | \boldsymbol{x}_\tau)}_{\text{likelihood}}. \tag{6}$$

The likelihood term in (6) is derived from a known measurement model, typically with some additive noise, e.g. $p(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}; \mathcal{A}(\boldsymbol{x}), \sigma_{\boldsymbol{n}}^2 \boldsymbol{I})$, where $\mathcal{A}$ is some measurement operator. DPS then approximates the likelihood score at step τ using the Tweedie estimate $\hat{\boldsymbol{x}}_0$ computed during the prior

step. With Gaussian measurement noise, this becomes:

$$\nabla_{\boldsymbol{x}_\tau} \log p_\tau(\boldsymbol{y} \mid \boldsymbol{x}_\tau) \simeq -\frac{1}{\sigma_{\boldsymbol{n}}^2 I} \nabla_{\boldsymbol{x}_\tau} ||\boldsymbol{y} - \mathcal{A}(\hat{\boldsymbol{x}}_0)||_2^2. \quad (7)$$

Adding the gradient in equation (7) to $\boldsymbol{x}_\tau$ constitutes the *likelihood step*. DPS alternates between prior and likelihood steps during inference, leading to samples that accord with the measurements while remaining plausible under the prior.

III. RELATED WORK

Subsampling methods have long been employed in medical imaging to decrease costs associated with acquisition. These methods typically consist of two important parts: the *subsampling mask*, choosing which part of the signal to sample, and the *reconstruction method*, recovering the target signal from the subsampled signal. Many approaches to implementing each part have been proposed in the literature. In general, the subsampling mask may be random or data-driven; it may also be fixed across samples or sample-adaptive. Similarly, the reconstruction model may be learned from data using machine learning, or hand-crafted using classical optimization techniques and simple priors. In what follows, we highlight recent work in subsampling for medical imaging, in each case categorising the approach according to the taxonomy above, and relating it to our proposed method.

In ultrasound imaging, a number of methods for subsampling channel data have been proposed, with the aim of decreasing data volume and increasing frame rates. Compressed sensing was initially employed to this end [18], [19], with more recent methods relying on deep learning. A popular deep-learning-based approach has been to employ fixed subsampling masks designed using domain knowledge, e.g., sparse array designs [20], and Convolutional Neural Network (CNN) reconstruction models to map the subsampled data to fully-sampled data [21]–[23]. The approach by Huijben *et al.* [24] instead learns subsampling masks jointly with a CNN reconstruction model, employing the Gumbel-Max trick [25] to backpropagate through the subsampling operation. Afrakteh *et al.* [26] tackle the problem of focused scan-line subsampling, using tensor-completion methods to inpaint the data-cubes containing the subsampled frames. We tackle the same problem in this paper, but use a data-driven prior in the form of a diffusion model with an adaptive subsampling mask, as opposed to the nuclear-norm tensor regularization and random subsampling mask used by Afrakteh *et al.*

A wide range of subsampling methods has been proposed for MRI acceleration, spurred in part by high-quality open-access datasets such as fastMRI [27]. The most successful of these methods use deep learning, typically with CNN-based architectures for reconstruction. Initial approaches opted for fixed masks, some hand-crafted [28] and some learned from data [29], [30]. Some more recent methods instead actively design the subsampling mask, leading to input-specific masks and improved reconstruction accuracy [31]–[33]. Of particular relevance to this work is dynamic MRI, which more closely resembles ultrasound data due to the presence of temporal correlation. A recent work by Yiasemis *et al.*[34] leverages this temporal correlation by creating an active subsampling

model for dynamic MRI, training a U-Net[35] based model to iteratively select which k-space lines to select per frame.

In this work, we identify the task of recovering fully-sampled ultrasound frames from a subset of scanned lines as being akin to *inpainting*, a popular task in computer vision and image generation: in both cases, the goal is to optimally recover the missing portion of the signal. We therefore choose to use diffusion models, which have shown excellent performance in inpainting [17], [36], to solve this problem. This modelling choice is further motivated by recent success in applying diffusion models to the domain of ultrasound, for synthetic data generation [37], dehazing [38], and beamforming [39].

IV. METHOD

In this section, we present our proposed method in terms of its two primary components: (i) *perception*, in which a posterior distribution over the possible states of the tissue is inferred from a partial observation, and (ii) *action*, in which this perceived distribution is used to select the next transmit lines. An overview of the method is shown in Figure 1.

A. Perception

The goal of the perception step is to infer a posterior distribution over the tissue state $\boldsymbol{x}_t$ at time t given the history of observations and actions until that point, i.e. the distribution $p(\boldsymbol{x}_t \mid \boldsymbol{y}_{<t}, A_{<t})$, where the shorthand $< t$ indicates $1 \ldots t$. We implement this inference procedure using the DPS algorithm described in Section II-C. Given that ultrasound video exhibits strong temporal dependencies between frames, it is important to model the conditional relationship between $\boldsymbol{x}_t$ and past measurements $\boldsymbol{y}_{<t}$. In order to model such dependencies, we fit the diffusion model on sequences of W consecutive frames $\mathbf{X} = [\boldsymbol{x}_{t-W}, \ldots, \boldsymbol{x}_t]$ sampled at random from the training set, learning a prior over tensors $\mathbf{X} \in \mathbb{R}^{N \times W}$. In other words, the model has a temporal context window of size W. This amounts to a prior model with a $W-$order Markov assumption on ultrasound video, where W can be chosen to balance the benefits in predictive ability with the cost of increasing training data sparsity and inference compute as W increases. For the models presented in this work, we use $W = 3$.

During inference, at each time step t we generate a set of N_p tensors $\mathbf{X}$ in parallel. The final image $\mathbf{X}[W]$ in each tensor represents one possible state of $\boldsymbol{x}_t$. These images, dubbed *particles*, can then be used to approximate the posterior distribution $p(\boldsymbol{x}_t \mid \boldsymbol{y}_{<t})$. Throughout the paper, we refer to this set of particles $\{\boldsymbol{x}_t^{(i)}\}_{i=1}^{N_p}$ as the agent's *belief distribution* at time t, with differences across particles indicating uncertainty in the state of $\boldsymbol{x}$. Throughout our experiments, we use $N_p = 4$.

