Sequential Posterior Sampling with Diffusion Models

Tristan S.W. Stevens*, Oisín Nolan*, Jean-Luc Robert†, Ruud J.G. van Sloun*

*Dept. of Electrical Engineering, Eindhoven University of Technology, The Netherlands
†Philips Research North America, Cambridge MA, USA

***Abstract*—Diffusion models have quickly risen in popularity for their ability to model complex distributions and perform effective posterior sampling. Unfortunately, the iterative nature of these generative models makes them computationally expensive and unsuitable for real-time sequential inverse problems such as ultrasound imaging. Considering the strong temporal structure across sequences of frames, we propose a novel approach that models the transition dynamics to improve the efficiency of sequential diffusion posterior sampling in conditional image synthesis. Through modeling sequence data using a video vision transformer (ViViT) transition model based on previous diffusion outputs, we can initialize the reverse diffusion trajectory at a lower noise scale, greatly reducing the number of iterations required for convergence. We demonstrate the effectiveness of our approach on a real-world dataset of high frame rate cardiac ultrasound images and show that it achieves the same performance as a full diffusion trajectory while accelerating inference 25×, enabling real-time posterior sampling. Furthermore, we show that the addition of a transition model improves the PSNR up to 8% in cases with severe motion. Our method opens up new possibilities for real-time applications of diffusion models in imaging and other domains requiring real-time inference.**

***Index Terms*—temporal diffusion prior, generative models, sequential data, cardiac ultrasound, posterior sampling**

I. Introduction

Deep generative models are celebrated for their ability to model complex distributions. Their use in inverse problem solving has unlocked new applications involving high-dimensional data. Diffusion Models (DMs) are particularly attractive generative models due to their interpretable denoising score matching objective and stable sampling procedure. Despite these benefits, the iterative nature of sampling from prior and posterior distributions with diffusion models inhibits their use in demanding real-time imaging applications with high data-rates such as cardiac ultrasound [1], [2] or automotive radar [3], [4].

There have been several works on accelerating DMs. These can be roughly categorized in two lines of research. On the training end, [5] proposes a *progressive distillation* method that augments the training of the DMs with a student-teacher model setup. In doing this, they are able to drastically reduce the number of sampling steps. Some methods aim to execute the the diffusion process in a reduced space to accelerate the diffusion process. While [6] restricts diffusion through projections onto subspaces, [7] and [8] run the diffusion in the latent space. On the other side of the spectrum, the sampling procedure itself can be altered. Inspired by momentum methods in sampling, [9] introduces a momentum sampler for DMs, which leads to increased sample quality with fewer function evaluations. More related to this work is a sampling strategy known as *Come-Closer-Diffuse-Faster* (CCDF) [10], which leverages a neural network based estimate of the posterior mean to reduce the number of reverse diffusion steps needed. Nonetheless, CCDF and the other aforementioned methods do not exploit the temporal structure across frames in sequential data which we demonstrate improves the solvability of inverse problems.

Video diffusion models extent on previous works by training a diffusion prior jointly on a sequence of frames [11]. While they have been extensively explored for tasks such as *text-to-video* [12] and *image-to-video* [13] generation, there has been limited research on their application to video reconstruction tasks. Some works have investigated the use of DMs for time-series; [14], for example, proposes a conditional diffusion model for time series forecasting. However, these works do not consider the temporal structure across frames for accelerating the sampling process, rendering them too slow for real-time inference.

In this work, we propose a novel autoregressive method for initializing successive diffusion trajectories for reconstruction of sequence data. We provide two flavors named *SeqDiff* and *SeqDiff+* which both leverage the temporal correlation across frames, by using the diffusion model output of previous frames as a starting point for the current posterior sampling procedure. SeqDiff straightforwardly initializes with the previous frame, which we show is often reasonable given high frame rates. Expanding on this idea, SeqDiff+ specifically models the transition between subsequent frames using a *Video Vision Transformer* (ViViT) [15] for a more accurate initialization, mitigating the effect of severe motion across frames.

To evaluate our method, we turn to echocardiography, which is the imaging of the heart using medical ultrasound. The real-time nature and high data-rates resulting from this sensory data encapsulate the challenges targeted by the proposed method. DMs have been effectively applied to cardiac ultrasound, from removing multipath scattering (dehazing) [2] to segmentation [1] and beyond. However, accurate and fast image reconstruction using DMs remains a challenge.

