Patient-Adaptive Echocardiography 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 transmits are the most commonly used transmit strategy for echocardiograms, but suffer 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 and receive 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 acquiring the most informative transmits. This cognitive ultrasound modality outperforms random and equispaced subsampling in terms of distortion and perceptual metrics on the 2D EchoNet-Dynamic dataset and a 3D Philips dataset, where we actively select focused elevation planes. Furthermore, our method improves generalized contrast-to-noise ratio from 0.83 to 0.89 compared to the same number of diverging wave transmits on six in-house echocardiograms. Additionally, we can segment the left ventricle, with on average 0.91 Dice-Sørensen coefficient, through simulating using 2 out of 112 lines. Finally, our method can be run in real-time on GPU accelerators from 2023, increasing the maximum achievable frame-rate from 46 Hz to 58 Hz. The code is publicly available at https://tue-bmd.github.io/casl/.**

***Index Terms*— Beamforming, cognitive ultrasound, diffusion models, echocardiography**

I. Introduction

ULTRASOUND imaging is one of the most widely used medical imaging modalities. It offers 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].

Among its clinical applications, echocardiography is the first-line non-invasive technique for assessing cardiac structure and function. It is essential for diagnosing and monitoring major cardiovascular conditions, including heart failure, heart valve disease, cardiomyopathy, ischemic heart disease, and pulmonary hypertension [2].

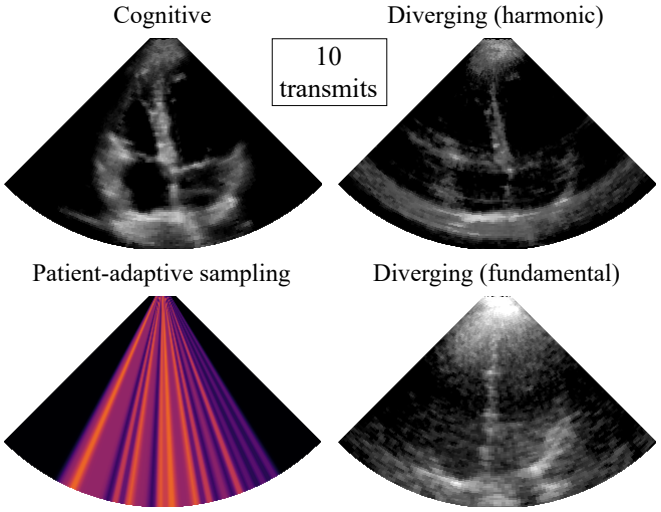

Fig. 1. Cognitive ultrasound Acquisition of Scan Lines (CASL) reduces the number of acquisitions needed to obtain a high-quality ultrasound image by actively selecting those measurements that are expected to be most informative. Each image is created using 10 transmit events.

There exist various techniques for ultrasound image formation. The transmit scheme can be designed using focused or unfocused events. Focusing permits the concentration of acoustic energy at specific locations within the body. Focusing the beam will provide enhanced lateral resolution, signal-to-noise ratio (SNR), and penetration depth in echocardiography [3]. Another important reason to opt for focused transmits is the generation of high-amplitude pressure fields, which are necessary for the generation of harmonic components used in harmonic imaging [4]. Harmonic imaging has become the gold standard for echocardiograms due to the superior image quality in hard-to-image patients, reducing reverberation artifacts, clutter, and near-field artifacts [3], [5]–[9]. Unfocused transmit events, such as diverging waves, are typically coherently compounded to improve SNR [10]. However, compounding can result in motion decorrelation in high-motion scenarios, which means that the tissue has moved a significant amount in between the transmit events, reducing image quality [11]–[13]. Due to these advantages, focused transmits are the most widely used transmit strategy in commercial ultrasound systems [14]. The downside is that the excited area is much smaller, meaning

This work was supported by the European Research Council (ERC) under the ERC starting grant nr. 101077368 (US-ACT).

Wessel L. van Nierop and Oisín Nolan contributed equally to this work.

Wessel L. van Nierop, Oisín Nolan, Tristan S.W. Stevens, and Ruud J.G. van Sloun are with the Department of Electrical Engineering, Eindhoven University of Technology, 5612 AZ Eindhoven, The Netherlands (email: w.l.v.nierop@tue.nl; o.i.nolan@tue.nl; t.s.w.stevens@tue.nl; r.j.g.v.sloun@tue.nl)

that more transmit events are necessary to cover a given field of view (FOV) compared to unfocused transmit events, reducing the temporal resolution.

For accurate diagnosis in echocardiography a frame rate of at least 40 Hz is required [15], while for certain medical conditions, such as tachycardia, over 80 Hz is advisable [16]. Temporal resolution in ultrasound image formation is determined by the number of transmit events, imaging depth, and the speed of sound. For an imaging depth of 15 cm and a typical sound speed of 1540 m/s (both common in echocardiography), each transmit requires 195 μs, which translates to a frame rate of over 5 kHz. Because focused transmits do not cover a wide FOV with high SNR and high-amplitude pressure fields, many transmit events are often necessary. For example, a phased-array transducer with an aperture of 20 mm and center frequency of 3 MHz, has an angular resolution of approximately 1.5°, which means 122 lines are needed for the full angular resolution (half beamwidth spacing) in a 90° sector [17]. This reduces the best achievable frame rate to 42 Hz, or 21 Hz when using pulse inversion for harmonic imaging [18]. In order to image phenomena with higher temporal resolution, for example valve dynamics, one might narrow the field of view or resort to M-mode imaging [19]. This, however, imposes a trade-off between field of view and temporal resolution. For 3D echocardiograms, even more focused transmit events are necessary, meaning that it is hard to obtain high-quality and fast 3D echocardiograms. This shows a need for a reduction in the number of transmit events while maintaining high image quality to ensure diagnostic accuracy.

In addition to accelerating frame rates, reducing the number of necessary transmit events also reduces certain cost factors associated with the acquisition. Such cost factors include power usage and on-device compute, which currently bottleneck imaging modalities that depend on battery power and require small form factor, such as wearable ultrasound patches for continuous monitoring [20], [21]. 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 [22].

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 (Fig. 1). This fits into the paradigm of *cognitive ultrasound*, recently proposed by van Sloun [23], in which the imaging process is modeled as an autonomous agent that actively designs future transmit events to maximize information gain. 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 measurements it observes. Based on these beliefs, the agent pursues focused acquisitions that have the highest expected information gain. This results in a drastic reduction in the number of focused transmit events per frame and thus increases the frame rate, offering a potential alternative to unfocused transmits. We refer to our method throughout the paper as *CASL*: ***C**ognitive ultrasound **A**cquisition of **S**can **L**ines*.

A. Related Work

Reducing the number of transmit events necessary to create an ultrasound image has the effect of reducing the data rate and increasing the maximum potential frame rate. In this section, we review a number of existing approaches to achieving these goals.

Firstly, we consider methods that optimize the focusing strategy to gain more information about the target per transmit. Multi-line transmission (MLT), for example, generates multiple focal points during transmission, such that the number of transmit events is reduced. However, the parallel focused beams do suffer from interference, which can result in visible artifacts in the image [24]. Furthermore, methods such as simultaneous axial multifocal imaging (SAMI) [25], improve the level of focusing along the axial dimension, decreasing the need for focusing along different depths using multiple transmits. MLT and SAMI share the same goal as this paper, but are complementary to our method, which can use any focused transmit.

Çakiroğlu *et al.* [26] take an alternative approach, optimizing a single set (non-adaptive) of transmit delays and weights through backpropagation to improve the recovery of imaging parameters across a large dataset of simulated examples. CASL, in contrast, takes a cognitive approach, optimizing transmit parameters adaptively.

Another popular approach to reducing data rates has been to use compressed sensing (CS) [27]. For instance, Chernyakova *et al.* [28] propose a Fourier-domain formulation, recognizing that the Fourier coefficients of the received signals can be obtained from their low-rate samples. CS has also been applied to recover fully-sampled RF channel data from subsampled measurements (e.g. RF samples or entire channels/elements), for example by leveraging sparsity in the wave atom basis [29], [30]. Beyond subsampling, other work has proposed compressive multiplexing of channels [31], [32]. For reconstruction beyond conventional sparsity priors, deep learning (DL) [33] has become increasingly attractive. In ultrasound, it has been used to recover beamformed scan lines and fully-sampled RF channel data from e.g. sparse array designs [34], [35]. As with MLT and SAMI, such approaches to reducing the channel data rates can also be used in combination with CASL, which reduces the number of transmit events (transmit rates).

