Frequency-Dependent Scheduled Schrödinger Bridge for Underwater Acoustic Signal Denoising

Pengsen Zhu[1], **Lina Gao**[1*], **Yulong Huang**[1*], **Lifeng Liu**[2], **Zeru Yang**[2, 3], **Yonggang Zhang**[1*]

[1]College of Intelligent Systems Science and Engineering, Harbin Engineering University, China
[2]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[3]Energy Research Institute @ NTU, Interdisciplinary Graduate Programme, Nanyang Technological University, Singapore
zhu_pengsen@163.com

Abstract

Schrödinger Bridge-based diffusion models have demonstrated promising performance in signal denoising. However, since ground truth signals are unavailable during the sampling process, neural networks must be employed to learn the mapping, which breaks the theoretical coupling between diffusion and sampling processes. This paper reveals a critical inconsistency between the theoretical diffusion path and the learned sampling trajectory across different frequency bands. This diffusion-sampling inconsistency directly undermines denoising effectiveness. To address this limitation, we propose the **F**requency-**D**ependent **S**cheduled **S**chrödinger **B**ridge (**FDSSB**), which leverages power spectral density to adaptively schedule diffusion processes across frequencies. This mechanism assigns asynchronous diffusion schedules to different frequency components, correcting the diffusion schedule to better match the sampling process. As a result, FDSSB effectively mitigates the mismatch and enhances the consistency between diffusion and sampling processes. Extensive experiments demonstrate that FDSSB achieves state-of-the-art performance, with an average scale-invariant signal-to-noise ratio improvement of 7.9066 dB over competitive approaches.

1 Introduction

Marine environments introduce diverse noise sources that significantly degrade the signal-to-noise ratio (SNR) of underwater acoustic signals. Effective underwater acoustic signal denoising is therefore crucial for marine defense and related engineering applications (Cong et al. 2021).

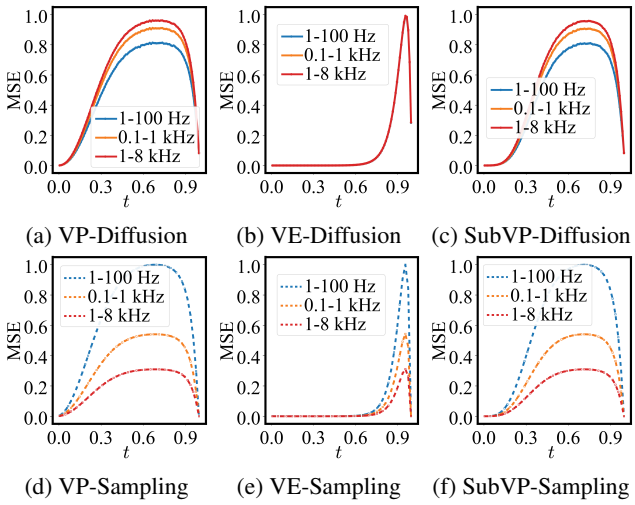
Traditional methods focus on signal decomposition and threshold-based filtering strategies (Li et al. 2018; Xing et al. 2021; Ou, Allen, and Syrmos 2011; Veeraiyan, Velayutham, and Philip 2013). Discriminative deep learning methods later emerged, utilizing long short-term memory (LSTM) and convolutional neural network (CNN) within encoder-decoder frameworks for noise reduction (Zhang et al. 2021; Ju et al. 2022; Duan, Shen, and Wang 2022; Zhou et al. 2023a,b; Duan, Shen, and Wang 2023; Koh, Chia, and Tan 2020). Recently, generative adversarial networks (GANs) have become the dominant generative approach for underwater acoustic signal denoising(Ashraf, Jeong, and Lee 2021; Ashraf et al. 2022). The success of diffusion models (Song et al. 2020) in speech and image denoising (Kulikov et al. 2023; Gong et al. 2024; Kimura et al. 2024; Yuan et al. 2025; Lu et al. 2025) has inspired underwater acoustic signal processing. The Schrödinger Bridge (SB) (Schrödinger 1932) advances beyond conventional diffusion models by enabling optimal transport between arbitrary distributions. However, generative methodologies remain underexplored for underwater acoustic signal denoising.

The effectiveness of SB frameworks critically depends on noise schedule design (Bunne et al. 2023), which governs noise scaling behavior throughout the diffusion process. Current implementations typically apply uniform schedules across all frequency bands, but underwater acoustic signals exhibit distinct frequency-dependent characteristics. According to (Zhu et al. 2023), underwater acoustic signals exhibit characteristics across three frequency bands: low-frequency (1-100 Hz) dominated by propeller rotation sounds, mid-frequency (100-1k Hz) characterized by propeller singing, and high-frequency (1k-8k Hz) containing cavitation noise. Each band's distinct spectral characteristics create inter-band variations, making uniform diffusion schedules prone to diffusion-sampling process mismatches.

To investigate this mismatch, we conducted preliminary experiments using the established frequency band divisions. We quantify the diffusion-sampling inconsistencies using normalized mean squared error (MSE). Specifically, we evaluate three widely adopted noise schedules from (Song et al. 2020): Variance Exploding (VE), Variance Preserving (VP), and SubVP. Fig. 1 presents MSE distributions for the diffusion process on the training set and the sampling process on the validation set. During diffusion, VP and SubVP schedules exhibited frequency-dependent variations, with minimal MSE in the 1-100 Hz band, while VE showed uniform characteristics across all bands. Theoretically, lower diffusion MSE in 1-100 Hz suggests easier reconstruction for this band. However, during sampling, all three schedules consistently showed the same pattern: largest reconstruction errors in the 1-100 Hz band, followed by 100-1k Hz, then 1k-8k Hz. This sampling consistency differs from diffusion observations, revealing a fundamental mismatch between diffusion and sampling behavior.

To address this, we propose Frequency-Dependent Scheduled Schrödinger Bridge (FDSSB). Our key insight is to

*Corresponding Author



(a) VP-Diffusion (b) VE-Diffusion (c) SubVP-Diffusion

(d) VP-Sampling (e) VE-Sampling (f) SubVP-Sampling

Figure 1: Normalized MSE distributions across frequency bands for different noise schedules.

adjust the diffusion process to better match the sampling process across different frequency bands. FDSSB incorporates power spectral density (PSD)-based adjustments and employs straightforward time modulation to achieve asynchronous schedules. Our method enables appropriate time modulation for distinct frequency bands, enabling differentiated diffusion schedule per frequency band, reducing diffusion-sampling process inconsistencies. Our contributions can be summarized as follows:

- We identify diffusion-sampling process inconsistencies across different frequency bands that undermine denoising effectiveness.
- We propose frequency-dependent diffusion schedules based on signal characteristics to reduce the inconsistencies between diffusion and sampling processes.
- We introduce PSD-based time modulation to achieve asynchronous diffusion schedule in a simple manner.
- Extensive experiments demonstrate that our approach achieves state-of-the-art performance and exhibits robust effectiveness across various noise schedules and challenging test scenarios.