Our main contributions can be summarized as follows:

- We propose autoregressive tracking of posterior samples across the noise manifolds in diffusion models to accelerate reconstruction of sequential data.
- We provide two variants, SeqDiff and SeqDiff+, both of which rely on previous diffusion posterior estimates for initialization. SeqDiff+ further leverages a Video Vision Transformer to model the transitions between frames.

- We evaluate our method on compressed sensing echocardiography, showing that our method improves image quality while accelerating the sampling process.

The remainder of this paper is organized as follows. In Section II we provide background on both posterior sampling with DMs as well as sequence modeling. In Section III we proceed with introduction of our methods, which are subsequently evaluated and concluded in sections IV and V, respectively.

II. Background

A. Diffusion Models

Diffusion Models (DMs) are a class of probabilistic generative models that learn the reversal of a forward corruption process, which add progressively increasing levels of Gaussian noise until the data $\mathbf{x}_0 \equiv \mathbf{x} \sim p(\mathbf{x})$ is transformed into a base distribution $\mathbf{x}_\mathcal{T} \sim \mathcal{N}(0, \mathbf{I})$. The continuous forward process $\mathbf{x}_0 \to \mathbf{x}_\tau \to \mathbf{x}_\mathcal{T}$, with diffusion time $\tau \in [0, \mathcal{T}]$ can be formally described by a variance preserving stochastic differential equation (VP-SDE) [16] $\mathrm{d}\mathbf{x} = -\frac{1}{2}\beta(\tau)\mathbf{x}\mathrm{d}\tau + \sqrt{\beta(\tau)}\mathrm{d}\mathbf{w}$, where $\beta(\tau)$ is the noise schedule, and $\mathbf{w}$ a standard Wiener process. Diffused samples from $p(\mathbf{x}_\tau|\mathbf{x}_0) = \mathcal{N}(\alpha_\tau \mathbf{x}_0, \sigma_\tau^2 \mathbf{I})$ can be directly generated by the following parameterization:

$$\mathbf{x}_\tau = \alpha_\tau \mathbf{x}_0 + \sigma_\tau \epsilon, \quad \epsilon \in \mathcal{N}(0, \mathbf{I}), \tag{1}$$

where $\sigma_\tau = 1 - e^{-\int_0^\tau \beta(s)\mathrm{d}s}$ and $\alpha_\tau = \sqrt{1-\sigma_\tau^2}$ are the noise and signal rates, respectively. The objective of generative models is to generate samples from the distribution of interest given samples from some tractable distribution. Accordingly, a corresponding reverse-time SDE can be constructed to achieve this:

$$\mathrm{d}\mathbf{x} = \left[-\frac{1}{2}\beta(\tau)\mathbf{x} - \beta(\tau)\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau)\right]\mathrm{d}\tau + \sqrt{\beta(\tau)}\mathrm{d}\bar{\mathbf{w}}, \tag{2}$$

where $\mathrm{d}\tau$ and $\mathrm{d}\bar{\mathbf{w}}$ are now processes running backwards in diffusion time. From this reverse SDE the gradient of the log-likelihood of the data arises $\nabla_\mathbf{x} \log p_t(\mathbf{x})$, also known as the *score function* which provides information on how to adjust $\mathbf{x}_\tau$ to move it towards $\mathbf{x}_0$ and can be modeled using neural network parameters θ leading to the following approximation: $s_\theta(\mathbf{x}_\tau, \tau) \approx \nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau)$. As shown in [17], the score model s_θ can be learned with the denoising score matching objective

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}_0 \sim p(\mathbf{x}), \tau \sim \mathcal{U}[0,\mathcal{T}]}\left[\|s_\theta(\mathbf{x}_\tau, \tau) - \nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau|\mathbf{x}_0)\|_2^2\right], \tag{3}$$

which essentially trains a conditional denoising network at each diffusion timestep τ. Finally, discretization of continuous process (2) into N equispaced diffusion steps is required to numerically approximate the reverse diffusion process and sample from the target distribution.