We must then specify a likelihood function to guide generation with DPS. We start by stacking our acquired scan-line measurements in a measurement buffer $\mathbf{Y} = [\boldsymbol{y}_{t-W}, \ldots, \boldsymbol{y}_t]$. Then, we define a measurement model simulating focused line-scanning. This model assumes that for each focused transmit, a single line of pixels extending along the focus line is beamformed, and that a frame is created by concatenating a string of such lines. The measurement model is thus a masking
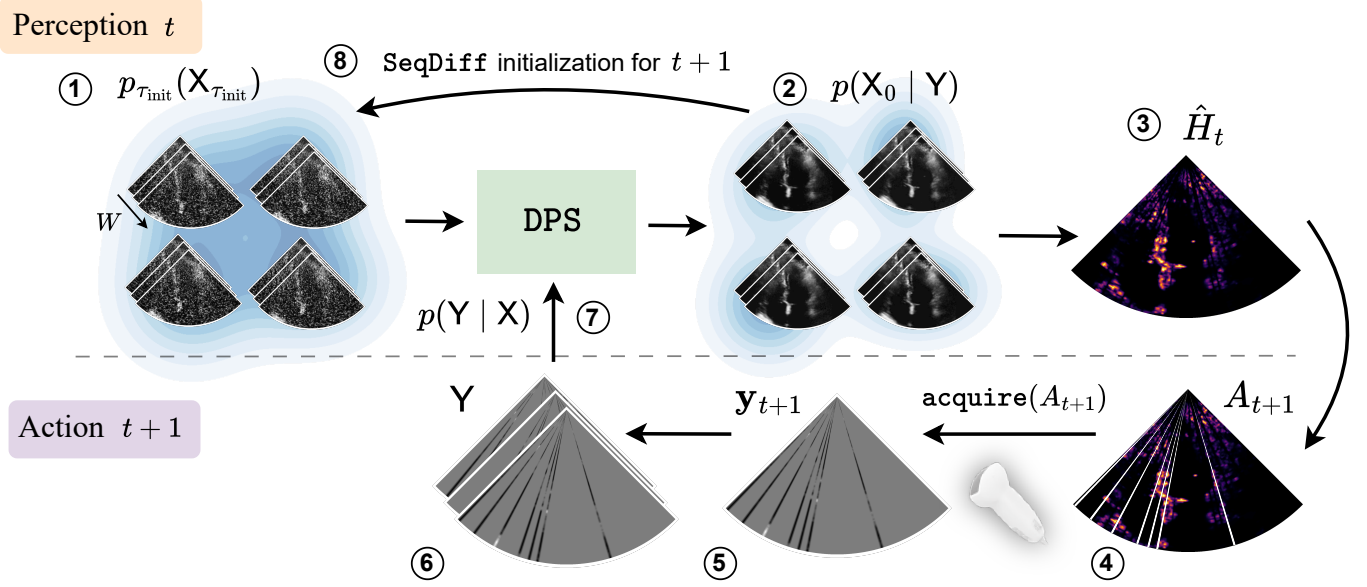
Fig. 1. ① Initialize the particles with noise at $t = 1$ or partially-noised previous samples for $t > 1$. ② Generate posterior samples using DPS. ③ Compute pixel-wise entropy from belief distribution. ④ Select next actions A_{t+1} using *K-Greedy Entropy Minimization.* ⑤ Acquire the next measurement. ⑥ Add new measurements to the measurement buffer. ⑦ Use the updated measurement buffer to run DPS at time $t + 1$. ⑧ Initialize the samples to be generated at time $t + 1$ using those generated at time t.

operation, wherein the full frame is mapped to a set of measurements by revealing only those that were acquired. In particular, $\mathbf{A} \in \mathbb{R}^{N \times W}$ is a measurement mask extending across the context window containing ones at the pixel locations measured by the acquired scan lines, and zeros elsewhere. Since this measurement model is deterministic, its likelihood is a Dirac delta distribution, i.e. $p(\mathbf{Y}|\mathbf{X}, \mathbf{A}) = \delta(\mathbf{Y} - \mathbf{A} \odot \mathbf{X})$. In order to ensure smooth gradients for DPS, however, we instead use a Gaussian distribution, which is a continuous relaxation of the Dirac delta. This yields the following likelihood, where the variance $\sigma_{\boldsymbol{n}}^2 = \gamma^{-1}$ is a hyperparameter:

$$p(\mathbf{Y} \mid \mathbf{X}, \mathbf{A}) = \mathcal{N}(\mathbf{Y}; \mathbf{A} \odot \mathbf{X}, \sigma_{\boldsymbol{n}}^2 I). \tag{8}$$

Computing the score of this likelihood function produces the following guidance step in DPS for diffusion step τ:

$$\nabla_{\mathbf{X}_\tau} \log p_\tau(\mathbf{Y} \mid \mathbf{X}_\tau) \simeq -\gamma \nabla_{\mathbf{X}_\tau} ||\mathbf{Y} - \mathbf{A} \odot \hat{\mathbf{X}}_0||_2^2. \tag{9}$$

In the case where the beamforming grid is specified in the polar domain, we fit the diffusion model on polar domain data, such that the model remains the same on polar and Cartesian grids, in each case simply revealing or masking vertical lines of pixels. In order to accelerate inference and create a temporally consistent video, we employ SeqDiff [40] initialization. Given that our DM is trained on stacks of images $\mathbf{X}$, the SeqDiff initialization becomes $\mathbf{X}_{t,\tau_{\text{init}}} \leftarrow \alpha_{\tau_{\text{init}}} \mathbf{X}_{t-1} + \sigma_{\tau_{\text{init}}} \boldsymbol{\epsilon}$. Finally, we return for each frame a single *reconstruction* image, $\tilde{\boldsymbol{x}}_t$, which is chosen to be the first particle $\tilde{\boldsymbol{x}}_t := \boldsymbol{x}_t^{(1)}$ of the belief distribution.

B. Action

The action step aims to choose a set of actions to take at time $t + 1$ given the belief distribution at time t. The action

space in this case is a discrete set of possible focused scan locations $\{A^\ell \mid \ell = 1, 2, \ldots, L\}$, where there are L total scan locations. Each action A^ℓ then denotes the set of indices of the pixels that are measured by that action, facilitating the creation of a corresponding measurement mask $\mathcal{M}(A^\ell)$, where $\mathcal{M}$ creates a matrix containing ones at the indices specified by A^ℓ and zeros elsewhere. The actions should be chosen to maximize information gain with respect to the tissue state, following the objective described in Section II-B. Starting with the expected information gain objective provided in (2), and following Van Sloun [6], we derive our action selection policy, substituting in the likelihood function specified in (8):

$$I(\boldsymbol{x}_t; \boldsymbol{y}_t \mid A^\ell, \boldsymbol{y}_{<t}) = H(\boldsymbol{y}_t \mid A^\ell, \boldsymbol{y}_{<t}) - H(\boldsymbol{y}_t \mid \boldsymbol{x}_t, A^\ell, \boldsymbol{y}_{<t})$$
$$= H(\boldsymbol{y}_t \mid A^\ell, \boldsymbol{y}_{<t}) - H(\boldsymbol{n}). \tag{10}$$

The second entropy term $H(\boldsymbol{y}_t \mid \boldsymbol{x}_t, A^\ell)$ is the entropy of our likelihood function, whose only source of uncertainty is the additive noise $\boldsymbol{n}$. $H(\boldsymbol{n})$ then drops out when we take the argmax with respect to the action A^ℓ, yielding the following objective:

$$\arg\max_\ell I(\boldsymbol{x}_t; \boldsymbol{y}_t \mid A^\ell, \boldsymbol{y}_{<t}) = \arg\max_\ell H(\boldsymbol{y}_t \mid A^\ell, \boldsymbol{y}_{<t}). \tag{11}$$