Other methods subsample at the level of the transmit, as CASL does. For example, the approach by Huijben *et al.* [36] learns subsampling masks for both channels and slow-time frames jointly with a convolutional neural network (CNN) reconstruction model, using the Gumbel-Max trick [37] to backpropagate through the subsampling operation. Similarly, Lorintiu *et al.* [38] employ dictionary learning to reconstruct 3D ultrasound volumes from subsampled scan lines. Afrakteh *et al.* also tackle scan-line subsampling, using random masks with tensor completion for reconstruction [39]. Each of the above methods uses either a fixed mask, optimized for a particular task or group of patients, or a random mask. CASL, in contrast, designs *patient-adaptive* sampling masks, in real-time. Given the variability that exists across patient anatomies,

patient-adaptive subsampling is essential to minimize redundancy and therefore maximize information gain. While patient-adaptive subsampling algorithms have been proposed for other medical imaging modalities, such as MRI [40]–[43] and X-Ray CT [44], this work is, to our knowledge, the first application of patient-adaptive subsampling to ultrasound video recovery.

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 recover the missing portion of the signal. We therefore choose to use diffusion models, which have shown excellent performance in inpainting [45], [46] to solve this problem, outperforming traditional CS and supervised DL methods [47]. This modeling choice is further motivated by recent success in applying diffusion models to the domain of ultrasound, for synthetic data generation [48], dehazing [49], and image reconstruction [50].

B. Contributions

To summarize, this paper presents the following main contributions. (1) We propose a method for reconstructing ultrasound video using a minimal number of transmit events with a temporal diffusion model that exploits the sequential nature of ultrasound. (2) We propose an algorithm for designing a transmit scheme which maximizes information gain in a computationally efficient way. (3) We present experimental results showing that acquiring patient-adaptive focused transmits outperforms diverging waves for the same number of transmit events in terms of generalized contrast-to-noise ratio (gCNR).

II. Theoretical Background

A. Active Perception

The goal of sensing is to acquire measurements 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 [51]. 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 [23]. 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 [52]. The aforementioned goal of sensing may then be formalized in Bayesian terms, where H is the entropy functional, $p(\cdot)$ denotes a probability density function, $\boldsymbol{x}$ denotes the environmental parameters to be estimated, A is the set of sensing actions, and $\boldsymbol{y}$ denotes the resulting observations [53]:

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

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

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 [53]:

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

where $I(\boldsymbol{x}; \boldsymbol{y} \mid A)$ denotes the conditional mutual information of $\boldsymbol{x}$ and $\boldsymbol{y}$ given A. Active perception is often performed *greedily*, and *iteratively*, first selecting the optimal sensing action according to (2), performing inference on $\boldsymbol{x}$ given the new observations $\boldsymbol{y}$, 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*.

B. Posterior Sampling with Diffusion Models

As mentioned in Section II-A, 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 [54]–[56]. 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 $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. 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 $[\boldsymbol{x}_0, \ldots, \boldsymbol{x}_\tau, \ldots, \boldsymbol{x}_{\tau_{\text{max}}}]$ interpolating between $\boldsymbol{x}_0$ and $\boldsymbol{x}_{\tau_{\text{max}}} = \boldsymbol{\epsilon}$. 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 [57]. 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 [58] 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

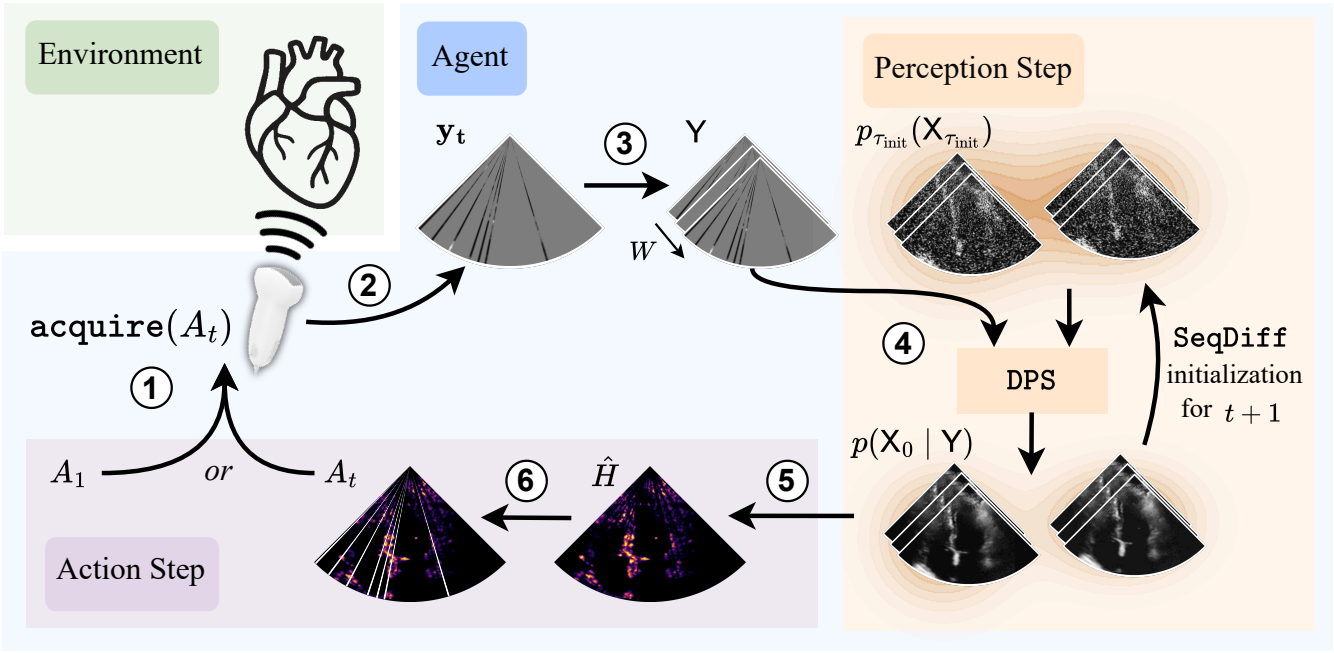

Fig. 2. ① Acquire scan lines at locations specified by $\boldsymbol{A_t}$ or initial locations $\boldsymbol{A_1}$. ② Beamform the acquired scan lines in a zero-filled target region to produce $\boldsymbol{y_t}$. ③ Place $\boldsymbol{y_t}$ in the measurement buffer $\mathbf{Y}$. ④ Perform posterior sampling using diffusion posterior sampling (DPS). ⑤ Compute pixel-wise entropy from the posterior distribution. ⑥ Select the next scan locations using *K-Greedy Entropy Minimization*, and repeat.

produced by the minimum mean squared error denoiser:

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

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 [45] solves this problem by formulating a posterior score function:

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

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:

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

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. Method

In this section, we present CASL 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 2.

A. Perception

Perception, as defined in Section II-A, can be formalized as Bayesian posterior inference. In our case, this amounts to inferring the tissue state $\boldsymbol{x}_t$ at time t given the history of observations and actions until that point, i.e., the distribution $p(\boldsymbol{x}_t \mid \boldsymbol{h}_t)$, where $\boldsymbol{h}_t$ indicates the observation history at time t, consisting of the actions $A_{1:t}$ taken so far, and their resulting observations $\boldsymbol{y}_{1:t}$. Concretely, the tissue state $\boldsymbol{x}_t \in \mathbb{R}^N$ is represented by a fully-sampled image, and the measurements $\boldsymbol{y}_t \in \mathbb{R}^N$ are partially-sampled images containing zeros at unmeasured tissue locations. Functionally, the perception step in CASL produces two important quantities. Firstly, a point estimate (e.g. a posterior sample or the posterior mean), which can serve as the *reconstruction* image for a given frame, and secondly, a pixel-wise uncertainty map (derived from multiple samples) which is used to drive information-maximizing action selection, discussed in Section III-B. Both of these quantities can be estimated from a set of posterior samples, motivating our use of the DPS algorithm, introduced in Section II-B. 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 $\boldsymbol{y}_{1:t}$. To model such dependencies, we fit the diffusion model on sequences of W consecutive frames $\mathbf{X} = [\boldsymbol{x}_{t-W+1}, \ldots, \boldsymbol{x}_t]$ 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 $W = 3$. The denoiser $\boldsymbol{\epsilon}_\theta$ is implemented using a U-Net architecture [59] of 2M parameters, with further details available on the GitHub repository.

During inference, at each time step t we generate a set of N_p tensors $\{\mathbf{X}_0^{(i)}\}_{i=1}^{N_p}$ in parallel. The final image $\mathbf{X}_0^{(i)}[W]$ in each generated tensor represents one possible state of $\boldsymbol{x}_t$. These images, dubbed *particles*, can then be used to approximate the posterior distribution $p(\boldsymbol{x}_t \mid \boldsymbol{h}_t)$. Throughout the paper, we refer to this set of particles $\{\boldsymbol{x}_t^{(i)}\}_{i=1}^{N_p}$ as the agent's posterior distribution at time t, with differences across particles indicating uncertainty in the state of $\boldsymbol{x}$. Throughout our experiments, we use $N_p = 2$, leading to minimal computational overhead.