B. Posterior Sampling

Shifting our focus to inverse problems solving, which seeks to retrieve underlying signals $\mathbf{x}$ from corrupted observations $\mathbf{y}$, we can define a general linear forward model as follows:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}, \quad \mathbf{y}, \mathbf{n} \in \mathbb{R}^m, \mathbf{x} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{m \times n}. \tag{4}$$

DMs can be extended to perform posterior sampling $p(\mathbf{x}_0|\mathbf{y})$, through substitution of a conditional score into (2), which can be factorized into the pretrained score model and a noise perturbed likelihood score through Bayes' rule: $\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_0|\mathbf{y}) \approx s_\theta(\mathbf{x}_\tau, \tau) + \nabla_{\mathbf{x}_\tau} \log p(\mathbf{y}|\mathbf{x}_\tau)$. The intractability of the latter term has lead to several approaches to approximate it [19], [20]. Among the methods is Diffusion Posterior Sampling (DPS) [18], which approximates the troubling $p(\mathbf{x}_0|\mathbf{x}_\tau)$, which leads to tractability of $p(\mathbf{y}|\mathbf{x}_\tau)$, as follows:

$$p(\mathbf{x}_0|\mathbf{x}_\tau) \approx \mathbb{E}[\mathbf{x}_0|\mathbf{x}_\tau] \approx \frac{1}{\alpha_\tau}(\mathbf{x}_\tau + \sigma_\tau^2 s_\theta(\mathbf{x}_\tau, \tau)) \tag{5}$$

where the first approximation is substitution of the posterior mean for $\mathbf{x}_0$, and the second approximation the learned score model for the actual unconditional score function.

C. Sequential inverse problems

In this work, we seek to address sequential inverse problems, also known as *dynamic inverse problems* [21], which involve reconstructing from a sequence of time-dependent measurements $\mathbf{y}^t = \mathbf{A}^t\mathbf{x}^t + \mathbf{n}^t$ with a clear dependency between $\mathbf{x}^t$ and $\mathbf{x}^{t-1}$. To capture the intricate dynamics of temporal data, we look to sequence modeling which has become a fundamental task in applications such as speech recognition, natural language processing, and video analysis. We are interested in predicting future frames given past observations:

$$p(\mathbf{x}^{t+1} \mid \mathbf{x}^t, \mathbf{x}^{t-1}, ..., \mathbf{x}^{t-K}), \tag{6}$$

where K is the context window size. In the context of cardiac ultrasound this would translate to predicting a future frame given K past frames. Traditional approaches to modeling sequences include hidden Markov models (HMMs), recurrent neural networks (RNNs), amongst which convolutional LSTMs (ConvLSTMs) [22] which have proven to work well for spatio-temporal data. More recently, transformer models have excelled especially in natural language processing tasks through self-attention mechanisms that capture long-range dependencies. The Video Vision Transformer (ViViT) [15] extends this capability to video data by treating a stack of subsequent frames. Specifically, ViViTs extract non-overlapping, spatio-temporal tubes (3D patches), also known as tubelet embeddings, to tokenize the input video and accordingly process using multi-headed self-attention blocks.

III. Methods

The temporal correlation across subsequent frames can be heavily exploited to accelerate sequential posterior sampling $p(\mathbf{x}^t|\mathbf{y}^t, \mathbf{x}^{t-K:t-1})$ using DMs. We propose two techniques to initialize the reverse diffusion process corresponding to the current frame based on past observations in an efficient manner. In other words, given the diffusion posterior samples $\mathbf{x}_0$ of past frames $\left\{\mathbf{x}_0^t, \mathbf{x}_0^{t-1}, ..., \mathbf{x}_0^{t-K}\right\}$ we would like to estimate $p(\mathbf{x}^{t+1} \mid \mathbf{x}_0^{t-K:t})$ such that the number of diffusion steps necessary is minimized. Since this is again a complex distribution, we instead estimate $p(\mathbf{x}_{\tau'}^{t+1} \mid \mathbf{x}_0^{t-K:t})$, and assume it follows a tractable Gaussian with diagonal covariance. The