The remaining entropy values for each line measurement $H(\boldsymbol{y}_t \mid A^\ell, \boldsymbol{y}_{<t})$ can be decomposed into a sum of pixel-wise entropy values by modeling the pixels as independent variables. Given that pixels masked by A^ℓ have zero entropy, the measurement entropy can be computed as a function of pixel entropies in $\boldsymbol{x}_t$, where $\boldsymbol{x}_t[\mathrm{i}]$ denotes the ith pixel of $\boldsymbol{x}_t$:

$$H(\boldsymbol{y}_t \mid A^\ell, \boldsymbol{y}_{<t}) = \sum_{\mathrm{i} \in A^\ell} H(\boldsymbol{x}_t[\mathrm{i}] \mid A^\ell, \boldsymbol{y}_{<t}) \tag{12}$$

In practice, we first compute a pixel-wise entropy map in the data domain $\boldsymbol{x}_t$, $\hat{H} = [\hat{H}[0], \ldots, \hat{H}[i], \ldots \hat{H}[N]]^\top$, where $\hat{H}[i] = H(\boldsymbol{x}_t[i] \mid A^\ell, \boldsymbol{y}_{<t})$. Given $\hat{H}$, we can sum the pixels corresponding to each action A^ℓ in order to get the line-wise measurement entropies, choosing the maximum entropy line as the next action. Using the variational entropy approximation proposed by Hershey *et al.* [41], the pixel-wise entropy map $\hat{H}$ can be computed by taking the element-wise squared error between each pair of particles in the belief distribution $\{\boldsymbol{x}_t^{(i)}\}_{i=1}^{N_p}$, as follows:

$$\hat{H} = -\sum_i \frac{1}{N_p} \log \sum_j \frac{1}{N_p} \exp \left[-\frac{(\boldsymbol{x}_t^{(i)} - \boldsymbol{x}_t^{(j)})^2}{2\sigma_{\boldsymbol{x}}^2} \right]. \quad (13)$$

Intuitively, this entropy map will have high values in regions where the images in the belief distribution *disagree* with one another, indicating uncertainty. Selecting the maximum entropy line ℓ from this entropy map then amounts to:

$$\arg\max_\ell H(\boldsymbol{y}_t \mid A^\ell, \boldsymbol{y}_{<t}) \approx \arg\max_\ell \sum_{i \in A^\ell} \hat{H}[i]. \quad (14)$$

We could proceed with the above as our policy, selecting one line at a time, performing the perception step for the resulting measurement, and repeating. However, the perception step requires executing some reverse diffusion steps. If this perception procedure is slower than the time taken to acquire the line, then it would bottleneck the frame rate. In order to prevent this, we propose an approximate algorithm, called **K-Greedy Entropy Minimization**. K-Greedy Entropy Minimization approximates the decrease in entropy that would result from conditioning on a given measurement using a radial basis function (RBF) around the measurement location. This effectively assumes that measuring a line ℓ will provide information about nearby lines, decreasing exponentially with distance. The algorithm proceeds by selecting the maximum entropy line, reweighting the entropies of the neighboring lines according to the RBF, and repeating, for K total lines. For a formal presentation of this algorithm, see the *action* step in Algorithm 1.

V. EXPERIMENTS

A comprehensive evaluation of the model's performance is provided through a series of experiments. First, we test our method on the Echonet-Dynamic dataset, which is an image dataset from which we simulate subsampling transmits using a masking measurement model. Next, we use an in-house dataset where we can directly subsample the transmit events in the channel data, and beamform those transmits to independent lines of pixels. Lastly, we show that our method can also be applied to 3D echocardiography, where we subsample elevation planes. We implement our active perception agent using zea, the cognitive ultrasound toolbox [42].

A. EchoNet-Dynamic

Here we train a diffusion model on the EchoNet-Dynamic dataset [43]. The EchoNet-Dynamic dataset consists of over 10k echocardiograms. As we do not have access to how

---

Algorithm 1 Focused Transmit Perception-Action Loop

Require: SeqDiff initial diffusion step τ_{SeqDiff}; total diffusion steps τ_{max}; number of focused transmit locations L; number of particles N_p; number of focused transmits per frame K; initial transmit indices A_1; diffusion schedule $\{\alpha_\tau, \sigma_\tau\}_{\tau=0}^{\tau_{\text{init}}}$; guidance weight γ; posterior variance $\sigma_{\boldsymbol{x}}^2$; RBF width w; temporal window size W.

Ensure: Sequence $\{\tilde{\boldsymbol{x}}_t\}_{t=1}^T$ of reconstructed frames.

1: **for** $t \in [1, ..., T]$ **do**
2:      $\boldsymbol{y}_t \leftarrow \texttt{acquire}(A_t)$ // Acquire measurements
3:      $\mathbf{Y} \leftarrow [\boldsymbol{y}_{t-W}, \ldots, \boldsymbol{y}_t]$ // Measurement buffer
4:      $\mathbf{A} \leftarrow [\mathcal{M}(A_{t-W}), \ldots, \mathcal{M}(A_t)]$ // Mask buffer
5:      **if** $t = 1$ **then**
6:          $\tau_{\text{init}} = \tau_{\text{max}}$
7:      **else**
8:          $\tau_{\text{init}} = \tau_{\texttt{SeqDiff}}$

     Perception Step

9:      **for** each $i \in \{1, ..., N_p\}$ in parallel **do**
10:          $\mathbf{X} \leftarrow [\boldsymbol{x}_{t-W-1}^{(i)}, \ldots, \boldsymbol{x}_{t-1}^{(i)}]$
11:          $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ // Initial noise
12:          $\mathbf{X}_{\tau_{\text{init}}} \leftarrow \alpha_{\tau_{\text{init}}} \mathbf{X} + \sigma_{\tau_{\text{init}}} \boldsymbol{\epsilon}$ // Initial samples
13:          **for** $\tau \in [\tau_{\text{init}}, ..., 0)$ **do**
14:              $\hat{\boldsymbol{\epsilon}} \leftarrow \boldsymbol{\epsilon}_\theta(\mathbf{X}_\tau, \sigma_\tau^2)$ // Predict Noise
15:              $\hat{\mathbf{X}}_0 \leftarrow (\mathbf{X}_\tau - \sigma_\tau \hat{\boldsymbol{\epsilon}})/\alpha_\tau$ // Tweedie Estimate
16:              $\mathbf{X}'_{\tau-1} \leftarrow \alpha_{\tau-1}\hat{\mathbf{X}}_0 + \sigma_{\tau-1}\hat{\boldsymbol{\epsilon}}$ // Prior step
17:              $\mathbf{X}_{\tau-1} \leftarrow \mathbf{X}'_{\tau-1} - \gamma \nabla_{\mathbf{X}_\tau} ||\mathbf{Y} - \mathbf{A} \odot \hat{\mathbf{X}}_0||_2^2$
18:          $\boldsymbol{x}_t^{(i)} \leftarrow \mathbf{X}_0[W]$ // Belief distribution
19:      $\tilde{\boldsymbol{x}}_t \leftarrow \boldsymbol{x}_t^{(1)}$ // Choose first as reconstruction