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 $\mathbf{Y} = [\boldsymbol{y}_{t-W+1}, \ldots, \boldsymbol{y}_t]$. Then, we define a measurement model simulating focused line-scanning. This model assumes that for each focused transmit, a set 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 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., $p(\mathbf{Y}|\mathbf{X}, \mathbf{A}) = \delta(\mathbf{Y} - \mathbf{A} \odot \mathbf{X})$, where $\odot$ denotes an element-wise product. 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:

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

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 [60] initialization. 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 posterior distribution. The first particle is chosen as a random sample from the posterior distribution to ensure that the reconstruction image is on-manifold [45], enhancing perceptual quality. Alternate reconstruction functions may, however, be used depending on the goal, for example, to minimize distortion we can approximate the posterior mean by averaging posterior samples [61]. Finally, we *hard-project* the measurements onto the reconstruction, setting the values at measured pixels to be identical to those in the measurement. Throughout this work, we compute image-quality metrics on the images in the polar domain, before scan conversion.

B. Action

The action step aims to choose a set of actions to take at time $t + 1$ given the posterior distribution at time t. An action in this case corresponds to firing a focused transmit beam at a particular steering angle. The action space is therefore represented by a discretized set of possible focused scan locations $\{A^\ell \mid \ell = 1, 2, \ldots, L\}$, where there are L total locations. We denote by A^ℓ 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. For brevity, we use the term *action* to refer to the pixel indices A^ℓ corresponding to that action, and we use a time subscript A_t^ℓ to denote a specific action to be taken at time t. The actions should be chosen to maximize information gain with respect to the tissue state, following the objective described in Section II-A. Starting with the expected information gain objective provided in (2), and following van Sloun [23], we derive our action selection policy, substituting in the likelihood function specified in (8). Because the conditional mutual information is symmetric, we can compute the expected reduction in uncertainty about $\boldsymbol{x}_t$ in terms of uncertainty about the observations under future actions, avoiding the need for an additional perception step on hypothetical observations. Then, for computational efficiency, rather than optimizing directly for the future action A_{t+1}^ℓ, which would require simulating a future posterior distribution, we make use of the present posterior distribution, and instead optimize for the hypothetical action $A_t^{\ell'}$ that would result in the most informative additional observation $\boldsymbol{y}_t'$ given the history $\boldsymbol{h}_t$, leading to the following measure of information gain:

$$\begin{aligned} I(\boldsymbol{y}_t'; \boldsymbol{x}_t \mid \boldsymbol{h}_t) &= H(\boldsymbol{y}_t' \mid A_t^{\ell'}, \boldsymbol{h}_t) - H(\boldsymbol{y}_t' \mid \boldsymbol{x}_t, A_t^{\ell'}, \boldsymbol{h}_t) \\ &= H(\boldsymbol{y}_t' \mid A_t^{\ell'}, \boldsymbol{h}_t) - H(\boldsymbol{n}). \end{aligned} \tag{10}$$

The second entropy term $H(\boldsymbol{y}_t' \mid \boldsymbol{x}_t, A_t^{\ell'}, \boldsymbol{h}_t)$ 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 $A_t^{\ell'}$, yielding the following objective, selecting the most informative location ℓ':

$$\arg\max_{\ell'} I(\boldsymbol{y}_t'; \boldsymbol{x}_t \mid A_t^{\ell'}, \boldsymbol{h}_t) = \arg\max_{\ell'} H(\boldsymbol{y}_t' \mid A_t^{\ell'}, \boldsymbol{h}_t). \tag{11}$$

The remaining entropy values for each line measurement $H(\boldsymbol{y}_t' \mid A_t^{\ell'}, \boldsymbol{h}_t)$ can be decomposed into sums of pixel-wise entropies by modeling the pixels as independent variables. Given that pixels masked by $A_t^{\ell'}$ 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 i^{th} pixel of $\boldsymbol{x}_t$:

$$H(\boldsymbol{y}_t' \mid A_t^{\ell'}, \boldsymbol{h}_t) = \sum_{\mathrm{i} \in A_t^{\ell'}} H(\boldsymbol{x}_t[\mathrm{i}] \mid A_t^{\ell'}, \boldsymbol{h}_t). \tag{12}$$

In practice, we first compute a pixel-wise entropy map in the image domain $\boldsymbol{x}_t$, $\hat{H} = [\hat{H}[1], \ldots, \hat{H}[\mathrm{i}], \ldots, \hat{H}[N]]^\top$, where $\hat{H}[\mathrm{i}] \approx H(\boldsymbol{x}_t[\mathrm{i}] \mid A_t^{\ell'}, \boldsymbol{h}_t)$. Given $\hat{H}$, we can sum the pixels corresponding to each action $A_t^{\ell'}$ in order to get the line-wise measurement entropies $\hat{H}^\ell$, choosing the maximum entropy line as the next action. Using the variational entropy approximation proposed by Hershey *et al.* [62], the pixel-wise entropy map $\hat{H}$ can be computed by taking the element-wise squared error between each pair of particles in the posterior distribution $\{\boldsymbol{x}_t^{(i)}\}_{i=1}^{N_p}$, as follows, where the exp and log functions are applied element-wise:

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

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

$$\arg\max_{\ell'} H(\boldsymbol{y}_t' \mid A_t^{\ell'}, \boldsymbol{h}_t) \approx \arg\max_{\ell'} \sum_{\mathrm{i} \in A_t^{\ell'}} \hat{H}[\mathrm{i}]. \quad (14)$$

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, decreasing the frame rate. In order to prevent this, we propose an approximate algorithm, called **K-Greedy Entropy Minimization**, where we batch the measurements that will be used to generate each frame and perform perception on that batch. This algorithm 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. In the algorithm, we use the notation A_t to gather the set of pixel indices from all selected lines.

IV. Experiments

A comprehensive evaluation of CASL'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. Lastly, we show that our method can also be applied to 3D echocardiography, where we subsample elevation planes. In all experiments we retrospectively subsample all lines to simulate acquiring only the selected lines. CASL will be compared to equispaced and random subsampling, using the same diffusion model for perception. The equispaced subsampler "rolls" the selected lines from left to right, such that over time the full imaging area is measured. Random sampling

Algorithm 1 CASL Perception-Action Loop

Require: `SeqDiff` initial diffusion step $\tau_{\texttt{SeqDiff}}$; total diffusion steps $\tau_{\max}$; 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 $\{\alpha_\tau, \sigma_\tau\}_{\tau=0}^{\tau_{\max}}$; guidance weight γ; posterior variance $\sigma_{\boldsymbol{x}}^2$; temporal window size W; $\forall t < 1 : \boldsymbol{x}_t \leftarrow \mathbf{0}, \boldsymbol{y}_t \leftarrow \mathbf{0}$.