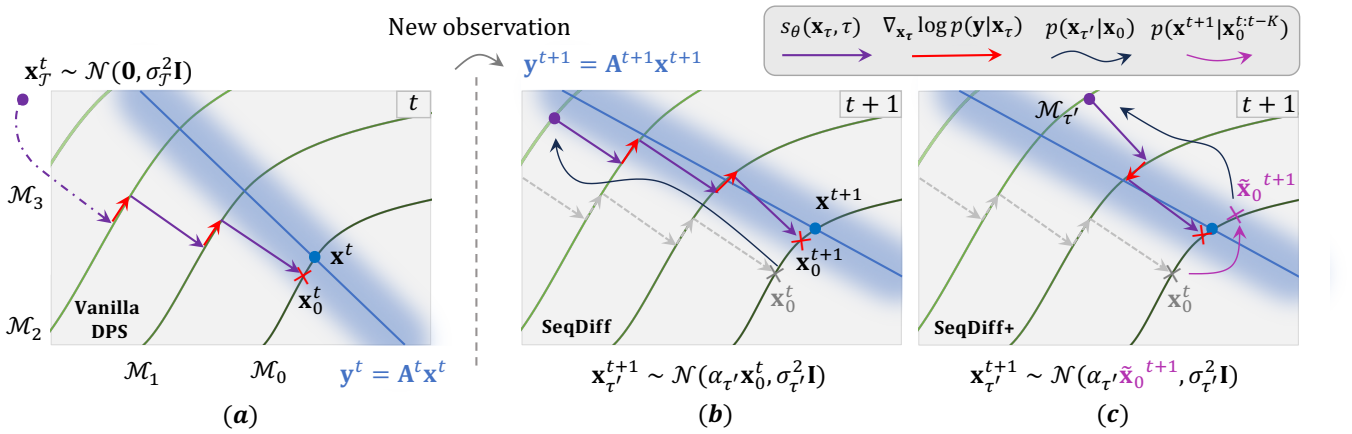

Fig. 1: Geometric representation of the reverse diffusion process and corresponding manifolds $\mathcal{M}_\tau$ for each diffusion timestep τ. In **(a)** a standard conditional reverse diffusion trajectory starting from a Gaussian sample $\mathbf{x}_\mathcal{T} \sim \mathcal{N}$ is shown with DPS as guidance rule [18]. For initialization of the next frame $t+1$, we propose two different methods SeqDiff and SeqDiff+, depicted in **(b)** and **(c)** respectively. In the first option we initialize the trajectory from a noised version of the Tweedie estimate of the previous frame, $p(\mathbf{x}_{\tau'}^{t+1}|\mathbf{x}_0^t)$ with $\tau' \ll \mathcal{T}$. The second option improves upon this by predicting the next frame with $\tilde{\mathbf{x}}_0^{t+1} \approx f(\cdot)$, accounting for any motion between frames. This leads to the initialization $p(\mathbf{x}_{\tau'}^{t+1}|\tilde{\mathbf{x}}_0^{t+1})$, with $\tau'_{\text{SeqDiff+}} < \tau'_{\text{SeqDiff}}$.

	Initialization $p(\mathbf{x}_{\tau'}^{t+1} \mid \mathbf{x}_0^{t+1})$ $\mathcal{N}(\quad \mu \quad , \quad \sigma \quad)$		**Sequence modeling**
Vanilla DPS	$\mathbf{0}$	$\sigma_\mathcal{T}^2$	✗
CCDF	$\alpha_{\tau'} g(\mathbf{y}^{t+1})$	$\sigma_{\tau'}^2$	✗
SeqDiff	$\alpha_{\tau'}\mathbf{x}_0^t$	$\sigma_{\tau'}^2$	~
SeqDiff+	$\alpha_{\tau'} f_\theta(\mathbf{x}_0^{t-K:t})$	$\sigma_{\tau'}^2$	✓

TABLE I: Comparison of the different initialization methods for accelerating reverse diffusion trajectories.

challenge is to estimate the parameters of this distribution, as well as the diffusion time point τ' for which the Gaussian approximation is accurate. We define the initialization diffusion scale as τ' which lies somewhere on the diffusion timeline $0 < \tau' \ll \mathcal{T}$. Rather than starting each diffusion trajectory from scratch at $\tau = \mathcal{T}$ with a Gaussian sample $\mathbf{x}_\mathcal{T} \sim \mathcal{N}(\mathbf{0}, \sigma_\mathcal{T}^2\mathbf{I})$, we use an appropriate estimate $\tilde{\mathbf{x}}$ based on past observations which we can diffuse forward up to $\tau = \tau'$. The initialization of the (shortened) diffusion trajectory then becomes $\mathbf{x}_{\tau'} \sim \mathcal{N}(\alpha_{\tau'}\tilde{\mathbf{x}}, \sigma_{\tau'}^2\mathbf{I})$. For the discretized case, this reduces the number of steps to $N' \ll N$, with $N' = N\tau'/\mathcal{T}$.