     Action Step

20:      $A_{t+1} \leftarrow \emptyset$ // Initialize action set for next transmit
21:      $\hat{H} \leftarrow -\sum_i \frac{1}{N_p} \log \sum_j \frac{1}{N_p} \exp \left[-\frac{(\boldsymbol{x}_t^{(i)} - \boldsymbol{x}_t^{(j)})^2}{2\sigma_{\boldsymbol{x}}^2} \right]$
22:      $\hat{H}^\ell \leftarrow \sum_{i \in A^\ell} \hat{H}[i]$ // Line-wise entropy
23:      **for** $k \in [1, \ldots, K]$ **do**
24:          $\ell^* \leftarrow \arg\max_\ell \hat{H}^\ell$ // Select max entropy action
25:          $A_{t+1} \leftarrow A_{t+1} \cup A^{\ell^*}$ // Gather selected actions
26:          $\hat{H}^\ell \leftarrow \hat{H}^\ell * -\exp\left(-\frac{(\ell-\ell^*)^2}{w}\right)$ // Reweight

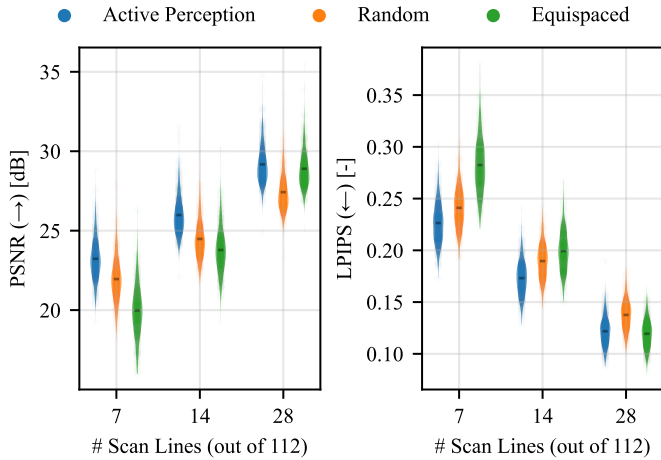   **return** $\{\tilde{\boldsymbol{x}}_t\}_{t=1}^T$

Fig. 2. Reconstruction performance for EchoNet-Dynamic in terms of PSNR and LPIPS as a function of the number of scanned lines for various action selection policies. The figure shows a distribution over the data samples and includes the mean as a gray line.
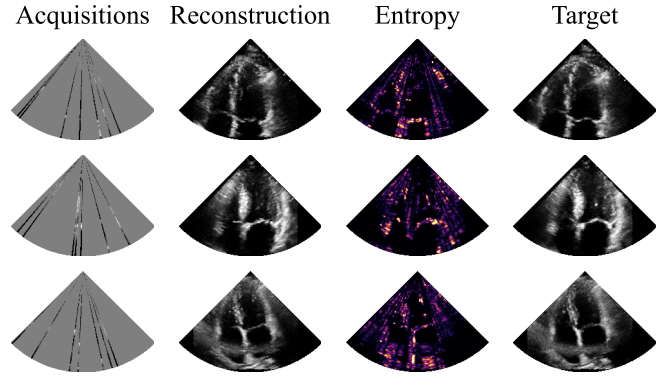

Fig. 3. Qualitative results on the EchoNet-Dynamic dataset. The figure shows the acquisitions and reconstructions for 7 / 112 lines compared to the target. Additionally shows the posterior entropy, which drives action selection.
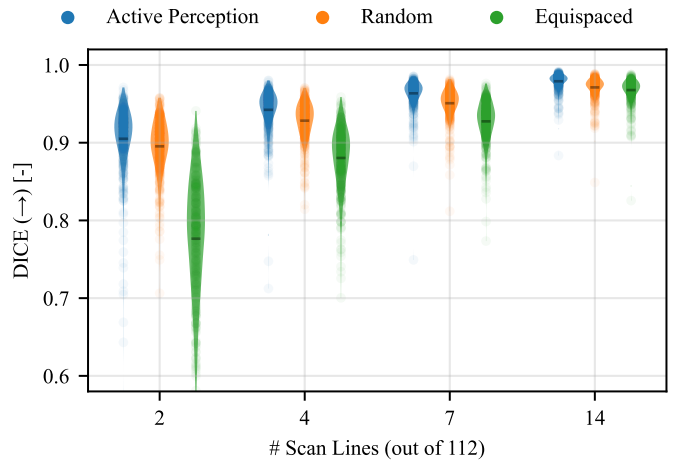

Fig. 4. Segmentation performance in terms of DICE of EchoNet-Dynamic on subsampled images for various action selection policies. The figure shows a distribution over the data samples and includes the mean as a gray line.

the data was beamformed or the channel data, we opted to simulate scan-lines as a column of pixels of the 112×112 images. To that extent, we have converted the dataset from scan-converted images back to the polar domain. In the process, we excluded 2,044 samples because their scan-converted images were generated using a different method or parameters, which prevented consistent conversion to the polar format used for the rest of the dataset. The rest of the data we have randomly split on the patient level into 6985 train sequences, 500 validation sequences, and 500 test sequences. While we used the full sequences to train our model, we use 100 frames per patient for the metrics to ensure every patient gets weighted equally in the metrics.

The active perception agent will be compared to equispaced and random subsampling, using the same diffusion model. The equispaced subsampler 'rolls' the selected lines from left to right, such that over time the full imaging area is measured. Random sampling means that the selected lines were sampled from a uniform distribution.

1) Reconstruction quality: Figure 2 shows the reconstruction quality in terms of peak signal-to-noise ratio (PSNR) and learned perceptual image patch similarity (LPIPS) as distributions over all the patients in the test dataset. It can be seen that active perception subsampling outperforms the other subsampling strategies, especially for lower subsampling rates. For 7 out of 112 lines, which is just over 6% of the image, the agent still achieves a PSNR of 23.23 on average, which consists of a 5.8% improvement over random sampling and an impressive 16.3% improvement over equispaced sampling.

The qualitative results are shown in Figure 3. Here, the 20th frame in four random sequences is used for three random patients in the test data. We show the acquired lines, the reconstruction, the entropy of the posterior samples, and the fully observed target images. The reconstructions are visually very similar to the targets, while using only 7 out of 112 scan-lines.

2) Left ventricle segmentation: A common parameter extracted from an echocardiogram is the ejection fraction, which measures the amount of blood pumped out of the heart's left ventricle with each heartbeat. The EchoNet-Dynamic model [43] can segment the left ventricle with high accuracy. In this experiment, we will evaluate how the subsampled reconstructions affect the ability to segment the left ventricle. We will use DICE to compare the segmentations of the subsampled images and the fully observed images. We exclude failure cases from the fully observed image sequences in which the segmentation model generates multiple disconnected components in at least five consecutive frames. Figure 4 shows that the active perception agent consistently produces the best left ventricle segmentations compared to equispaced and random subsampling. The performance for 2 out of 112 still reaches a DICE of 0.91 on average.