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

for $t \in [1, ..., T]$ **do**
  $\boldsymbol{y}_t \leftarrow \texttt{acquire}(A_t)$ // Acquire measurements
  $\mathbf{Y} \leftarrow [\boldsymbol{y}_{t-W+1}, \ldots, \boldsymbol{y}_t]$ // Measurement buffer
  $\mathbf{A} \leftarrow [\mathcal{M}(A_{t-W+1}), \ldots, \mathcal{M}(A_t)]$ // Mask buffer
  if $t = 1$ **then**
    $\tau_{\text{init}} \leftarrow \tau_{\max}$
  else
    $\tau_{\text{init}} \leftarrow \tau_{\texttt{SeqDiff}}$
  Perception Step
  for each $i \in \{1, ..., N_p\}$ in parallel **do**
    $\mathbf{X} \leftarrow [\boldsymbol{x}_{t-W}^{(i)}, \ldots, \boldsymbol{x}_{t-1}^{(i)}]$ // $\mathbf{X}$ shorthand for $\mathbf{X}_t^{(i)}$
    $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ // Initial noise
    $\mathbf{X}_{\tau_{\text{init}}} \leftarrow \alpha_{\tau_{\text{init}}} \mathbf{X} + \sigma_{\tau_{\text{init}}} \boldsymbol{\epsilon}$ // Initial samples
    for $\tau \in [\tau_{\text{init}}, ..., 0)$ **do**
      $\hat{\boldsymbol{\epsilon}} \leftarrow \boldsymbol{\epsilon}_\theta(\mathbf{X}_\tau, \sigma_\tau^2)$ // Predict noise
      $\hat{\mathbf{X}}_0 \leftarrow (\mathbf{X}_\tau - \sigma_\tau \hat{\boldsymbol{\epsilon}})/\alpha_\tau$ // Tweedie Estimate
      $\mathbf{X}'_{\tau-1} \leftarrow \alpha_{\tau-1} \hat{\mathbf{X}}_0 + \sigma_{\tau-1} \hat{\boldsymbol{\epsilon}}$ // Prior step
      $\mathbf{X}_{\tau-1} \leftarrow \mathbf{X}'_{\tau-1} - \gamma \nabla_{\mathbf{X}_\tau} ||\mathbf{Y} - \mathbf{A} \odot \hat{\mathbf{X}}_0||_2^2$
    $\boldsymbol{x}_t^{(i)} \leftarrow \mathbf{X}_0[W]$ // Posterior distribution t
  $\tilde{\boldsymbol{x}}_t \leftarrow \boldsymbol{x}_t^{(1)}$ // Choose first as reconstruction
  Action Step
  $A_{t+1} \leftarrow \emptyset$ // Initialize action set for next transmit
  $\hat{H} \leftarrow -\sum_i \frac{1}{N_p} \log \sum_j \frac{1}{N_p} \exp\left[-\frac{(\boldsymbol{x}_t^{(i)} - \boldsymbol{x}_t^{(j)})^2}{2\sigma_{\boldsymbol{x}}^2}\right]$
  $\forall \ell : \hat{H}^\ell \leftarrow \sum_{\mathrm{i} \in A^\ell} \hat{H}[\mathrm{i}]$ // Line-wise entropy
  for $k \in [1, \ldots, K]$ **do**
    $\ell^* \leftarrow \arg\max_\ell \hat{H}^\ell$ // Select max entropy action
    $A_{t+1} \leftarrow A_{t+1} \cup A^{\ell^*}$ // Gather selected actions
    $\forall \ell : \hat{H}^\ell \leftarrow \hat{H}^\ell * \left(1 - \exp\left(-\frac{(\ell - \ell^*)^2}{2}\right)\right)$
return $\{\tilde{\boldsymbol{x}}_t\}_{t=1}^T$

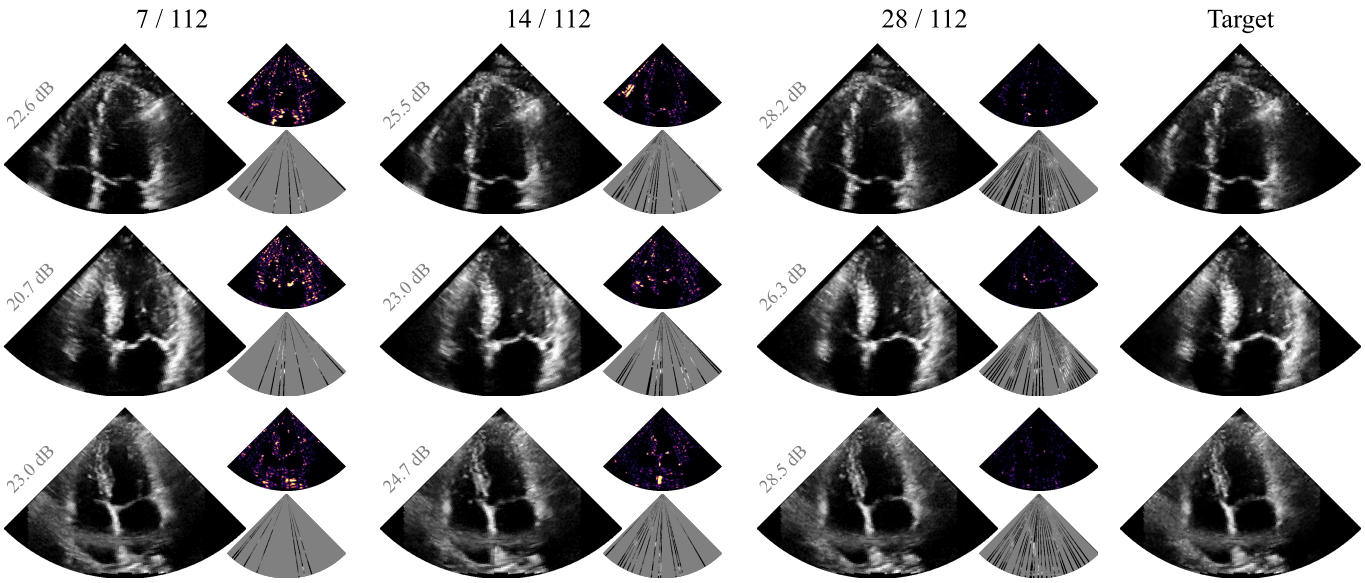

Fig. 3. Qualitative results on the EchoNet-Dynamic dataset. The figure shows the reconstructions for 7, 14, and 28 lines compared to the target, for three different subjects in the test set. Additionally, it shows the reconstruction quality, in terms of PSNR, the acquired lines, and the posterior entropy, which drives action selection.

means that the selected lines were sampled from a uniform distribution. We implement CASL using `zea`, the cognitive ultrasound toolbox [63].

A. EchoNet-Dynamic

Here we train a diffusion model on the EchoNet-Dynamic dataset [64]. The EchoNet-Dynamic dataset consists of over 10k echocardiograms. As we do not have access to how 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 end, 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.

1) Reconstruction quality: The qualitative results are shown in Figure 3. Here, the 20th frame is used for three random patients in the test data. We show reconstructions for three subsampling rates. Additionally, we show the acquired lines, the entropy of the posterior samples, and the fully observed target images. The reconstructions are visually very similar to the targets, while in the extreme case, using only 7 out of 112 scan lines.

Figure 4 shows the reconstruction quality in terms of peak signal-to-noise ratio (PSNR) and learned perceptual image patch similarity (LPIPS) [65] as distributions over all the patients in the test dataset. The brackets in the figure indicate the win-rate of our method over the baselines. It can be seen

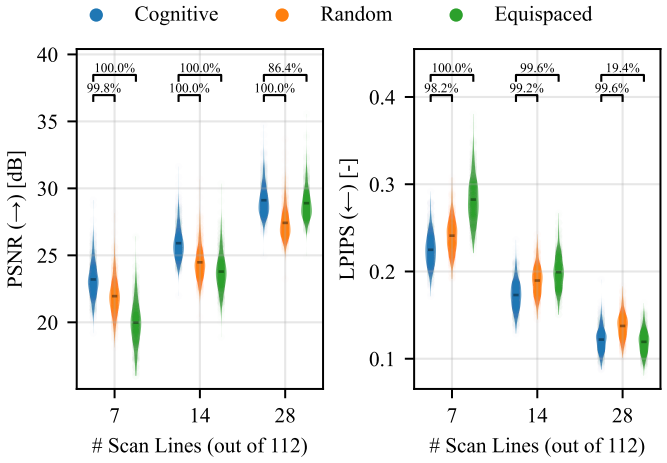

Fig. 4. Reconstruction performance for EchoNet-Dynamic in terms of PSNR and LPIPS [65] as a function of the number of scanned lines for various action selection policies. The figure shows a distribution over the data samples and includes the mean as a gray line. The brackets show the win-rate over the baselines. Cognitive is found to be significantly better ($\boldsymbol{p < 10^{-53}}$) than the other baselines for all subsampling rates and metrics using the Wilcoxon signed-rank test.

that cognitive 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, CASL still achieves a PSNR of 23.2 on average, which consists of a 5.7% improvement over random sampling and an impressive 16.2% improvement over equispaced sampling. The strong significance is explained by the result that CASL outperforms baselines for both metrics in virtually every cine loop.

2) Hyperparameters: Algorithm 1 relies on a number of hyperparameters, most notably the temporal window size W, guidance weight γ, and number of particles N_p. In order to measure the influence of these parameters on reconstruction

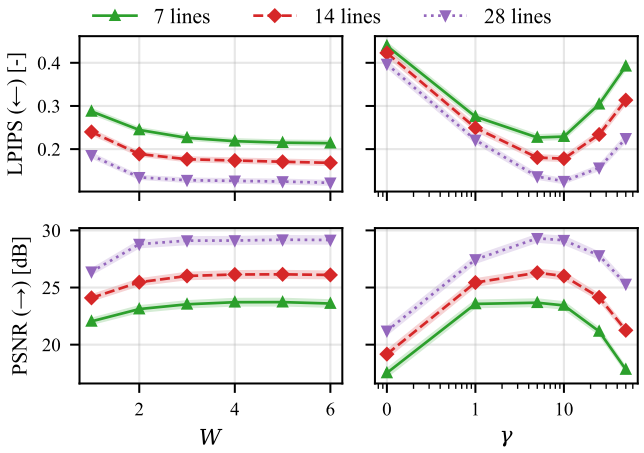

Fig. 5. Reconstruction quality on EchoNet-Dynamic validation samples as a function of hyperparameter values $\boldsymbol{W} \in [\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}, \mathbf{5}, \mathbf{6}]$ and $\gamma \in [\mathbf{0.1}, \mathbf{1}, \mathbf{5}, \mathbf{10}, \mathbf{25}, \mathbf{50}]$, while keeping the remaining parameters fixed. The mean result across 20 patients from the validation set with 100 frames each is plotted, with shaded regions indicating the standard error of the mean. For $\boldsymbol{W}$, performance plateaus at some point, depending on the number of scan lines, while γ exhibits a convex curve for both metrics with optima at $\gamma \approx \mathbf{10}$.