2 Related Work and Preliminary

In this section, we introduce related works and necessary Schrödinger Bridge preliminaries.

2.1 Discriminative Denoising

Discriminative methods directly learn mappings from noisy to clean signals. Early approaches utilized temporal networks within encoder-decoder frameworks for amplitude mask estimation (Zhang et al. 2021). Auto-encoders have enabled unsupervised enhancement for low SNR scenarios (Ju et al. 2022). Transformer architectures improved temporal modeling with U-Net for time-domain mask prediction (Duan, Shen, and Wang 2022). Attention mechanisms further advanced performance through fullband-subband networks for multi-frequency processing (Zhou et al. 2023a), dual-branch self-attention for spectral dependencies (Zhou et al. 2023b), and adaptive fusion in dual-branch architectures (Duan, Shen, and Wang 2023). Self-supervised methods like WaveN2N (Koh, Chia, and Tan 2020) have demonstrated effectiveness for multi-channel data.

However, discriminative methods are limited by direct noisy-clean mappings, which restrict complex signal distribution modeling and cause suboptimal performance under diverse noise conditions.

2.2 Generative Denoising

Generative methods model underlying data distributions to reconstruct clean signals probabilistically. GANs have emerged as the dominant paradigm, with early approaches modeling magnitude and phase components using U-Net generators (Ashraf, Jeong, and Lee 2021). Specialized architectures include ambient noise-free GANs for environmental noise suppression (Ashraf et al. 2022) and noise simulators for heavy-tailed impulsive noise distributions (Zhou et al. 2021). However, GAN-based methods face training instability, mode collapse, and reconstruction difficulties.

Recent diffusion models and SB approaches demonstrate strong potential for generative tasks, with successful applications spanning speech denoising, image enhancement, and text generation (Jukić et al. 2024; Han et al. 2025; Chen et al. 2023). SB methods enable direct data-to-data transformations rather than data-to-noise mappings. However, existing SB approaches suffer from critical diffusion-sampling inconsistencies. Particularly in signal denoising applications, current methods apply uniform noise schedules across all frequency bands, thereby ignoring the inherent frequency-dependent characteristics. This creates fundamental mismatches between diffusion and sampling processes, ultimately undermining denoising effectiveness.

While recent work has explored asynchronous scheduling for image restoration (Han et al. 2025) by setting drift terms to zero and considering only diffusion components, this approach presents several limitations that restrict broader applicability. Such methods require restrictive noise schedule constraints, depend on pre-trained models for pixel weighting operations, lead to additional computational costs, and introduce supplementary supervisory information that compromises deployment flexibility. To address these limitations, we propose a more general formulation that leverages power spectral density-informed time modulation. Our approach eliminates restrictive noise schedule assumptions and achieves frequency-dependent asynchronous scheduling through direct temporal modulation, without requiring any additional models or external supervision.

2.3 Schrödinger Bridge Denoising Preliminary

The Schrödinger Bridge (SB) is an energy-normalized optimal transport problem that identifies the most efficient probabilistic transformation between initial and target distributions. This framework operates with two boundary probabil-

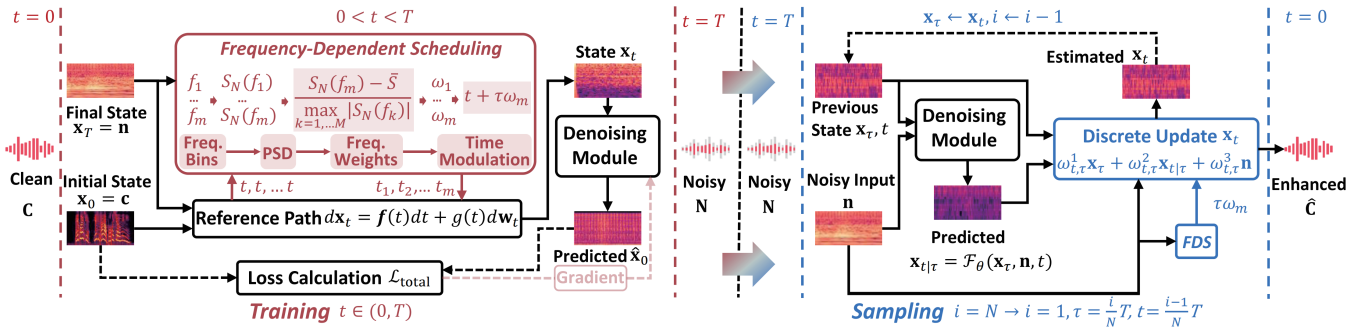


Figure 2: Overview of the proposed FDSSB. $\mathcal{L}_{\text{total}}$ denotes the training loss. During the sampling phase, N-step discrete updates are performed using coefficients $\omega^1_{t,\tau}$, $\omega^2_{t,\tau}$, and $\omega^3_{t,\tau}$, which correspond to the previous state $\mathbf{x}_\tau$, the network prediction $\mathbf{x}_{t|\tau}$, and the noisy observation $\mathbf{n}$, respectively. The time modulation modifies the uniform time step t into a set of frequency-dependent indices $t_1, t_2, \ldots, t_m$, each associated with a frequency component f_m. The modulation is governed by the intensity parameter τ and a normalized weight ω_m, which is derived from the power spectral density $S_{\mathbf{N}}(f_m)$ of the noisy input.

ity distributions p_{data} and p_{prior}, formulating the optimization as a Kullback-Leibler divergence minimization:

$$\min_{p\in\mathcal{P}_{[0,T]}} D(p||p_{\text{ref}}) \quad \text{s. t.} \quad p_0 = p_{\text{data}}, p_T = p_{\text{prior}}, \tag{1}$$

where $\mathcal{P}_{[0,T]}$ denotes the probability measure space over continuous trajectories connecting p_{data} and p_{prior} within the temporal interval $[0, T]$. For siganl denoising tasks, these boundary conditions represent clean and noisy signal distributions, respectively. The baseline transformation is established through the reference path measure p_{ref}, constructed by a forward stochastic differential equation (SDE):

$$\mathrm{d}\mathbf{x}_t = \boldsymbol{f}\big(\mathbf{x}_t, t\big)\,\mathrm{d}t + g(t)\,\mathrm{d}\mathbf{w}_t, \tag{2}$$

where t ranges from 0 to T, with $\mathbf{x}_t$ representing the magnitude spectrum at diffusion time t. The diffusion process is characterized by the vector drift function $\boldsymbol{f}\left(\mathbf{x}_t, t\right)$, scalar diffusion function $g(t)$, and the standard Wiener process $\mathbf{w}_t$.