A. SeqDiff

One straightforward method of initialization given past past observations is to directly use the previous diffusion posterior estimate $\mathbf{x}_0^t$ as an estimate for the mean of $\mathbf{x}^{t+1}$. This would lead to the following diffusion initialization for $t+1$:

$$\mathbf{x}_{\tau'}^{t+1} \sim p(\mathbf{x}_{\tau'}^{t+1} \mid \mathbf{x}_0^{t+1}) \approx \mathcal{N}(\alpha_{\tau'}\mathbf{x}_0^t, \sigma_{\tau'}^2\mathbf{I}). \tag{7}$$

This assumes a simple linear sequential model, which we show is reasonable in case of high frame rate scenarios where the temporal correlation across subsequent frames is strong.

B. SeqDiff+

In cases of severe motion or lower frame rates we leverage a ViViT network $f_\phi(\cdot)$ to model the system dynamics and predict the mean of the next frame for improved initialization. This allows us to improve on (7), as follows:

$$\mathbf{x}_{\tau'}^{t+1} \sim p(\mathbf{x}_{\tau'}^{t+1} \mid \mathbf{x}_0^{t+1}) \approx \mathcal{N}(\alpha_{\tau'}\tilde{\mathbf{x}}_0^{t+1}, \sigma_{\tau'}^2\mathbf{I}), \tag{8}$$

where $\tilde{\mathbf{x}}_0^{t+1}$ is predicted by the transformer model f_ϕ, parameterized with ϕ, which takes as input a sequence of past posterior estimates and outputs a prediction of the next frame as follows:

$$\tilde{\mathbf{x}}_0^{t+1} = f_\phi(\mathbf{x}_0^t, \mathbf{x}_0^{t-1}, ..., \mathbf{x}_0^{t-K}). \tag{9}$$

A full comparison of all diffusion initialization methods is listed in Table I, and illustrated in Fig. 1

IV. Results

To evaluate our methods, we test conditional diffusion trajectories (vanilla DPS) with and without SeqDiff and SeqDiff+ initialization strategies on the EchoNet-Dynamic dataset [23] with approximately 7000 sequences of around 80 to 300 frames each of which we reserve 100 sequences for evaluation. We map all images to a polar grid to retrieve the original scanning lines and resize to 128×128. After inference the images are scan converted back to cartesian grid for display and metrics calculation. Subsampling is a compressed sensing technique frequently used in medical imaging to reduce data rates [24]–[27]. As a reconstruction task for the diffusion model we consider a scan-line undersampling task which can be used in ultrasound imaging to reduce acquisition time and is essentially subsampling of the image columns. For SeqDiff+ we use a ViViT architecture with context length $K = 4$, two transformer layers with 8 heads for encoder and decoder each and a tublet size of $(2, 16, 16)$. Unless specified otherwise, results are generated with only $N' = 4$ diffusion steps. In Fig. 2 a visual comparison of the initialization methods is shown, with the proposed methods clearly outperforming both vanilla DPS given the same number of diffusion steps, as well as full diffusion trajectory with $25\times$ fewer steps. This is