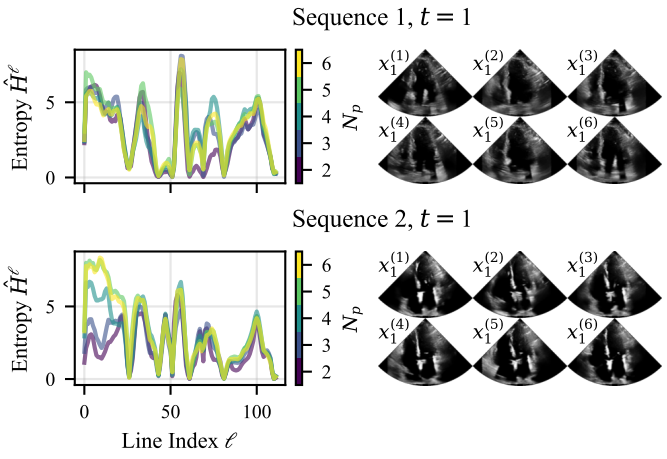

Fig. 6. Linewise entropy estimates generated using different numbers of particles $\boldsymbol{N_p}$ for the same input measurements on the first frame of two different sequences from the EchoNet-Dynamic validation set, where uncertainty will be highest. It is clear that while some differences exist between the estimates, the peaks lie in similar locations, leading to similar actions.

quality, a hyperparameter sweep was performed for each on a set of 20 patients from the EchoNet-Dynamic validation set with 100 frames each. In each case, all parameters aside from the one being swept over are fixed to the optimal values. Both W and γ were found to exert some influence on reconstruction quality, as illustrated in Figure 5. For W, we find that the quality plateaus at $W = 3$, thereby making $W = 3$ a sensible choice, since the required compute increases with W. We do however observe that the plateau appears later as the number of scan lines decreases, indicating that models with longer context windows may improve performance under more extreme undersampling. γ shows convex curves for both PSNR and LPIPS with optima around $\gamma = 10$. Interestingly,

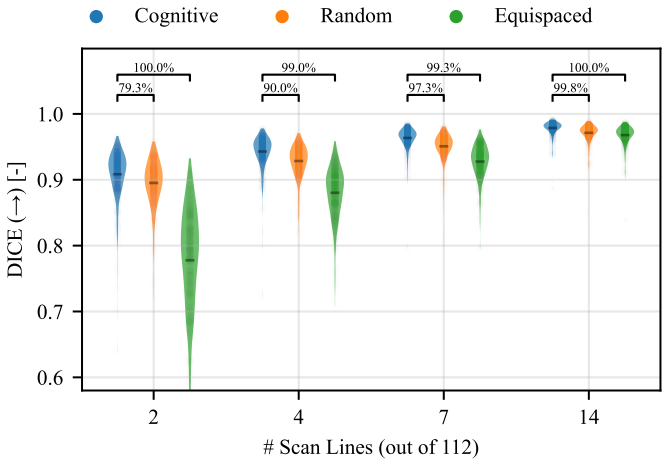

Fig. 7. Segmentation performance in terms of DICE of EchoNet-Dynamic on subsampled images for various action selection policies. The figure shows a distribution over the data samples and includes the mean as a gray line. The brackets show the win-rate over the baselines. Cognitive is found to be significantly better ($\boldsymbol{p} < \mathbf{10^{-29}}$) than the other baselines for all subsampling rates and metrics using the Wilcoxon signed-rank test.

N_p was found to have a negligible effect on reconstruction performance, with $N_p = 2$ performing almost identically to $N_p \in \{3, 4, 5, 6\}$. One explanation for this is the coarse nature of the action space, which sums the entropy along an entire scan line, meaning that subtle differences between particles along the vertical axis do not affect the overall linewise entropy estimates. This effect is illustrated by Fig. 6, which plots the linewise entropy estimates for the same frame using increasing values for N_p. It is clear in the figure that the entropy estimates are similar for increasing N_p, leading to similar actions. The result is that using $N_p = 2$ is sufficient to drive uncertainty-minimizing scan line selection without requiring excessive computational resources. While these entropy estimates lead to similar actions, it is worth commenting on *why* they differ for different N_p. Since the particles are randomly drawn from the posterior distribution, it can be that the set of particles agree with one another in a certain region *by coincidence*, resulting in an underestimation of the entropy in that region. Similarly, particles could coincidentally disagree in a region due to an unlikely mode of the distribution being over-sampled.

3) 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 [64] 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 the Dice-Sørensen coefficient (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 7 shows that CASL consistently produces the best left ventricle segmentations compared to equispaced and random subsampling. The performance for 2

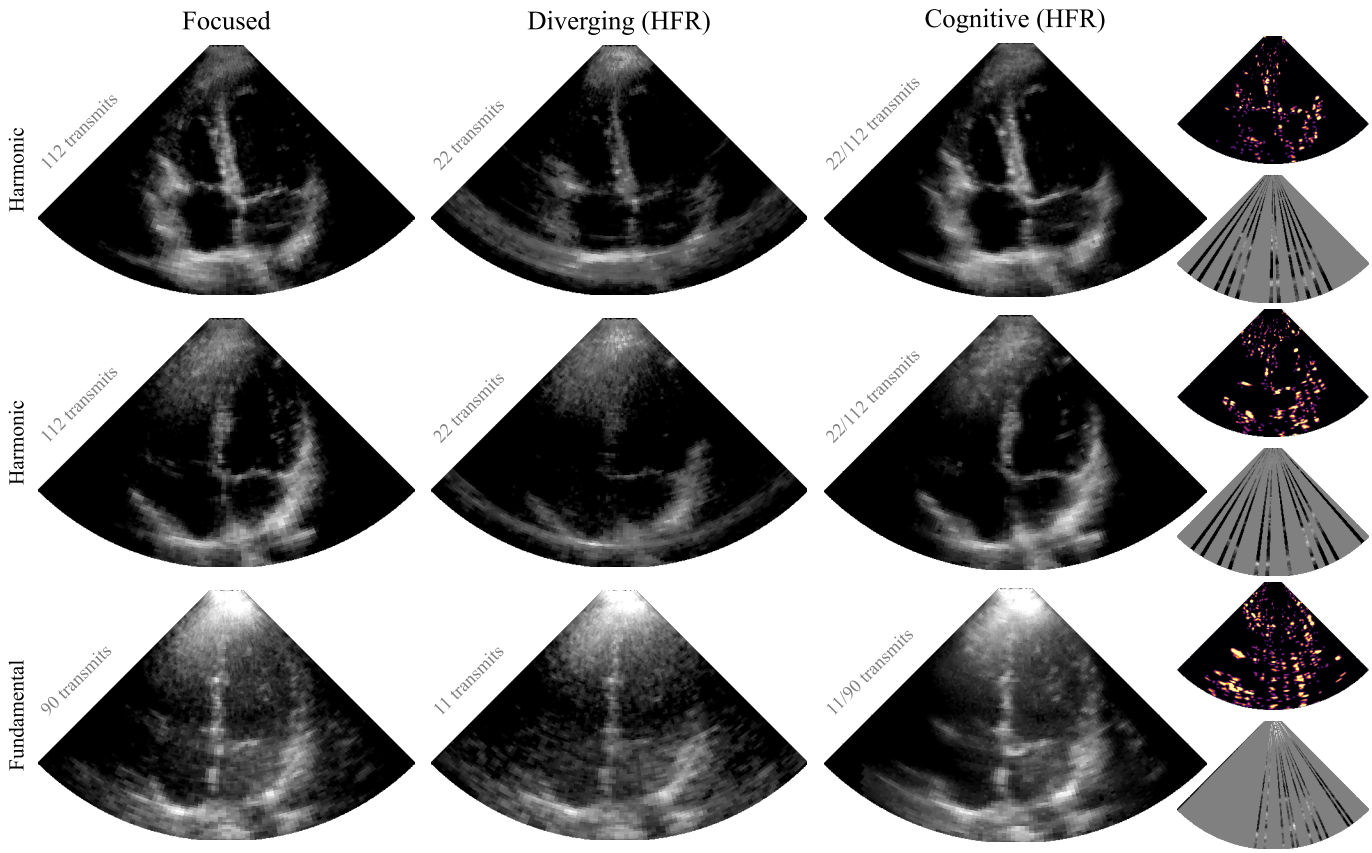

Fig. 8. Qualitative results on the in-house echocardiograms. The top rows show frames from an echocardiogram using harmonic imaging; the third shows another subject using fundamental imaging. During the scan, both focused and diverging waves were acquired. Each row shows the fully observed focused acquisition, high frame-rate (HFR) acquisition using diverging waves, and HFR using focused cognitive transmits. On the right, the acquisitions, entropy and reconstructions are shown for the subsampled focused transmits. All images were 112×112 pixels before scan conversion and histogram matched to the fully observed focused data.