The optimal solution within the SB paradigm is characterized by a pair of coupled forward and reverse SDEs (Lu et al. 2022) that control the probabilistic transport between boundary distributions. The forward SDE is given by:

$$\mathrm{d}\mathbf{x}_t = [\boldsymbol{f}(\mathbf{x}_t, t) + g^2(t)\nabla \log \Psi(t, \mathbf{x}_t)]\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t, \tag{3}$$

and the corresponding reverse SDE as:

$$\mathrm{d}\mathbf{x}_t = [\boldsymbol{f}(\mathbf{x}_t, t) - g^2(t)\nabla \log \bar{\Psi}(t, \mathbf{x}_t)]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}_t. \tag{4}$$

Here, $\boldsymbol{f}(\mathbf{x}_t, t)$, $g(t)$, and $\mathbf{w}_t$ preserve their meanings established in (2), with $\bar{\mathbf{w}}_t$ representing the reverse-time Wiener process. The terms $\nabla \log \Psi(t, \mathbf{x}_t)$ and $\nabla \log \bar{\Psi}(t, \mathbf{x}_t)$ represent additional nonlinear drift components. These terms satisfy the coupled partial differential equations (PDEs) (Chen, Liu, and Theodorou 2021):

$$\begin{cases} \frac{\partial \Psi}{\partial t} = -\nabla_{\mathbf{x}} \Psi^\top \boldsymbol{f}(\mathbf{x}_t, t) - \frac{1}{2}\,\mathrm{Tr}\big(g^2(t)\nabla^2_{\mathbf{x}}\Psi\big), \\ \frac{\partial \bar{\Psi}}{\partial t} = -\nabla_{\mathbf{x}}(\bar{\Psi}\,\boldsymbol{f}(\mathbf{x}_t, t)) + \frac{1}{2}\,\mathrm{Tr}\big(g^2(t)\nabla^2_{\mathbf{x}}\bar{\Psi}\big). \end{cases} \tag{5}$$

Ψ and $\bar{\Psi}$ represent $\Psi(t, \mathbf{x})$ and $\bar{\Psi}(t, \mathbf{x})$, respectively. These functions satisfy boundary constraints $\Psi(0, \cdot)\bar{\Psi}(0, \cdot) = p_{\text{data}}$ and $\Psi(T, \cdot)\bar{\Psi}(T, \cdot) = p_{\text{prior}}$.

As shown in (Chen, Liu, and Theodorou 2021), the marginal distribution p_t at any time t can be expressed as $\Psi(t, \cdot)\bar{\Psi}(t, \cdot) = p_t$, which gives the relationship:

$$\nabla \log \Psi(t, \mathbf{x}_t) + \nabla \log \bar{\Psi}(t, \mathbf{x}_t) = \nabla \log p_t(\mathbf{x}_t). \tag{6}$$

When $\nabla \log \Psi(t, \mathbf{x}_t) = 0$, substituting (6) into (3) and (4) shows that score-based generative models (SGMs) are a special case of the SB framework. In subsequent sections, we will derive analytical solutions for our proposed asynchronous scheduling and utilize these closed-form solutions to perform diffusion and sampling.

3 Methodology

In this section, we first formulate the signal denoising problem and introduce a frequency-dependent scheduling strategy. Then,we derive the analytical solution for this framework. Finally, we introduce the training process and the sampling scheme. Fig. 2 presents an overview of FDSSB.

3.1 Problem Formulation

Let $\mathbf{C} \in \mathbb{R}^L$ denote the clean underwater acoustic signal of length L. During propagation, this signal is corrupted by noise, resulting in the observed noisy signal $\mathbf{N} \in \mathbb{R}^L$. The denoising task aims to recover an estimate $\hat{\mathbf{C}} \in \mathbb{R}^L$ that approximates the clean signal. We formulate the problem using magnitude spectral representations. The short-time Fourier transform is applied to obtain the magnitude spectra:

$$\mathbf{c} = |\mathcal{T}(\mathbf{C})|, \quad \mathbf{n} = |\mathcal{T}(\mathbf{N})|, \tag{7}$$

where $\mathbf{c}, \mathbf{n} \in \mathbb{R}^D$ are the clean and noisy magnitude spectra, respectively. D denotes the frequency-domain dimension, $\mathcal{T}$ represents the Fourier transform operator, and $|\cdot|$ denotes the magnitude operation. The denoising model estimates the clean magnitude spectrum from the noisy input, and the final time-domain output $\hat{\mathbf{C}}$ is obtained through phase reconstruction and inverse Fourier transformation.

3.2 Frequency-Dependent Scheduling

As showed in Fig. 1, the existing schedule exhibits a critical mismatch: during the diffusion process, noise injection

intensity decreases from high frequency to mid frequency to low frequency bands. However, in the sampling process, we observe that low frequency bands exhibit the highest reconstruction difficulty. To address this scheduling inconsistency, we propose FDSSB. Our approach leverages PSD characteristics of different frequency bands as the basis for adaptive schedule adjustment. We employ a temporal shift strategy that offsets the diffusion process across frequency bands, enabling the diffusion to better align with the sampling.