3) Robustness across patients: An essential feature of any image reconstruction method in medical imaging is robustness against outliers, ensuring that the performance is consistent across patients. In order to evaluate this in our approach, we
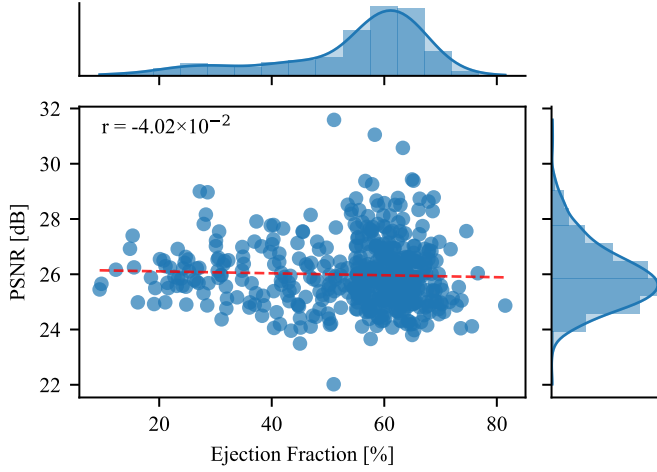
Fig. 5. Reconstruction quality (PSNR) plotted against patient ejection fraction. The lack of correlation indicates that reconstruction performance is consistent across varying ejection fractions, suggesting no bias against outlier patients.
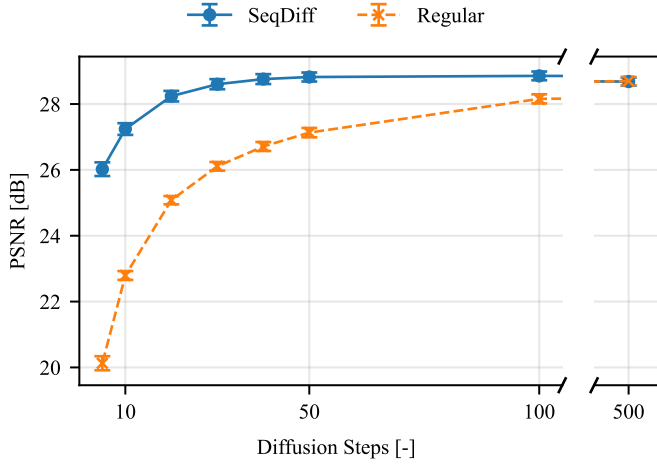

Fig. 6. Reconstruction quality for SeqDiff [40] and regular diffusion as a function of the diffusion steps, i.e., the acceleration. The reconstruction quality was computed for a single sequence with active perception and a subsampling rate of 25%. The error bars show the standard error over the frames.

TABLE I
INFERENCE SPEED OPTIMIZATIONS COMPUTED ON THE RTX 2080 TI GPU (NVIDIA, SANTA CLARA, CA, USA) FOR 112×112 PIXELS.

Optimization	Frame	
	Time [ms]	Freq. [Hz]
Unoptimized (*500 steps*)	3868	0.26
+ SeqDiff [40] (*25 steps*)	365.2	2.74
+ Just-in-time compilation	151.6	6.6
+ Parallel posterior sampling (*multi-gpu*)	80.56	12.41
+ Mixed precision (*float16*)	61.67	16.22
All (*on 1x Nvidia H100*)	24.02	**41.64**
Frame acquisition (28 lines)	5.46	183.2

further, we applied a group of optimizations as shown in Table I. First, we chose 25 SeqDiff steps as a good balance between reconstruction quality and inference speed. Then we applied just-in-time compilation using the JAX library [44]. Furthermore, we parallelized the computation of the posterior samples across multiple GPUs when needed. Finally, the diffusion model, trained in 32-bit floating point precision, can be run in mixed precision using 16 bits. When these optimizations are applied on a single H100 GPU (NVIDIA, Santa Clara, CA, USA) from 2023, the active perception agent can be run with over 41 Hz.

B. In-house echocardiograms

The in-house dataset consists of 90 focused transmits, which were interleaved with 11 diverging transmits for comparison. All subjects in this dataset provided informed consent at the time of data collection, and the study was approved by the local Institutional Review Board. We apply active perception by subsampling certain transmit events from the (fundamental) channel data and independently beamforming only those transmit events to columns of pixels, giving us y_t. The pretrained prior will be used to generate reconstructions $\hat{x}_t$.

To demonstrate the effectiveness of our method, we compute the gCNR metric between the ventricle and the myocardium as well as between the ventricle and the valve. The gCNR is calculated relative to the fully sampled focused acquisition, which allows us to compare active perception to diverging waves for the same number of transmits.

Figure 7 shows the gCNR over time between the valve and the ventricle for two subjects. It can be seen that active perception almost always outperforms diverging waves. Active perception generally has slightly higher gCNR compared to focused transmits, while for diverging waves it is slightly lower. In Figure 8 we show the distribution of gCNR over the frames between the myocardium and ventricle for three subjects. This highlights again that active perception outperforms diverging waves for all subjects, and shows fewer outliers.

The qualitative results are shown in Figure 9. Here, we see the fully sampled focused and diverging waves scans, combined with the acquired focused lines (11 out of 90)

ran active perception on the first 100 frames of each of the 500 sequences in the unseen EchoNet-Dynamic test set, with a measurement budget of 14 lines per frame. In Figure 5, we plot the reconstruction quality as measured by PSNR against the ejection fraction of each patient, examining the correlation between the two. Figure 5 shows that the reconstruction quality is independent of the patient's ejection fraction, indicating a lack of bias against outlier patients.

4) Inference speed: As mentioned before, we employ SeqDiff [40], which not only improves temporal consistency of posterior samples, it also massively reduces the required number of function evaluations for sequential signals. Figure 6 shows the relation of the number of diffusion steps to the reconstruction quality in terms of PSNR for regular and SeqDiff, which motivates employing SeqDiff for enhanced reconstruction quality and speed. To improve inference speed
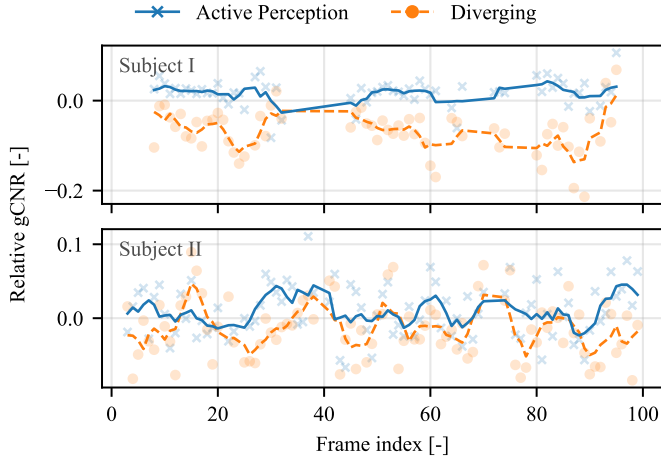
Fig. 7. Generalized contrast-to-noise ratio (gCNR) for two subjects over time relative to a focused acquisition of 90 transmits. The gCNR was measured between the **valve** and the **ventricle**. Both active perception and diverging use 11 transmits.
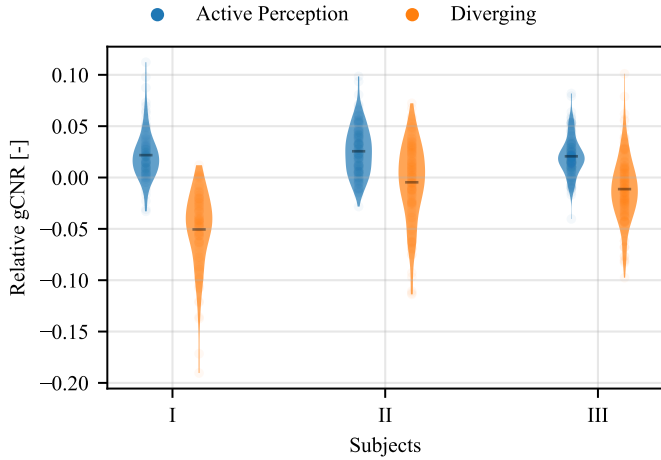