TABLE I
INFERENCE SPEED OPTIMIZATIONS COMPUTED ON THE NVIDIA L40S FOR 112×112 PIXELS AND $N_p = 2$.

Optimization	Frame Time [ms]	Frame Freq. [Hz]
Base (*500 steps*)	858.7	1.16
+ SeqDiff [60] (*25 steps*)	76.0	13.16
+ Just-in-time compilation	36.0	27.81
+ Mixed precision (*float16*)	17.3	**57.65**
Physical *subsampled* acquisition (28 transmits)	5.46	183.2
Physical *full* acquisition (112 transmits)	21.82	45.83

out of 112 lines still reaches a DICE of 0.91 on average.

4) Inference speed: As mentioned before, we employ SeqDiff [60], which not only improves temporal consistency of posterior samples, but it also massively reduces the required number of neural function evaluations to generate sequential signals. To improve inference speed 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 to the function encapsulating both the "Perception Step" and "Action Step" using the JAX library [66]. Finally, the diffusion model, trained in 32-bit floating point precision, can be run in mixed precision using 16 bits. All these optimizations combined yield a frame rate of 58 Hz on the Nvidia L40S. We measured GPU memory usage to be approximately 720 MB. The inference speed of the reconstruction model is independent of the number of transmits. The physical time it takes to acquire 112 transmits is 21.82 ms, meaning the maximal achievable frame rate is 46 Hz. Given that our model can run faster, this would enable increasing the frame rate to 58 Hz through subsampling 89/112 focused transmits. We have also listed the frame rate for 28 transmits, indicating that improvements in inference speed could further increase frame rates.

B. In-house echocardiograms

The in-house dataset (Figure 8) was recorded on a Verasonics Vantage 256 with an S5-1 Philips transducer with a center frequency of 3.125 MHz. The scans were performed by an experienced sonographer. The dataset consists of six subjects, who provided informed consent at the time of data collection,

and the study was approved by the local Institutional Review Board. We included two types of acquisitions: fundamental mode imaging was used for three subjects, and harmonic imaging with pulse inversion was used for the other three subjects. The fundamental acquisition consists of 90 focused transmits, which were interleaved with 11 diverging transmits for comparison, at a pulse-repetition frequency (PRF) of 4.32 kHz, and the transmit frequency is equal to the probe's center frequency. The harmonic acquisition consists of 56 $\times$ 2 focused transmits, one for each polarity, interleaved with 11 $\times$ 2 diverging transmits, at a PRF of 5.07 kHz. The transmit frequency is 1.9531 MHz, and the demodulation frequency is double the transmit frequency, to capture the harmonic signal. The recorded sequences consist of 70 to 100 frames in the apical four-chamber view. The dynamic range is matched to the EchoNet-Dynamic dataset using percentile-based clipping. Subsampling is implemented by taking a subset of transmit events from the channel data and independently beamforming only those transmit events to lines. The width of the beamformed lines is 1 or 2 pixels for the fundamental and harmonic data, respectively. Both modalities are beamformed to 2 pixels per wavelength axially. Subsequently, the lines are downsampled to 112 pixels, using linear interpolation and a triangular filter for anti-aliasing, to match the training dataset, giving us $\boldsymbol{y}_t$. The pretrained (EchoNet-Dynamic) prior was used to generate reconstructions $\tilde{\boldsymbol{x}}_t$.

1) Contrast: To demonstrate the effectiveness of our method, we compute the gCNR metric [67] between the ventricle and the myocardium as well as between the ventricle and the valve. These regions of interest are manually annotated, using the fully observed image (all focused transmits), by one of the authors. The same annotations are used for all the reconstruction methods. The gCNR is calculated relative to the fully sampled focused acquisition, which allows us to compare CASL to diverging waves for the same number of transmits.

Figure 9 shows the gCNR over time between the valve and the ventricle for two subjects. It can be seen that CASL almost always outperforms diverging waves. CASL generally has slightly higher gCNR compared to focused transmits, while for diverging waves it is slightly lower. In Figure 10 we show the distribution of gCNR over the frames between the myocardium and ventricle for six subjects. This highlights again that CASL outperforms diverging waves for all subjects, and shows fewer outliers.

2) Reconstruction quality: Figure 11 shows the reconstruction quality in terms of PSNR and LPIPS [65] for three subsampling rates. CASL outperforms random and equispaced undersampling for all subsampling rates and every subject in the dataset. The qualitative results are shown in Figure 8. Here, we see the focused transmits, diverging transmits, and reconstructions using CASL for two subjects, where one uses the fundamental mode and the other uses harmonic imaging. 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.

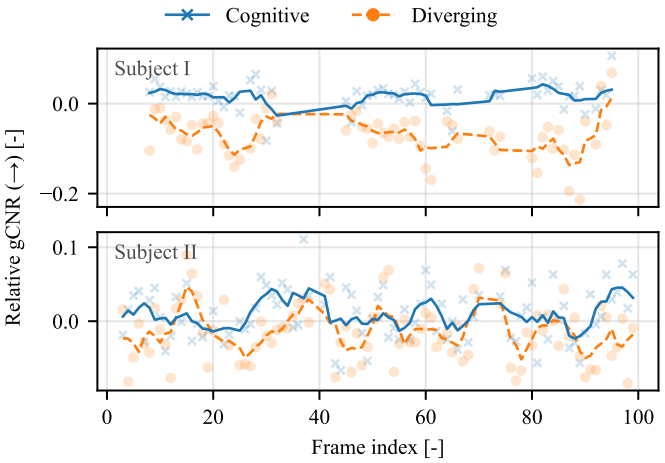

Fig. 9. Generalized contrast-to-noise ratio (gCNR) for two subjects over time relative to a focused acquisition of 90 transmits. The gCNR was measured between the **valve** and the **ventricle**. Both cognitive and diverging use 11 transmits.

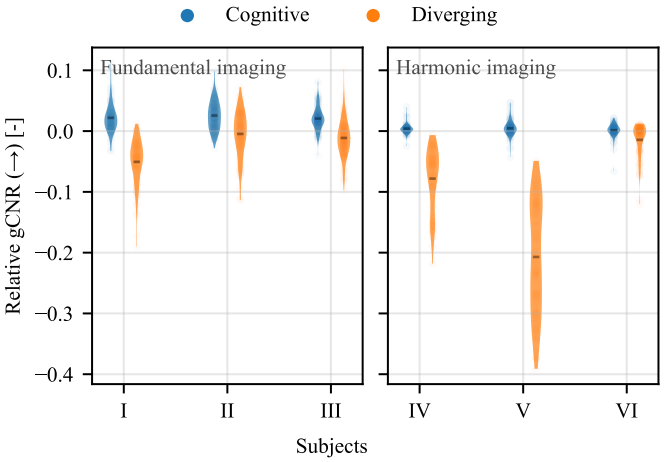

Fig. 10. Generalized contrast-to-noise ratio (gCNR) for six subjects relative to a focused acquisition. The gCNR was measured between the **myocardium** and the **ventricle**. Both cognitive and diverging use the same number of transmits, 11 for fundamental and 22 for harmonic. The figure shows a distribution over the frames and includes the mean as a gray line. Cognitive is found to be significantly better ($\boldsymbol{p < 0.05}$) than diverging using the Wilcoxon signed-rank test.