Given the time-domain noisy signal $\mathbf{N} \in \mathbb{R}^L$, we first compute the PSD $S_{\mathbf{N}}(f)$ at frequency f. We then calculate the mean PSD across all frequency bins:

$$\bar{S} = \frac{1}{M} \sum_{m=1}^{M} S_{\mathbf{N}}(f_m), \tag{8}$$

where f_m represents the center frequency of the m-th frequency bin and M denotes the total number of frequency bins. To quantify spectral characteristics, we compute normalized frequency weights:

$$\omega_m = \frac{S_{\mathbf{N}}(f_m) - \bar{S}}{\max_{k=1,\ldots,M} |S_{\mathbf{N}}(f_k)|}, \tag{9}$$

where ω_m represents the normalized frequency weights of the m-th frequency bin. Based on these weights, we propose a frequency-dependent time modulation by applying a temporal shift to the original diffusion time t:

$$t_m = t + \tau \omega_m, \tag{10}$$

where $\tau \in \mathbb{R}^+$ controls the modulation intensity, and t_m represents the modulated diffusion time for each frequency component. t_m is clamped such that it does not exceed the maximum time step T.

Positive weights $\omega_m > 0$ in power-concentrated regions yield $t_m > t$, resulting in accelerated diffusion. Since underwater acoustic signals typically exhibit energy concentration in the low-frequency range (Zhu et al. 2023), this modulation naturally accelerates diffusion in those regions, aligning well with their sampling characteristics shown in Fig. 1. Conversely, negative weights $\omega_m < 0$ in power-dispersed regions produce $t_m < t$, leading to decelerated diffusion, which also conforms to our design objective of slowing down the diffusion process of high-frequency bands. Through this mechanism, we can easily adjust the diffusion time t for different frequency bands, thereby achieving asynchronous scheduling tailored to the spectral characteristics of each frequency component.

3.3 Analytical Solution Derivation

To obtain analytical solutions, boundary conditions are constructed by adding Gaussian noise with variances ε_1^2 and ε_2^2 to the boundary points $\mathbf{c}$ and $\mathbf{n}$, respectively. For linear drift terms $\boldsymbol{f}(\mathbf{x}_t, t) = f(t)\mathbf{x}_t$, these variances satisfy:

$$\varepsilon_2 = \exp\left(\int_0^1 f(t)dt\right)\varepsilon_1. \tag{11}$$

This yields the boundary distributions:

$$p_{\text{data}} = \mathcal{N}(\mathbf{c}, \varepsilon_1^2\mathbf{I}), \quad p_{\text{prior}} = \mathcal{N}(\mathbf{n}, \varepsilon_2^2\mathbf{I}). \tag{12}$$

Under these boundary conditions and linear drift form $\boldsymbol{f}(\mathbf{x}_t, t) = f(t)\mathbf{x}_t$. As $\varepsilon_1 \to 0$, this formulation enables us to derive a complete analytical solution (Bunne et al. 2023; Chen, Liu, and Theodorou 2021) for (5):

$$\bar{\Psi} = \mathcal{N}(\alpha_t \mathbf{c}, \alpha_t^2\sigma_t^2\mathbf{I}), \quad \Psi = \mathcal{N}(\bar{\alpha}_t \mathbf{n}, \alpha_t^2\bar{\sigma}_t^2\mathbf{I}), \tag{13}$$

where α_t and σ_t^2 represent the diffusion parameters:

$$\alpha_t = e^{\int_0^t f(\tau)d\tau}, \quad \sigma_t^2 = \int_0^t \frac{g^2(\tau)}{\alpha_\tau^2}d\tau, \tag{14}$$

while $\bar{\alpha}_t$ and $\bar{\sigma}_t^2$ are also diffusion parameters that satisfy the following relationships:

$$\bar{\alpha}_t = \alpha_t \alpha_T^{-1}, \quad \bar{\sigma}_t^2 = \sigma_T^2 - \sigma_t^2. \tag{15}$$

In our framework, the temporal parameter t is modified to incorporate frequency-dependent diffusion characteristics. For each frequency component f_m, we apply the modulated time parameter t_m defined in (10). To establish frequency-dependent diffusion parameters, we define $\beta(t, f_m)$, $\bar{\beta}(t, f_m)$, $\delta^2(t, f_m)$, and $\bar{\delta}^2(t, f_m)$ as the frequency-dependent counterparts of the diffusion parameters α_t, $\bar{\alpha}_t$, σ_t^2, and $\bar{\sigma}_t^2$, respectively. The relationships between these parameters are expressed as:

$$\beta(t, f_m) = \alpha_{t_m}, \quad \bar{\beta}(t, f_m) = \bar{\alpha}_{t_m}, \tag{16}$$

$$\delta^2(t, f_m) = \sigma_{t_m}^2, \quad \bar{\delta}^2(t, f_m) = \bar{\sigma}_{t_m}^2. \tag{17}$$

According to (Chen, Liu, and Theodorou 2021), the marginal distribution p_t given frequency f_m follows:

$$p_t = \mathcal{N}\left(\omega_{\mathbf{c}}(t, f_m)\mathbf{c}_{f_m} + \omega_{\mathbf{n}}(t, f_m)\mathbf{n}_{f_m}, \Sigma_{t,f_m}\mathbf{I}\right), \tag{18}$$

where $\mathbf{c}_{f_m}$ and $\mathbf{n}_{f_m}$ denote the clean and noisy signals at frequency f_m. $\mathbf{I}$ is the identity matrix. The mean is computed as a weighted combination of clean and noisy signals, with the weight for the clean signal $\omega_{\mathbf{c}}(t, f_m)$ defined as:

$$\omega_{\mathbf{c}}(t, f_m) = \frac{\beta(t, f_m)\bar{\delta}^2(t, f_m)}{\sigma_T^2}. \tag{19}$$

By substituting(16) and (17), we obtain:

$$\omega_{\mathbf{c}}(t, f_m) = \frac{\alpha_{t_m}\bar{\sigma}_{t_m}^2}{\sigma_T^2}. \tag{20}$$

Correspondingly, the weight for the noisy signal, $\omega_{\mathbf{n}}(t, f_m)$, and the variance Σ_{t,f_m} are given by:

$$\omega_{\mathbf{n}}(t, f_m) = \frac{\bar{\alpha}_{t_m}\sigma_{t_m}^2}{\sigma_T^2}, \quad \Sigma_{t,f_m} = \frac{\alpha_{t_m}^2\bar{\sigma}_{t_m}^2\sigma_{t_m}^2}{\sigma_T^2}. \tag{21}$$