Fig. 8. Generalized contrast-to-noise ratio (gCNR) for three subjects relative to a focused acquisition of 90 transmits. The gCNR was measured between the **myocardium** and the **ventricle**. Both active perception and diverging use 11 transmits. The figure shows a distribution over the frames and includes the mean as gray line.

and the reconstruction using our method. Even though the diffusion model was trained on a different dataset, the method still reconstructs well using limited measurements. For the same number of transmits as diverging waves, it shows certain details, such as the valve, more clearly.

C. 3D echocardiograms

In this section, we apply active perception to 3D echocardiography. Following Stevens *et al.* [45], we consider a measurement model in which the elevation dimension is sparsely sampled, leading to a small set of acquired focused elevation planes from which the full volume must be recovered. Building on the reconstruction model implemented by Stevens *et al.*, we too train a DM on 2D slices taken along the axial (ax) and elevation (el) axes, but we extend this model in the temporal direction as with our EchoNet model described in Section IV.
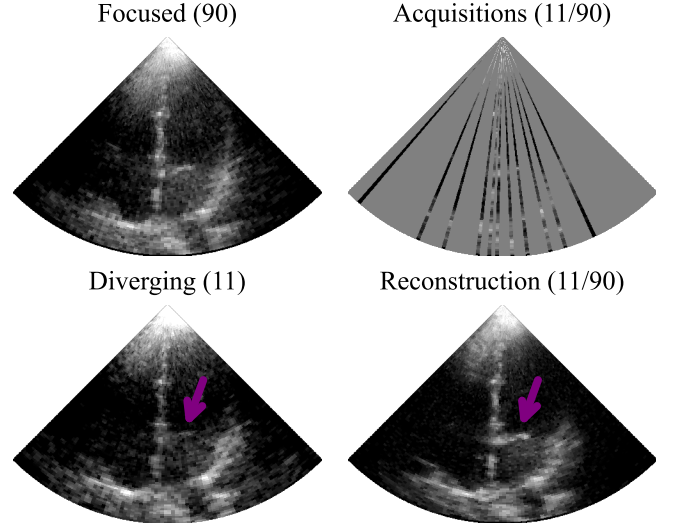

Fig. 9. Qualitative results on the in-house echocardiograms. On the left, the figure shows a focused acquisition that was interleaved with a diverging wave acquisition. On the right, the acquisitions and reconstructions are shown for 11/90 focused transmits. All images were 112×112 pixels prior to scan conversion.

Our prior is therefore approximating the joint distribution $p(\mathbf{X})$ where $\mathbf{X} \in \mathbb{R}^{N_{ax} \times N_{el} \times W}$. The DM was trained on samples of size $N_{ax} = 400$, $N_{el} = 48$, and $W = 3$. The dataset consists of 100 *in-vivo* volume sequences across 16 patients, acquired using a Philips EPIQ scanner with an X5-1C matrix probe. A set of 7 volume sequences across 3 patients is held out for testing. For posterior sampling, a guidance weight of $\gamma = 3$ was used, with $N_p = 4$, $\tau_{\max} = 500$, and $\tau_{\texttt{SeqDiff}} = 450$, and initial planes A_1 selected uniformly at random.

In order to perform the action step on 3D volumes, the K-Greedy Entropy Minimization algorithm was modified to first average the entropy map across azimuthal angles to produce a 2D entropy map along the axial and elevation axes. Given this 2D entropy map, the algorithm proceeds as in the 2D case, selecting a series of lines, now representing elevation planes, aiming to cover as much entropy as possible.

As with the experiments on EchoNet-Dynamic, we benchmarked reconstructions created with active perception against those created with baseline sampling strategies, with PSNR and LPIPS results plotted in Figure 10. Across the subsampling rates, it is clear that employing active perception results in more faithful reconstructions, particularly with more aggressive subsampling. The distributions of results are also more unimodal when using active perception, indicating a more stable performance across patients and volumes. In Figure 11, we provide qualitative examples in the form of bi-plane plots of volume reconstructions from 6/48 elevation planes, at the 4th frame in each sequence.

VI. DISCUSSION

It is clear throughout the results provided in Section V that the active perception strategy outperforms the equispaced and random baseline strategies. The degree of improvement varies across the experiments. In Section V-A, our results on
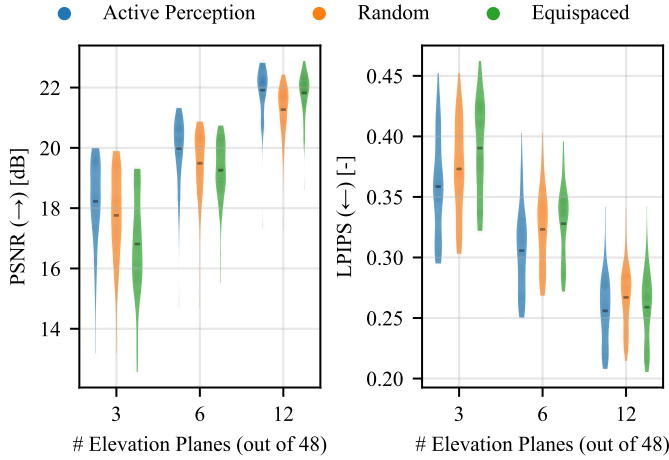
Fig. 10. Reconstruction performance for the 3D dataset in terms of PSNR and LPIPS as a function of the number of scanned lines for various action selection policies. The figure shows a distribution over the data samples and includes the mean as a gray line.
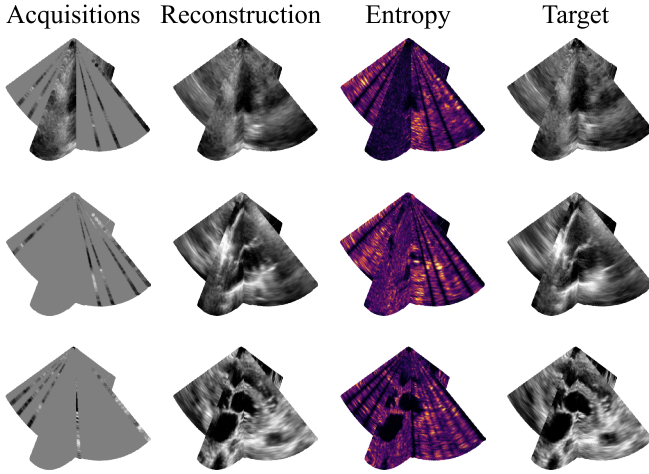