C. 3D echocardiograms

In this section, we apply CASL to 3D echocardiography. Following Stevens *et al.* [68], 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 III. Our prior is therefore approximating the joint distribution $p(\mathbf{X})$ where $\mathbf{X} \in \mathbb{R}^{N_{\text{ax}} \times N_{\text{el}} \times W}$. For this task, a new DM was trained, to better match the resolution and distribution of (ax, el) slices, which depart from the apical four-chamber views used in the

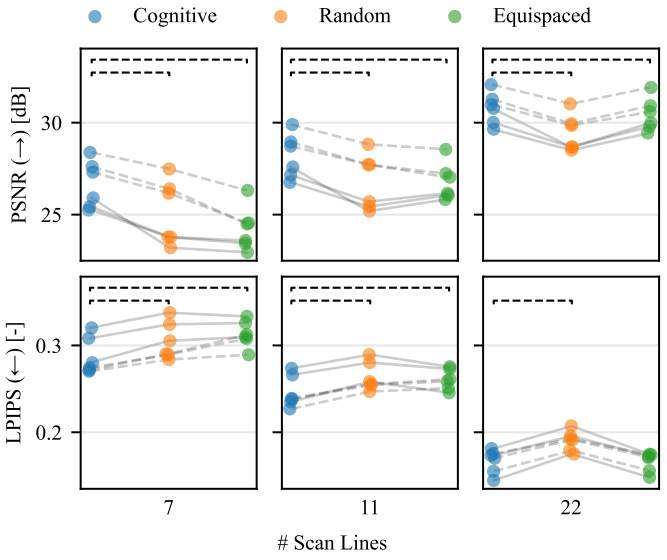

Fig. 11. Reconstruction performance for six subjects in terms of PSNR and LPIPS [65] as a function of the number of scanned lines for various action selection policies. Every line represents a subject, where a dashed line uses harmonic imaging, solid lines use fundamental imaging, and brackets indicate $\boldsymbol{p < 0.05}$ using the Wilcoxon signed-rank test.

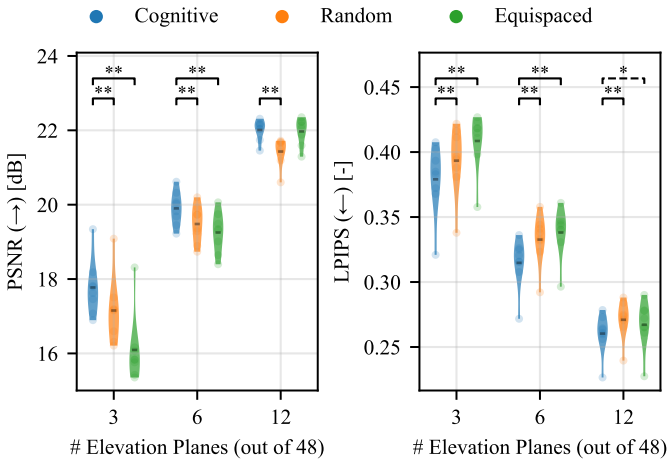

Fig. 12. Reconstruction performance for the 3D dataset in terms of PSNR and LPIPS [65] as a function of the number of scanned lines for various action selection policies. The figure shows a distribution over the data samples and includes the mean as a gray line ($^{**}\boldsymbol{p < 0.01}$, $^{*}\boldsymbol{p < 0.05}$ using the Wilcoxon signed-rank test).

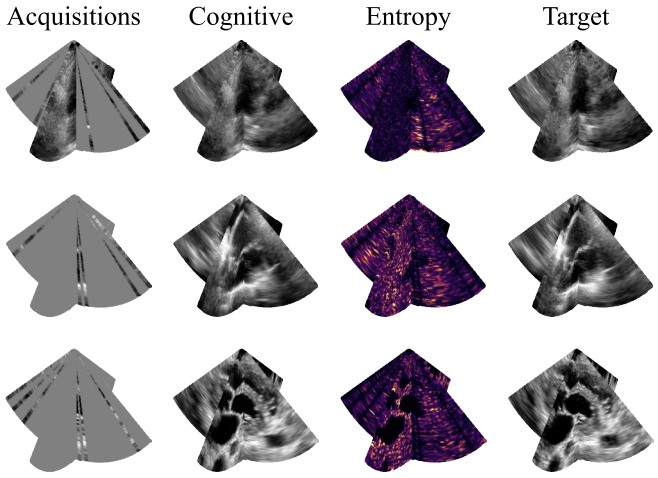

Fig. 13. Qualitative results on the 3D dataset. The figure shows the acquisitions and reconstructions for 6 / 48 elevation planes compared to the target. Additionally, it shows the posterior entropy, which drives action selection.

previous experiments. The DM was trained on samples of size $N_{\text{ax}} = 400$, $N_{\text{el}} = 48$, and $W = 3$. The dataset consists of 100 *in-vivo* B-mode volume sequences across 16 patients, acquired using a Philips EPIQ scanner with an X5-1c matrix probe (Philips Research, North America). The transducer has a frequency range of 1 MHz to 5 MHz. A set of 8 volume sequences across 3 patients is held out for testing. The test volumes contained 400 axial samples, 48 elevation planes, and varying numbers of azimuthal angles, ranging from 56 to 84. For posterior sampling, a guidance weight of $\gamma = 3$ was used, with $N_p = 2$, $\tau_{\text{max}} = 500$, 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 CASL against those created with baseline sampling strategies, with PSNR and LPIPS [65] results plotted in Figure 12. The test volume sequences also varied in length, from 6 to 39 volumes long, and so the metrics were averaged per sequence to avoid bias. Across the subsampling rates, it is clear that employing CASL results in more faithful reconstructions, particularly with more aggressive subsampling. In Figure 13, we provide qualitative examples in the form of bi-plane plots of volume reconstructions from 6/48 elevation planes, at the 4^{th} frame in each sequence.

V. Discussion

It is clear throughout the results provided in Section IV that CASL outperforms the equispaced and random baseline strategies. The degree of improvement varies across the experiments. In Section IV-A, our results on the 2D EchoNet-Dynamic dataset show significant benefits to using CASL, achieving an improved trade-off between reconstruction quality and subsampling rate. In our experiments on 3D data in Section IV-C, we also find that CASL outperforms fixed sampling strategies across a range of sampling budgets, achieving a better trade-off between volume rate and reconstruction accuracy than equispaced and random strategies.

It is noteworthy that CASL achieves strong performance on the in-house echocardiograms using measurements from both fundamental and harmonic imaging modes, while using a generative prior trained on EchoNet-Dynamic. This showcases

an instance of *generalization*, wherein the model recovers a target sample from a new, unseen distribution. Indeed, Jia *et al.* have shown that, if enough measurements have been acquired, DPS can recover target images from distributions that are vastly different from the prior [69]. In this case, the in-house echocardiograms have in common with EchoNet-Dynamic that they use the apical four-chamber view. To explore the limits of generalization with CASL, we have included additional results using measurements from parasternal long-axis echocardiograms, and a static CIRS phantom on the GitHub repository.

A. Future work

Future work towards improving performance in the 2D regime might develop approaches to generative modeling that can support longer context windows, more efficient inference, and higher spatial resolution. For example, we now use a data prior trained on the EchoNet-Dynamic dataset, which consists of 112×112 pixels, for our in-house dataset, which has the potential to be beamformed at a much higher resolution. The algorithm could also be improved by lowering reconstruction error. While it was shown in Section IV that the reconstruction error did not significantly hamper performance on a left ventricle segmentation task, the impact of the specific character of reconstruction errors introduced by CASL on further downstream tasks could be interesting in future work.

Our encouraging 3D results highlight opportunities for further enhancement. 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, focusing in both the elevation and azimuthal directions, as compared to elevation plane selection, 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 cognitive subsampling 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 $p(\boldsymbol{y} \mid \boldsymbol{x}, A^\ell)$ becomes more challenging and no longer corresponds to an inpainting task. The same holds for other types of spatially overlapping reconstructions, such as is common for diverging waves or plane waves. Future work could explore how to better leverage the image quality benefits of RTBF.

In our experiments we retrospectively subsample the selected lines. To fully leverage CASL, the algorithm must operate in real-time with the frame acquisition, which could improve the results due to higher temporal correlation. We found that with GPU hardware from 2023, our algorithm could increase frame rate from 46 Hz to 58 Hz through subsampling to 89/112 lines. In Table I, we computed that 28 scan lines result in a physical frame acquisition time of 5.46 ms. Therefore, to achieve higher subsampling rates, and higher frame rates, the algorithm still requires an approximate $3\times$ speed-up. A final avenue for future work could involve exploring clinical advantages of the additional frame-rate offered by CASL.

VI. Conclusion

We proposed a patient-adaptive focused transmit and receive scheme that reduces the number of acquisitions needed for a high-quality echocardiogram by actively selecting the most informative measurements. Our method leverages posterior sampling with a temporal diffusion model and acquires new measurements where the approximated posterior shows the most entropy. We have shown that our method outperforms baselines on the 2D EchoNet-Dynamic dataset, a 2D in-house dataset consisting of both fundamental and harmonic imaging modes, and a 3D cardiac dataset, especially in cases with very few focused transmits. We have shown that using cognitive ultrasound with focused transmits improves gCNR compared to diverging waves with the same number of transmits. For 112×112 images, the method can be run in real-time at 58 Hz on GPU accelerators from 2023.

Acknowledgment

The authors would like to thank Danique de Bruijn for performing the ultrasound scans.