We now derive the analytical formulations under our noise schedules. For underwater acoustic signals, we design the diffusion process as a combination of two components: signal transformation, governed by the drift term, and noise injection, managed by the diffusion term. Motivated by the exponential attenuation of acoustic pressure during propagation (Jensen et al. 2011), the drift term is formulated as:

$$\boldsymbol{f}(\mathbf{x}_t, t) = \beta_1 e^{kt}\mathbf{x}_t, \tag{22}$$

Method	Type	Year	SI-SNRi (dB) ↑					SDRi (dB) ↑				
			Mixed	Rain	Wave	Wind	Average	Mixed	Rain	Wave	Wind	Average
DCUnet-20	D	2018	5.2002	4.1787	4.3869	4.6787	4.6111	5.9244	5.3067	5.4715	5.0724	5.4438
Conv-TasNet	D	2019	7.3368	6.1781	6.4705	7.8162	6.9504	7.9119	7.1396	7.3314	8.1301	7.6283
FullSubNet	D	2020	4.3575	3.3046	3.4864	4.1826	3.8328	5.0905	4.3123	4.4252	4.9861	4.7035
DCCRN	D	2020	6.1141	5.0147	5.1926	6.0566	5.5945	6.8042	6.1727	6.2795	6.6071	6.4659
DPTNet	D	2020	7.2938	6.0858	6.3238	6.8787	6.6455	8.1442	7.5223	7.5389	7.5125	7.6795
DPRNN	D	2020	7.6406	6.6435	6.6316	6.9288	6.9611	8.5192	8.0995	7.8819	7.5761	8.0192
FullSubNet+	D	2022	0.0213	-0.0004	0.0088	-0.0171	0.0032	0.0222	0.0016	0.0101	-0.0153	0.0046
CleanUNet	D	2022	0.8176	-0.4224	-0.2124	-0.1282	0.0137	1.1847	-0.0186	0.1859	0.3185	0.4176
MP-SENet	D	2023	6.1764	5.1166	5.3459	6.0515	5.6726	6.9242	6.3120	6.3820	6.8172	6.6089
TDANet	D	2023	5.8353	4.5688	4.7271	5.3182	5.1124	6.5443	5.7855	5.7628	5.9466	6.0098
SGMSE	G	2023	9.5432	8.6350	8.6148	8.9235	8.9291	9.4410	8.5406	8.5296	8.9025	8.8534
OUVE	G	2023	7.1542	6.0290	6.1738	7.4005	6.6894	7.2290	6.2166	6.3296	7.5045	6.8199
SpatialNet	D	2024	6.8866	5.7526	5.9820	6.1168	6.1845	7.9691	7.4320	7.5395	6.7884	7.4323
SepReformer	D	2024	6.0042	4.7643	4.9238	5.6719	5.3411	6.8249	6.1425	6.1156	6.1388	6.3055
SEMamba	D	2024	6.0322	4.9500	5.2468	5.6665	5.4739	6.4180	5.5756	5.8226	5.8965	5.9282
CMGAN	G	2024	6.3120	5.1143	5.3348	6.3529	5.7785	7.4053	6.8752	6.8725	7.3361	7.1223
RGDNet	G	2025	11.1453	10.8872	10.7989	8.1963	10.2569	11.5960	11.5488	11.4514	8.5728	10.7923
FDSSB (Ours)	G	2025	**14.2108**	**13.8241**	**14.2054**	**11.5159**	**13.4391**	**14.3744**	**14.9748**	**13.9444**	**12.5523**	**13.9615**

Table 1: Performance comparison on ShipsEar dataset under -5 dB SNR condition. The best scores are highlighted in bold, the second-best scores are underlined, and ↑ indicates that higher scores denote better performance. G and D indicate generative and discriminative methods, respectively.

where β_1 is a negative constant that serves as an attenuation coefficient, controlling the intensity of signal decay. The term $k > 0$ denotes the exponential growth rate. The diffusion coefficient is designed to align with the exponential structure of the drift term:

$$g^2(t) = \beta_2 e^{kt}, \tag{23}$$

β_2 is a positive constant that regulates the diffusion intensity.

Substituting the designed (22) and (23) into (14), we can derive explicit formulas for the parameters:

$$\alpha_t = e^{\frac{\beta_1}{k}(e^{kt}-1)}, \quad \sigma_t^2 = \frac{\beta_2}{2\beta_1}(1 - e^{-2\frac{\beta_1}{k}(e^{kt}-1)}). \tag{24}$$

By substituting these expressions into equations (20), (21), we obtain analytical expressions for the diffusion process.

3.4 Model Training

Based on the closed-form solution from (18), the noisy signal $\mathbf{x}_t$ at time step t for frequency component f_m becomes:

$$\mathbf{x}_t = \omega_{\mathbf{c}}(t, f_m)\mathbf{c}_{f_m} + \omega_{\mathbf{n}}(t, f_m)\mathbf{n}_{f_m} + \sqrt{\Sigma_{t,f_m}}\mathbf{z}, \tag{25}$$

where $\mathbf{c}_{f_m}$ and $\mathbf{n}_{f_m}$ represent clean and noisy signals at frequency f_m, $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ denotes standard Gaussian noise, and $\omega_{\mathbf{c}}(t, f_m), \omega_{\mathbf{n}}(t, f_m), \Sigma_{t,f_m}$ constitute frequency-dependent temporal coefficients specified in (20) and (21). The network estimates clean signal $\hat{\mathbf{x}}_0$ as:

$$\hat{\mathbf{x}}_0 = \mathcal{F}_\theta(\mathbf{x}_t, \mathbf{n}, t), \tag{26}$$

$\mathcal{F}_\theta$ denotes the neural network with parameters θ.

Model training integrates multiple losses across both frequency and time domains. Specifically, the time-frequency domain loss $\mathcal{L}_{\text{tf}}$ computes the mean squared error between predicted and target spectrograms, while the time-domain loss $\mathcal{L}_{\text{l1}}$ directly applies the L1 loss. To further improve perceptual reconstruction quality, we incorporate the scale-invariant signal-to-distortion ratio (SI-SDR) and signal-to-noise ratio (SI-SNR), with their negative values used as losses $\mathcal{L}_{\text{sdr}}$ and $\mathcal{L}_{\text{snr}}$. The final loss is formulated as a weighted combination of all components:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{tf}}\mathcal{L}_{\text{tf}} + \lambda_{\text{l1}}\mathcal{L}_{\text{l1}} + \lambda_{\text{sdr}}\mathcal{L}_{\text{sdr}} + \lambda_{\text{snr}}\mathcal{L}_{\text{snr}}. \tag{27}$$

where λ_{tf}, λ_{l1}, λ_{sdr}, λ_{snr} serve as hyperparameters for loss term weighting.