Fig. 11. Qualitative results on the 3D dataset. The figure shows the acquisitions and reconstructions for 6 / 48 elevation planes compared to the target. Additionally shows the posterior entropy, which drives action selection.

the 2D EchoNet-Dynamic dataset show significant benefits to using active perception, achieving similar reconstruction performance with only half the sampling budget of the baselines. Future work towards improving performance in the 2D regime might develop approaches to generative modeling that can model longer temporal context windows without sacrificing inference speed, leading to improvements in quality even with very low sampling budgets.

In our experiments on 3D data in Section V-C, we also find that active perception outperforms fixed sampling strategies across a range of sampling budgets, achieving a better trade-off between volume rate and reconstruction accuracy than prior works. These encouraging preliminary results highlight opportunities for further enhancement through improvements in key areas. In particular, training on a substantially larger 3D dataset (e.g., millions of volumes) would likely improve the model's reconstruction quality and the informativeness

of our derived uncertainty estimates. Furthermore, acquiring data with focusing in both the elevation and azimuthal directions would significantly enlarge the action space and allow for more targeted, information-efficient acquisition. Together, these enhancements have the potential to significantly boost the effectiveness of active perception in 3D ultrasound.

In our experiment using in-house echocardiograms, we chose line-by-line beamforming, although retrospective transmit beamforming (RTBF) could potentially yield higher-quality images. However, with RTBF, the measurement model A^ℓ becomes more challenging and no longer corresponds to an inpainting task. Future work could explore to better leverage the image quality benefits of RTBF.

To fully leverage active perception, the algorithm must operate in real-time with the frame acquisition. Given an imaging depth of 15 cm and a typical sound speed of 1540 m/s (common in echocardiography), each scan-line requires 195 μs. Acquiring 28 scan-lines results in a physical frame acquisition time of 5.46 ms. Thus, to achieve real-time performance, the algorithm still requires an approximate $4\times$ speedup.

Our experiment indicates that our method does not show bias against outlier patients when reconstruction quality is compared to ejection fraction. While further experiments could enhance our confidence in the method's robustness, this experiment serves as promising evidence that the model performs well across patient subgroups.

VII. CONCLUSION

We proposed a patient-adaptive focused transmit scheme that reduces the number of acquisitions needed for a high-quality ultrasound image by actively selecting the most informative measurements. Our method leverages posterior sampling with a temporal diffusion model and designs new measurements where the approximated posterior shows the most entropy. We have shown to outperform baselines on the 2D EchoNet-Dynamic dataset and a 3D Philips dataset, especially in cases with very little focused transmits. We have shown that active perception with focused transmits has improved gCNR compared to diverging waves with the same number of transmits. The proposed method did not show bias against outlier patients and showed that ejection fraction can still be accurately determined with only 2% of the transmits. The method can be run in real-time at over 40 Hz on GPU accelerators from 2023.

REFERENCES

[1] J. Wise, "Everyone's a radiologist now," *BMJ : Br. Med. J.*, vol. 336, no. 7652, pp. 1041–1043, May 2008.

[2] L. Demi, M. D. Verweij, and K. W. Van Dongen, "Parallel transmit beamforming using orthogonal frequency division multiplexing applied to harmonic Imaging-A feasibility study," *IEEE Trans. on Ultrason. Ferroelectr. Freq. Control*, vol. 59, no. 11, pp. 2439–2447, Nov. 2012.

[3] H. Huang, R. S. Wu, M. Lin, and S. Xu, "Emerging wearable ultrasound technology," *IEEE Trans. on Ultrason. Ferroelectr. Freq. Control*, vol. 71, no. 7, pp. 713–729, 2023.

[4] N. Ottakath, S. Al-Maadeed, A. Bouridane, M. E. Chowdhury, and K. K. Sadasivuni, "Wearable ultrasound devices for continuous health monitoring: Current and future prospects," in *2024 IEEE 8th Energy Conf. (ENERGYCON)*, IEEE, 2024, pp. 1–6.

[5] H. Hadri, A. Fail, M. Sadik, and A. Essaken, "Ultrasound beamforming: Exploring cloud-native and edge computing solution," in *2024 4th Int. Conf. on Technol. Adv. Comput. Sci. (ICTACS)*, IEEE, 2024, pp. 1339–1343.

[6] R. J. Van Sloun, "Active inference and deep generative modeling for cognitive ultrasound," *IEEE Trans. on Ultrason. Ferroelectr. Freq. Control*, pp. 1–1, 2024.

[7] R. J. G. van Sloun, J. C. Ye, and Y. C. Eldar, "Deep learning for ultrasound beamforming," in *Deep Learning for Biomedical Image Reconstruction*, J. C. Ye, Y. C. Eldar, and M. Unser, Eds. Cambridge University Press, 2023, pp. 223–251.

[8] J. D. Thomas and D. N. Rubin, "Tissue harmonic imaging: Why does it work?" *J. Am. Soc. Echocardiogr.*, vol. 11, no. 8, pp. 803–808, Aug. 1998.

[9] R. Bajcsy, "Active perception," *Proc. IEEE*, vol. 76, no. 8, pp. 966–1005, 1988.

[10] D. Kersten, P. Mamassian, and A. Yuille, "Object perception as bayesian inference," *Annu. Rev. Psychol.*, vol. 55, no. 1, pp. 271–304, 2004.

[11] T. Rainforth, A. Foster, D. R. Ivanova, and F. Bickford Smith, "Modern bayesian experimental design," *Stat. Sci.*, vol. 39, no. 1, pp. 100–114, 2024.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Adv. neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[13] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based Generative Modeling through Stochastic Differential Equations," in *9th Int. Conf. on Learn. Represent. ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[14] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 8633–8646, 2022.

[15] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Int. Conf. on Learn. Represent.*, 2021.

[16] B. Efron, "Tweedie's formula and selection bias," *J. Am. Stat. Assoc.*, vol. 106, no. 496, pp. 1602–1614, 2011.

[17] H. Chung, J. Kim, M. T. McCann, M. L. Klasky, and J. C. Ye, "Diffusion Posterior Sampling for General Noisy Inverse Problems," in *The Eleventh Int. Conf. on Learn. Represent. ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

[18] A. Ramkumar and A. K. Thittai, "Strategic undersampling and recovery using compressed sensing for enhancing ultrasound image quality," *IEEE Trans. on Ultrason. Ferroelectr. Freq. Control*, vol. 67, no. 3, pp. 547–556, 2019.