References

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

[2] R. M. Lang *et al.*, "Recommendations for cardiac chamber quantification by echocardiography in adults: An update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging," *J. Am. Soc. Echocardiogr. Off. Publ. Am. Soc. Echocardiogr.*, vol. 28, no. 1, 1–39.e14, Jan. 2015.

[3] V. Kakkad, M. LeFevre, K. Roy Choudhury, J. Kisslo, and G. E. Trahey, "Effect of Transmit Beamforming on Clutter Levels in Transthoracic Echocardiography," *Ultrason. Imaging*, vol. 40, no. 4, pp. 215–231, Jul. 1, 2018.

[4] 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.

[5] 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.

[6] K. Hawkins, J. S. Henry, and R. A. Krasuski, "ORIGINAL INVESTIGATIONS: Tissue Harmonic Imaging in Echocardiography: Better Valve Imaging, But at What Cost?" *Echocardiography*, vol. 25, no. 2, pp. 119–123, 2008.

[7] M. MONAGHAN, "Second harmonic imaging: A new tune for an old fiddle?" *Heart*, vol. 83, no. 2, pp. 131–132, Feb. 2000.

[8] G. M. Matte, P. L. M. J. Van Neer, M. G. Danilouchkine, J. Huijssen, M. D. Verweij, and N. D. Jong, "Optimization of a phased-array transducer for multiple harmonic imaging in medical applications: Frequency and topology," *IEEE Trans. on Ultrason. Ferroelectr. Freq. Control*, vol. 58, no. 3, pp. 533–546, Mar. 2011.

[9] G. Schmitz and S. Dencks, "Ultrasound Imaging," in *Molecular Imaging in Oncology*, O. Schober, F. Kiessling, and J. Debus, Eds., Cham: Springer International Publishing, 2020, pp. 135–154.

[10] G. Montaldo, M. Tanter, J. Bercoff, N. Benech, and M. Fink, "Coherent plane-wave compounding for very high frame rate ultrasonography and transient elastography," *IEEE Trans. on Ultrason. Ferroelectr. Freq. Control*, vol. 56, no. 3, pp. 489–506, Mar. 2009.

[11] J. Grondin, V. Sayseng, and E. E. Konofagou, "Cardiac Strain Imaging with Coherent Compounding of Diverging Waves," *IEEE Trans. on Ultrason. Ferroelectr. Freq. Control*, vol. 64, no. 8, pp. 1212–1222, Aug. 2017.

[12] L. Nie, D. M. J. Cowell, T. M. Carpenter, J. R. Mclaughlan, A. A. Çubukçu, and S. Freear, "High-Frame-Rate Contrast-Enhanced Echocardiography Using Diverging Waves: 2-D Motion Estimation and Compensation," *IEEE Trans. on Ultrason. Ferroelectr. Freq. Control*, vol. 66, no. 2, pp. 359–371, Feb. 2019.

[13] L. Nie, D. M. J. Cowell, T. M. Carpenter, J. R. McLaughlan, A. A. Çubukçu, and S. Freear, "Motion Compensation for High-Frame-Rate Contrast-Enhanced Echocardiography Using Diverging Waves: Image Registration Versus Correlation-Based Method," in *2019 IEEE Int. Ultrason. Symp. (IUS)*, Oct. 2019, pp. 380–383.

[14] 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.

[15] M. Negoita *et al.*, "Frame rate required for speckle tracking echocardiography: A quantitative clinical study with open-source, vendor-independent software," *Int. J. Cardiol.*, vol. 218, pp. 31–36, 2016.

[16] V. Mor-Avi *et al.*, "Current and Evolving Echocardiographic Techniques for the Quantitative Evaluation of Cardiac Mechanics: ASE/EAE Consensus Statement on Methodology and Indications: Endorsed by the Japanese Society of Echocardiography," *J. Am. Soc. Echocardiogr.*, vol. 24, no. 3, pp. 277–313, Mar. 2011.

[17] O. T. Von Ramm and S. W. Smith, "Beam Steering with Linear Arrays," *IEEE Trans. on Biomed. Eng.*, vol. BME-30, no. 8, pp. 438–452, Aug. 1983.

[18] C.-C. Shen, Y.-H. Chou, and P.-C. Li, "Pulse Inversion Techniques in Ultrasonic Nonlinear Imaging," *J. Med. Ultrasound*, vol. 13, no. 1, pp. 3–17, Jan. 2005.

[19] Z. Zhong *et al.*, "Visualization of human aortic valve dynamics using magnetic resonance imaging with sub-millisecond temporal resolution," *J. Magn. Reson. Imaging*, vol. 54, no. 4, pp. 1246–1254, 2021.

[20] 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.

[21] 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.

[22] 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.

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

[24] L. Demi, "Practical Guide to Ultrasound Beam Forming: Beam Pattern and Image Reconstruction Analysis," *Appl. Sci.*, vol. 8, no. 9, p. 1544, 9 Sep. 2018.

[25] A. Ilovitsh, T. Ilovitsh, J. Foiret, D. N. Stephens, and K. W. Ferrara, "Simultaneous Axial Multifocal Imaging using a Single Acoustical Transmission: A Practical Implementation," *IEEE Trans. on Ultrason. Ferroelectr. Freq. Control*, vol. 66, no. 2, pp. 273–284, Feb. 2019.

[26] O. Çakiroğlu, E. Pérez, F. Roemer, and M. Schiffner, "Autoencoder-based learning of transmission parameters in fast pulse-echo ultrasound imaging employing sparse recovery," in *2023 IEEE 9th Int. Workshop on Comput. Adv. Multi-Sensor Adapt. Process. (CAMSAP)*, IEEE, 2023, pp. 516–520.

[27] D. L. Donoho, "Compressed sensing," *IEEE Trans. on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[28] T. Chernyakova and Y. C. Eldar, "Fourier-domain beamforming: The path to compressed ultrasound imaging," *IEEE Trans. on Ultrason. Ferroelectr. Freq. Control*, vol. 61, no. 8, pp. 1252–1267, 2014.

[29] 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.

[30] 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.

[31] A. Besson, D. Perdios, M. Arditi, Y. Wiaux, and J.-P. Thiran, "Compressive multiplexing of ultrasound signals," in *2018 IEEE Int. Ultrason. Symp. (IUS)*, IEEE, 2018, pp. 1–4.

[32] A. Besson, D. Perdios, Y. Wiaux, and J.-P. Thiran, "Joint sparsity with partially known support and application to ultrasound imaging," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 84–88, 2018.

[33] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] O. Lorintiu, H. Liebgott, M. Alessandrini, O. Bernard, and D. Friboulet, "Compressed sensing reconstruction of 3d ultrasound data using dictionary learning and line-wise subsampling," *IEEE Trans. on Med. Imaging*, vol. 34, no. 12, pp. 2467–2477, 2015.

[39] 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.

[40] 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.

[41] 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.

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

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

[44] C. Wang, K. Shang, H. Zhang, S. Zhao, D. Liang, and S. K. Zhou, "Active ct reconstruction with a learned sampling policy," in *Proc. 31st ACM Int. Conf. on Multimed.*, 2023, pp. 7226–7235.

[45] 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.

[46] 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.

[47] Y. Song, L. Shen, L. Xing, and S. Ermon, "Solving inverse problems in medical imaging with score-based generative models," in *The Tenth Int. Conf. on Learn. Represent. ICLR 2022, Virtual Event, April 25-29, 2022*, OpenReview.net, 2022.

[48] 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.

[49] 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.

[50] 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.

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

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

[53] 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.

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

[55] 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.

[56] 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.

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

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

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

[60] 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.

[61] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proc. IEEE Conf. on Comput. Vis. Pattern Recognit.*, 2018, pp. 6228–6237.

[62] 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.

[63] T. S. Stevens *et al.*, "Zea: A toolbox for cognitive ultrasound imaging," *J. Open Source Softw.*, vol. 11, no. 121, p. 9881, 2026.

[64] 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.

[65] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *2018 IEEE/CVF Conf. on Comput. Vis. Pattern Recognit.*, Salt Lake City, UT: IEEE, Jun. 2018, pp. 586–595.

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

[67] A. Rodriguez-Molares *et al.*, "The generalized contrast-to-noise ratio: A formal definition for lesion detectability," *IEEE Trans. on Ultrason. Ferroelectr. Freq. Control*, vol. 67, no. 4, pp. 745–759, Apr. 2020.

[68] T. S. Stevens, O. Nolan, O. Somphone, J.-L. Robert, and R. J. Van Sloun, "High volume rate 3d ultrasound reconstruction with diffusion models," *IEEE Trans. on Med. Imaging*, pp. 1–1, 2025.

[69] J. Jia, W. Yuan, S. Liu, L. Shen, and G. Wang, "Weak diffusion priors can still achieve strong inverse-problem performance," *arXiv preprint arXiv:2601.22443*, 2026.