3.5 Sampling Scheme

For the sampling process, we employ a deterministic sampler based on probability flow ordinary differential equations (ODEs) for fast sampling (Lu et al. 2022; Calvo and Palencia 2006; Hochbruck, Ostermann, and Schweitzer 2009). For each frequency component f_m, we apply the parameters established in (16) and (17). The update for $\mathbf{x}_t$ is formulated as a weighted combination of three key components: the previous state $\mathbf{x}_\tau$, the network's prediction $\mathbf{x}_{t|\tau} = \mathcal{F}_\theta(\mathbf{x}_\tau, \mathbf{n}, t)$, and the initial noisy observation $\mathbf{n}$:

$$\mathbf{x}_t = \omega_{t,\tau}^1\mathbf{x}_\tau + \omega_{t,\tau}^2\mathbf{x}_{t|\tau} + \omega_{t,\tau}^3\mathbf{n}, \tag{28}$$

where t and τ represent two adjacent time steps in the sampling process. For a total of N sampling steps, these time values are computed as:

$$\tau = \frac{i}{N}T, \quad t = \frac{i-1}{N}T, \tag{29}$$

for step index i, where i decreases from N to 1 during the sampling process, and T is the terminal time.

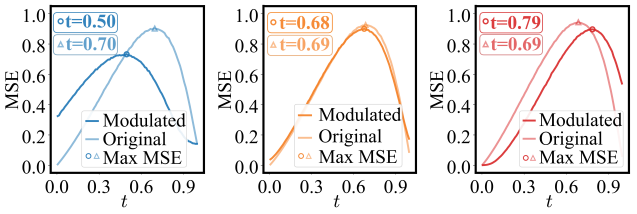
The frequency-modulated time steps t_m are obtained from (10), and τ_m is computed analogously. For each frequency

Method	SI-SNRi (dB) ↑		SDRi (dB) ↑	
	Mixed	Average	Mixed	Average
Conv-TasNet	9.6593	12.0705	9.5097	12.4264
DCCRN	7.9348	7.6271	7.9974	8.0720
DPTNet	9.5955	8.0651	9.6606	8.2203
DPRNN	10.2298	8.4761	10.3675	8.6678
TDANet	7.7469	9.1136	7.7612	9.2782
SpatialNet	9.1518	10.4745	9.6037	11.3828
SepReformer	7.8187	8.8096	7.8836	9.1999
CMGAN	8.7205	7.3224	9.0873	8.0764
RGDNet	<u>13.9406</u>	<u>13.0362</u>	<u>13.8661</u>	<u>13.2496</u>
FDSSB (Ours)	**17.8364**	**13.3168**	**17.9146**	**13.4734**

Table 2: Performance under -10 dB condition. Bold and underlined indicate best and second-best scores.

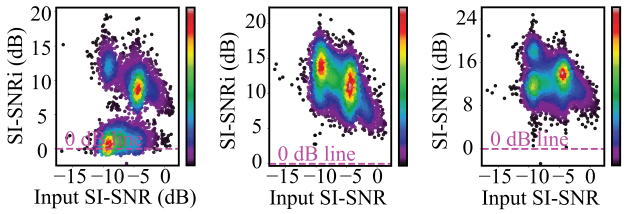
Method	SI-SNRi (dB) ↑		SDRi (dB) ↑	
	Mixed	Average	Mixed	Average
Conv-TasNet	8.6135	10.7950	8.8643	11.5227
DCCRN	7.0044	6.8967	7.4450	7.7274
DPTNet	8.5604	7.5805	9.0445	8.2610
DPRNN	9.0164	7.7946	9.5626	8.4741
TDANet	6.8848	8.4754	7.2803	9.2068
SpatialNet	8.0313	9.3084	8.8283	10.5454
SepReformer	7.0230	8.0830	7.4841	8.9776
CMGAN	7.5795	6.5608	8.3237	7.7180
RGDNet	<u>12.5282</u>	<u>11.7253</u>	<u>12.7689</u>	<u>12.2771</u>
FDSSB (Ours)	**15.9431**	**12.0374**	**16.2531**	**12.5684**

Table 3: Performance under -10 to -5 dB condition. Bold and underlined indicate best and second-best scores.



(a) 1-100 Hz (b) 100-1K Hz (c) 1K-8K Hz

Figure 3: Normalized MSE distributions across frequency bands during diffusion process.



(a) SGMSE (b) RGDNet (c) FDSSB (Ours)

Figure 4: SI-SNR improvement distribution. Points above the 0 dB line indicate improvement.

component f_m, the coefficient $\omega_{t,\tau}^1$ is given by:

$$\omega_{t,\tau}^1 = \frac{\alpha_{t_m}\sigma_{t_m}\bar{\sigma}_{t_m}}{\alpha_{\tau_m}\sigma_{\tau_m}\bar{\sigma}_{\tau_m}}. \tag{30}$$

The coefficient $\omega_{t,\tau}^2$ is expressed as:

$$\omega_{t,\tau}^2 = \frac{\alpha_{t_m}}{\sigma_T^2}\left(\bar{\sigma}_{t_m}^2 - \frac{\bar{\sigma}_{\tau_m}\sigma_{t_m}\bar{\sigma}_{t_m}}{\sigma_{\tau_m}}\right), \tag{31}$$

while $\omega_{t,\tau}^3$ is defined as:

$$\omega_{t,\tau}^3 = \frac{\alpha_{t_m}}{\alpha_{T_m}\sigma_T^2}\left(\sigma_{t_m}^2 - \frac{\sigma_{\tau_m}\sigma_{t_m}\bar{\sigma}_{t_m}}{\bar{\sigma}_{\tau_m}}\right). \tag{32}$$

The complete sampling process operates as follows: starting from the noisy observation $\mathbf{x}_T = \mathbf{n}$, we iteratively apply the update equation using the coefficients $\omega_{t,\tau}^1$, $\omega_{t,\tau}^2$, and $\omega_{t,\tau}^3$ until reaching $t = 0$. This yields the clean signal estimate $\hat{\mathbf{c}} = \mathbf{x}_0$, which is then converted back to the time domain via inverse Fourier transform.

4 Numerical Experiments

In this section, we first introduce the implementation details. Then, we conduct comprehensive comparative experiments, followed by ablation studies.