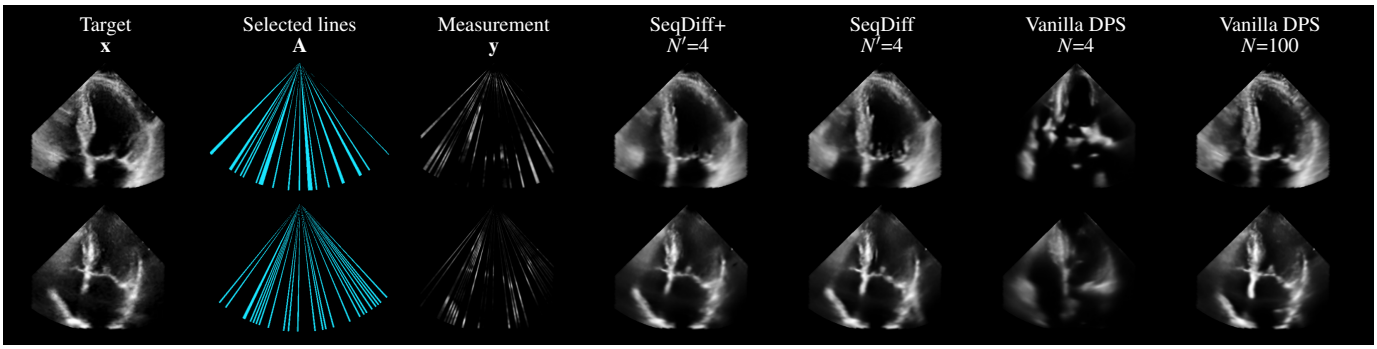

Fig. 2: Qualitative comparison of Vanilla DPS (for $N = 4$ and $N = 100$ steps), and the two proposed initialization methods SeqDiff and SeqDiff+ for only $N' = 4$ diffusion steps. Target images $\mathbf{x}^t$ are 80% masked by $\mathbf{A}^t$ to produce observation $\mathbf{y}^t$. Initialization with SeqDiff(+) is able to improve on full diffusion trajectories with $25\times$ speedup.

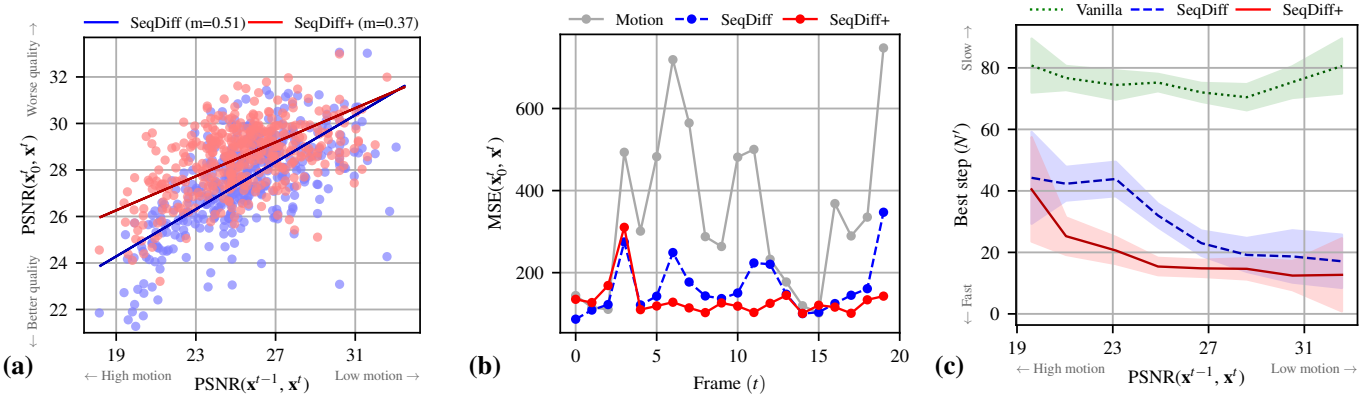

Fig. 3: Comparison of SeqDiff(+) performance in PSNR against various motion conditions. **(a)** For every sample in the test set. The advantage of using a transition model (SeqDiff+) is most advantageous with high motion (see linear fit m). **(b)** For a single sequence of frames. SeqDiff+ is less correlated with the motion, whereas the error of SeqDiff increases with more movement, emphasizing the importance of the transition model. **(c)** Best performing N' for each initialization method against motion. SeqDiff+ outperforms the other methods for all motion levels. For lower motion levels, SeqDiff is a valid option.

reflected in the metrics too, as seen in Fig. 4, where SeqDiff+ initialization outperforms its counterpart without transition model, especially for low N'. For high N' the performance tapers off as useful past information is *forgotten* due to the noise being added. The importance of an accurate transition model is highlighted in Fig. 3a, Fig. 3b and Fig. 3c which compare the performance against motion. We observe that in cases with higher motion it pays off to use the ViViT to account for the dynamics. Furthermore, based on the amount of motion, SeqDiff(+) offers a way to determine the optimal initialization point τ' as seen in Fig. 3c.