[19] D. Friboulet, H. Liebgott, and R. Prost, "Compressive sensing for raw rf signals reconstruction in ultrasound," in *2010 IEEE Int. Ultrason. Symp.*, IEEE, 2010, pp. 367–370.

[20] R. Cohen and Y. C. Eldar, "Sparse convolutional beamforming for ultrasound imaging," *IEEE Trans. on Ultrason. Ferroelectr. Freq. Control*, vol. 65, no. 12, pp. 2390–2406, 2018.

[21] A. Mamistvalov, A. Amar, N. Kessler, and Y. C. Eldar, "Deep-learning based adaptive ultrasound imaging from sub-nyquist channel data," *IEEE Trans. on Ultrason. Ferroelectr. Freq. Control*, vol. 69, no. 5, pp. 1638–1648, 2022.

[22] T. Di Ianni and R. D. Airan, "Deep-fus: A deep learning platform for functional ultrasound imaging of the brain using sparse data," *IEEE Trans. on Med. Imaging*, vol. 41, no. 7, pp. 1813–1825, 2022.

[23] D. Xiao, W. M. Pitman, B. Y. Yiu, A. J. Chee, and C. Alfred, "Minimizing image quality loss after channel count reduction for plane wave ultrasound via deep learning inference," *IEEE Trans. on Ultrason. Ferroelectr. Freq. Control*, vol. 69, no. 10, pp. 2849–2861, 2022.

[24] I. A. Huijben, B. S. Veeling, K. Janse, M. Mischi, and R. J. van Sloun, "Learning sub-sampling and signal recovery with applications in ultrasound imaging," *IEEE Trans. on Med. Imaging*, vol. 39, no. 12, pp. 3955–3966, 2020.

[25] I. A. Huijben, W. Kool, M. B. Paulus, and R. J. Van Sloun, "A review of the gumbel-max trick and its extensions for discrete stochasticity in machine learning," *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1353–1371, 2022.

[26] S. Afrakhteh, G. Iacca, and L. Demi, "High frame rate ultrasound imaging by means of tensor completion: Application to echocardiography," *IEEE Trans. on Ultrason. Ferroelectr. Freq. Control*, vol. 70, no. 1, pp. 41–51, 2022.

[27] J. Zbontar *et al.*, "Fastmri: An open dataset and benchmarks for accelerated MRI," *CoRR*, vol. abs/1811.08839, 2018.

[28] A. Sriram *et al.*, "End-to-end variational networks for accelerated mri reconstruction," in *Med. Image Comput. Comput. Assist. Interv. 2020: 23rd Int. Conf. Lima, Peru, Oct. 4–8, 2020, Proceedings, Part II 23*, Springer, 2020, pp. 64–73.

[29] C. D. Bahadir, A. V. Dalca, and M. R. Sabuncu, "Learning-based optimization of the under-sampling pattern in mri," in *Inf. Process. Med. Imaging: 26th Int. Conf. IPMI 2019, Hong Kong, China, June 2–7, 2019, Proc. 26*, Springer, 2019, pp. 780–792.

[30] I. A. Huijben, B. S. Veeling, and R. J. van Sloun, "Learning sampling and model-based signal recovery for compressed sensing mri," in *ICASSP 2020-2020 IEEE Int. Conf. on Acoust. Speech Signal Process. (ICASSP)*, IEEE, 2020, pp. 8906–8910.

[31] H. Van Gorp, I. Huijben, B. S. Veeling, N. Pezzotti, and R. J. Van Sloun, "Active deep probabilistic subsampling," in *Int. Conf. on Mach. Learn.*, PMLR, 2021, pp. 10 509–10 518.

[32] T. Yin, Z. Wu, H. Sun, A. V. Dalca, Y. Yue, and K. L. Bouman, "End-to-end sequential sampling and reconstruction for MRI," in *Mach. Learn. for Heal. ML4H@NeurIPS 2021, 04 Dec. 2021, Virtual Event*, S. Roy *et al.*, Eds., ser. Proceedings of Machine Learning Research, vol. 158, PMLR, 2021, pp. 261–281.

[33] O. Nolan, T. Stevens, W. L. van Nierop, and R. V. Sloun, "Active diffusion subsampling," *Trans. on Mach. Learn. Research*, 2025.

[34] G. Yiasemis, J. Sonke, and J. Teuwen, "End-to-end adaptive dynamic subsampling and reconstruction for cardiac MRI," *CoRR*, vol. abs/2403.10346, 2024.

[35] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Med. image computing computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Ger. Oct. 5-9, 2015, proceedings, part III 18*, Springer, 2015, pp. 234–241.

[36] L. Rout, N. Raoof, G. Daras, C. Caramanis, A. Dimakis, and S. Shakkottai, "Solving linear inverse problems provably via posterior sampling with latent diffusion models," *Adv. Neural Inf. Process. Syst.*, vol. 36, 2024.

[37] D. Stojanovski, U. Hermida, P. Lamata, A. Beqiri, and A. Gomez, "Echo from noise: Synthetic ultrasound image generation using diffusion models for real image segmentation," in *Int. Workshop on Adv. Simpl. Med. Ultrasound*, Springer, 2023, pp. 34–43.

[38] T. S. Stevens, F. C. Meral, J. Yu, I. Z. Apostolakis, J.-L. Robert, and R. J. Van Sloun, "Dehazing ultrasound using diffusion models," *IEEE Trans. on Med. Imaging*, 2024.

[39] Y. Zhang, C. Huneau, J. Idier, and D. Mateus, "Ultrasound image reconstruction with denoising diffusion restoration models," in *Int. Conf. on Med. Image Comput. Comput. Interv.*, Springer, 2023, pp. 193–203.

[40] T. S. Stevens, O. Nolan, J.-L. Robert, and R. J. Van Sloun, "Sequential posterior sampling with diffusion models," in *ICASSP 2025-2025 IEEE Int. Conf. on Acoust. Speech Signal Process. (ICASSP)*, IEEE, 2025, pp. 1–5.

[41] J. R. Hershey and P. A. Olsen, "Approximating the kullback leibler divergence between gaussian mixture models," in *2007 IEEE Int. Conf. on Acoust. Speech Signal Process.*, IEEE, vol. 4, 2007, pp. IV–317.

[42] T. S. Stevens *et al.*, *zea: A Toolbox for Cognitive Ultrasound Imaging*, version 0.0.1, Jul. 2025.

[43] D. Ouyang *et al.*, "Video-based AI for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, no. 7802, pp. 252–256, Apr. 2020.

[44] J. Bradbury *et al.*, *JAX: Composable transformations of Python+NumPy programs*, version 0.3.13, 2018.

[45] T. S. W. Stevens, O. Nolan, O. Somphone, J. Robert, and R. J. G. van Sloun, "High volume rate 3d ultrasound reconstruction with diffusion models," *CoRR*, vol. abs/2505.22090, 2025.