4.1 Implementation Details

The ShipsEar (Santos-Domínguez et al. 2016) dataset, an established benchmark for underwater acoustic signal processing extensively utilized in recent research (Zhou et al. 2023a,b; Zhang et al. 2021; Ashraf et al. 2022; Ashraf, Jeong, and Lee 2021; Song, Liur, and Shen 2022; Zhu et al. 2025), is employed for validation. Audio recordings undergo segmentation into 3-second samples with 8 kHz resampling. Dataset construction follows the methodologies from (Zhu et al. 2025) using available open-source implementations.

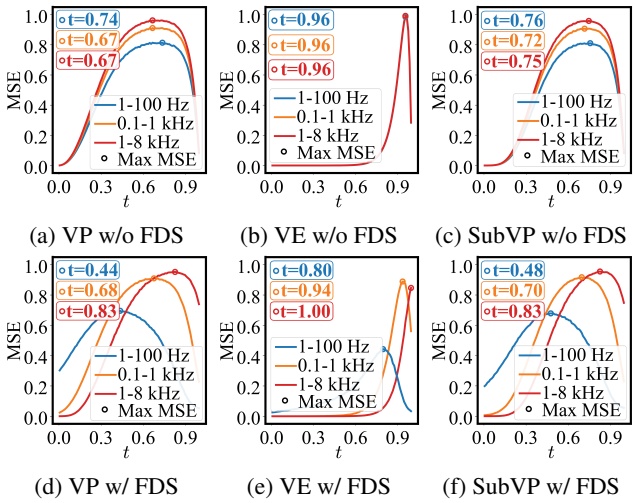
The experimental design encompasses training data with fixed SNR (-5 dB) across four distinct noise scenarios, while testing data incorporates three SNR levels across identical scenarios, yielding 12 test conditions. Noise scenarios include wind, wave, rain, and mixed conditions combining all three types. Testing SNR levels comprise -5 dB, -10 dB, and randomized values between -10 dB and -5 dB.

Performance assessment utilizes scale-invariant signal-to-noise ratio (SI-SNR) (Luo and Mesgarani 2018) and signal-to-distortion ratio (SDR) (Vincent, Gribonval, and Févotte 2006) metrics. We report SI-SNRi and SDRi, which measure the improvement in SI-SNR and SDR.

4.2 Comparative Experiments

To evaluate the proposed FDSSB, comparative experiments are conducted against 17 competitive methods. For the -5 dB SNR condition, we evaluate performance across all four noise scenarios. For the -10 dB and variable SNR conditions, we focus on the mixed scenario and overall average performance. For clear presentation, we report the top-10 baselines along with our method.

As shown in Table 1, under the -5 dB SNR condition, FDSSB achieves an average SI-SNRi of 13.44 dB and SDRi



(a) VP w/o FDS (b) VE w/o FDS (c) SubVP w/o FDS

(d) VP w/ FDS (e) VE w/ FDS (f) SubVP w/ FDS

Figure 5: Normalized MSE distribution for the diffusion process, comparing noise schedules with and without FDS.

Schedule	SI-SNRi (dB) ↑		SDRi (dB) ↑	
	w/o FDS	w/ FDS	w/o FDS	w/ FDS
VE	9.6596	**10.4751**	10.7842	**12.0271**
VP	8.0616	**8.4488**	8.8352	**9.4337**
SubVP	1.5446	**10.4051**	3.2172	**11.5373**
FDSSB	11.6437	**13.4391**	12.3935	**13.9615**

Table 4: Ablation study of FDS across different noise schedules. The best results are highlighted in bold.

of 13.96 dB across all noise scenarios, substantially outperforming all comparative methods. The improvements over RGDNet, the strongest baseline, reach 3.18 dB in SI-SNRi and 3.17 dB in SDRi. FDSSB performs particularly well in mixed noise conditions, achieving SI-SNRi and SDRi of 14.21 dB and 14.37 dB, respectively. Even in the most challenging wind noise scenario, FDSSB maintains competitive performance with 11.52 dB SI-SNRi and 12.55 dB SDRi.

Table 2 presents results for the -10 dB SNR condition. FDSSB consistently outperforms all baselines, with particularly strong performance in mixed noise scenarios and significant improvements over RGDNet. The method maintains consistent quality despite severely degraded input conditions. Under variable SNR conditions, as detailed in Table 3, FDSSB continues to demonstrate superior performance across all tested scenarios.

Fig. 3 shows the impact of frequency-dependent scheduling across different bands. The low-frequency band shows an leftward shift, indicating accelerated diffusion schedule, while the high-frequency band demonstrates a clear rightward shift, reflecting decelerated diffusion schedule. The mid-frequency band maintains relatively stable positioning with minimal curve movement. These temporal shifts are consistent with our theoretical analysis in Section 3.2.

To better understand the denoising characteristics at the sample level, we conducted a detailed analysis of SI-SNRi

τ	SI-SNRi (dB) ↑		SDRi (dB) ↑	
	Mixed	Average	Mixed	Average
0.3	12.4705	11.4971	12.7801	11.9103
0.7	12.6261	11.6001	12.9099	11.9766
1.0	**14.2108**	**13.4391**	**14.3744**	**13.9615**
1.5	13.1534	12.1553	13.4416	12.5402
2.0	12.5714	11.5005	12.9861	12.0130

Table 5: Performance comparison of different τ. The best results are highlighted in bold.

distribution patterns for the top-3 methods. As illustrated in Fig. 4, FDSSB demonstrates outstanding performance with SI-SNRi values concentrated in the range of 8-24 dB, forming a dense high-performance distribution well above the 0 dB baseline. RGDNet shows the second-best performance with improvements distributed across 5-20 dB, though with greater variability and lower peak values compared to FDSSB. SGMSE displays a mixed distribution pattern with some samples achieving improvements around 5-15 dB while others show limited enhancement near the 0 dB line. The distribution patterns clearly indicate that FDSSB not only achieves superior average performance but also maintains remarkable consistency across all input samples.

4.3 Ablation Studies

To validate the effectiveness of the proposed frequency-dependent scheduling (FDS), ablation studies are conducted comparing FDSSB against conventional diffusion schedules. In addition, we evaluate three widely adopted noise schedules: VE, VP, and SubVP(Song et al. 2020).

The results in Table 4 demonstrate that all schedules benefit from FDS, with SubVP exhibiting the most substantial enhancement, while VP and VE show moderate improvements. We also evaluate the impact of τ settings on performance. Table 5 shows that $\tau = 1.0$ achieves optimal results.