V. Conclusions

In this paper, we introduce a novel sequential posterior sampling approach, coined SeqDiff(+), to accelerate diffusion models in the context of sequence data. Our method capitalizes on the temporal structure between subsequent frames which enables autoregressive sampling based on previous posterior estimates. Additionally, we adapt a Video Vision Transformer (ViViT) to model the transition dynamics between frames for improved initialization of the diffusion process. Our approach effectively reduces the number of diffusion iterations with respect to full conditional diffusion trajectories up to $25\times$,

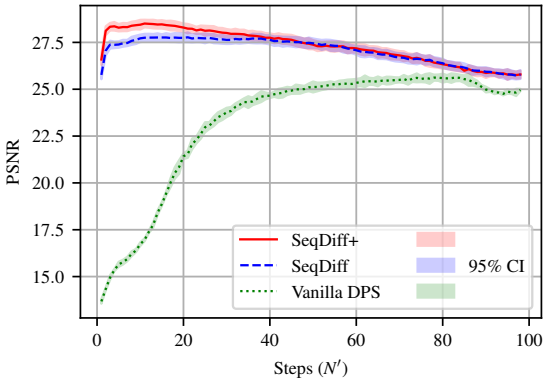

Fig. 4: PSNR against number of diffusion steps on sequences of frames from the test split of EchoNet-Dynamic dataset. Confidence Interval (CI) is taken over 3 splits with different masks and seeds. SeqDiff+ shows a notable improvement, particularly with fewer diffusion steps N'.

unlocking the use of diffusion models for real-time imaging applications such as ultrasound imaging. We evaluate our approach on scan-line undersampling in cardiac ultrasound frames and show that, especially in cases with severe motion, the addition of a transition model further improves performance.

REFERENCES

[1] 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 *International Workshop on Advances in Simplifying Medical Ultrasound*. Springer, 2023, pp. 34–43.

[2] 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 Transactions on Medical Imaging*, 2024.

[3] J. Wu, R. Geng, Y. Li, D. Zhang, Z. Lu, Y. Hu, and Y. Chen, "Diffradar: High-quality mmwave radar perception with diffusion probabilistic model," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 8291–8295.

[4] J. Overdevest, X. Wei, H. van Gorp, and R. van Sloun, "Model-based diffusion for mitigating automotive radar interference," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2024, pp. 284–288.

[5] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," in *International Conference on Learning Representations*, 2021.

[6] B. Jing, G. Corso, R. Berlinghieri, and T. Jaakkola, "Subspace diffusion generative models," *arXiv preprint arXiv:2205.01490*, 2022.

[7] A. Vahdat, K. Kreis, and J. Kautz, "Score-based generative modeling in latent space," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 287–11 302, 2021.

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[9] G. Daras, M. Delbracio, H. Talebi, A. G. Dimakis, and P. Milanfar, "Soft diffusion: Score matching for general corruptions," 2022.

[10] H. Chung, B. Sim, and J. C. Ye, "Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[11] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 8633–8646. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/39235c56aef13fb05a6adc95eb9d8d66-Paper-Conference.pdf

[12] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet *et al.*, "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[13] H. Ni, C. Shi, K. Li, S. X. Huang, and M. R. Min, "Conditional image-to-video generation with latent flow diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 18 444–18 455.

[14] K. Rasul, C. Seward, I. Schuster, and R. Vollgraf, "Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8857–8868.

[15] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[16] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[17] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[18] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye, "Diffusion posterior sampling for general noisy inverse problems," *arXiv preprint arXiv:2209.14687*, 2022.

[19] J. Song, A. Vahdat, M. Mardani, and J. Kautz, "Pseudoinverse-guided diffusion models for inverse problems," in *International Conference on Learning Representations*, 2023.

[20] M. Mardani, J. Song, J. Kautz, and A. Vahdat, "A variational perspective on solving inverse problems with diffusion models," *arXiv preprint arXiv:2305.04391*, 2023.

[21] A. Hauptmann, O. Öktem, and C. Schönlieb, "Image reconstruction in dynamic inverse problems with temporal models," *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging: Mathematical Imaging and Vision*, pp. 1–31, 2021.

[22] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[23] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, E. A. Ashley *et al.*, "Video-based ai for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, no. 7802, pp. 252–256, 2020.

[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 Transactions on Medical Imaging*, vol. 39, no. 12, pp. 3955–3966, 2020.

[25] T. Bakker, H. van Hoof, and M. Welling, "Experimental design for mri by greedy policy search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 954–18 966, 2020.

[26] T. S. Stevens, N. Chennakeshava, F. J. de Bruijn, M. Pekař, and R. J. van Sloun, "Accelerated intravascular ultrasound imaging using deep reinforcement learning," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 1216–1220.

[27] O. Nolan, T. S. Stevens, W. L. van Nierop, and R. J. van Sloun, "Active diffusion subsampling," *arXiv preprint arXiv:2406.14388*, 2024.