Fig. 5 illustrates the impact of FDS on the diffusion process. The analysis shows that the 1–100 Hz band consistently shifts leftward, the low-frequency band remains relatively stable, while the high-frequency band uniformly shifts rightward. This pattern confirms that FDS accelerates diffusion in low-frequency bands and decelerates it in high-frequency bands, effectively aligning the diffusion process with the sampling process.

5 Conclusion

In this work, we introduce frequency-dependent scheduling, which adaptively adjusts diffusion across frequency bands. By aligning diffusion and sampling, FDSSB achieves state-of-the-art performance. While it consistently improves results across noise schedules, applying it to other domains may face challenges due to differing signal characteristics. Nonetheless, this work offers a valuable perspective on frequency-dependent scheduling for related tasks.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62503126, U24B20184, and 62373118), and the Heilongjiang Provincial Natural Science Foundation of China (No. ZD2023F002).

References

Ashraf, H.; Jeong, Y.; and Lee, C. H. 2021. Underwater ambient-noise removing GAN based on magnitude and phase spectra. *IEEE Access*, 9: 24513–24530.

Ashraf, H.; Shah, B.; Soomro, A. M.; Halim, Z.; Shah, S. K.; et al. 2022. Ambient-noise free generation of clean underwater ship engine audios from hydrophones using generative adversarial networks. *Computers and Electrical Engineering*, 100: 107970.

Bunne, C.; Hsieh, Y.-P.; Cuturi, M.; and Krause, A. 2023. The schrödinger bridge between gaussian measures has a closed form. In *International Conference on Artificial Intelligence and Statistics*, 5802–5833. PMLR.

Calvo, M. P.; and Palencia, C. 2006. A class of explicit multistep exponential integrators for semilinear problems. *Numerische Mathematik*, 102: 367–381.

Chen, T.; Liu, G.-H.; and Theodorou, E. A. 2021. Likelihood training of schrödinger bridge using forward-backward sdes theory. *arXiv preprint arXiv:2110.11291*.

Chen, Z.; He, G.; Zheng, K.; Tan, X.; and Zhu, J. 2023. Schrödinger bridges beat diffusion models on text-to-speech synthesis. *arXiv preprint arXiv:2312.03491*.

Cong, Y.; Gu, C.; Zhang, T.; and Gao, Y. 2021. Underwater robot sensing technology: A survey. *Fundamental Research*, 1(3): 337–345.

Duan, Y.; Shen, X.; and Wang, H. 2022. Time-domain anti-interference method for ship radiated noise signal. *EURASIP Journal on Advances in Signal Processing*, 2022(1): 65.

Gong, K.; Johnson, K.; El Fakhri, G.; Li, Q.; and Pan, T. 2024. PET image denoising based on denoising diffusion probabilistic model. *European Journal of Nuclear Medicine and Molecular Imaging*, 51(2): 358–368.

Han, Z.; Zhang, B.; Zhang, L.; Feng, S.; Lin, K.; Liang, G.; and Ye, Y. 2025. AsyncDSB: Schedule-Asynchronous Diffusion Schrödinger Bridge for Image Inpainting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3374–3382.

Hochbruck, M.; Ostermann, A.; and Schweitzer, J. 2009. Exponential Rosenbrock-type methods. *SIAM Journal on Numerical Analysis*, 47(1): 786–803.

Jensen, F. B.; Kuperman, W. A.; Porter, M. B.; Schmidt, H.; and Tolstoy, A. 2011. *Computational ocean acoustics*, volume 2011. Springer.

Ju, D.; Chi, C.; Li, Z.; Li, Y.; Zhang, C.; and Huang, H. 2022. Deep-learning-based line enhancer for passive sonar systems. *IET Radar, Sonar & Navigation*, 16(3): 589–601.

Jukić, A.; Korostik, R.; Balam, J.; and Ginsburg, B. 2024. Schrödinger Bridge for Generative Speech Enhancement. *arXiv preprint arXiv:2407.16074*.

Kimura, R.; Nakatani, T.; Kamo, N.; Marc, D.; Araki, S.; Ueda, T.; and Makino, S. 2024. Diffusion model-based MIMO speech denoising and dereverberation. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 455–459. IEEE.

Koh, S.; Chia, C. S.; and Tan, B. A. 2020. Underwater signal denoising using deep learning approach. In *OCEANS*, 1–6. IEEE.

Kulikov, V.; Yadin, S.; Kleiner, M.; and Michaeli, T. 2023. Sinddm: A single image denoising diffusion model. In *International conference on machine learning*, 17920–17930. PMLR.

Li, Y.; Li, Y.; Chen, X.; Yu, J.; Yang, H.; and Wang, L. 2018. A new underwater acoustic signal denoising technique based on CEEMDAN, mutual information, permutation entropy, and wavelet threshold denoising. *Entropy*, 20(8): 563.

Lu, B.; Miao, Q.; Liu, Y.; Tamir, T. S.; Zhao, H.; Zhang, X.; Lv, Y.; and Wang, F.-Y. 2025. A Diffusion Model for Traffic Data Imputation. *IEEE/CAA Journal of Automatica Sinica*, 12(3): 606–617.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35: 5775–5787.

Luo, Y.; and Mesgarani, N. 2018. Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 696–700. IEEE.

Ou, H.; Allen, J. S.; and Syrmos, V. L. 2011. Frame-based time-scale filters for underwater acoustic noise reduction. *IEEE Journal of Oceanic Engineering*, 36(2): 285–297.

Santos-Domínguez, D.; Torres-Guijarro, S.; Cardenal-López, A.; and Pena-Gimenez, A. 2016. ShipsEar: An underwater vessel noise database. *Applied Acoustics*, 113: 64–69.

Schrödinger, E. 1932. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. In *Annales de l'institut Henri Poincaré*, volume 2, 269–310.

Song, Y.; Liur, F.; and Shen, T. 2022. Method of underwater acoustic signal denoising based on dual-path transformer network. *IEEE Access*.

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

Veeraiyan, V.; Velayutham, R.; and Philip, M. M. 2013. Frequency domain based approach for denoising of underwater acoustic signal using EMD. *Journal of Intelligent Systems*, 22(1): 67–80.

Vincent, E.; Gribonval, R.; and Févotte, C. 2006. Performance measurement in blind audio source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 14(4): 1462–1469.

Duan, Y.; Shen, X.; and Wang, H. 2023. Dual-branch time–frequency domain anti-interference method for ship radiated noise signal. *Ocean Engineering*, 279: 114537.

Xing, C.; Wu, Y.; Xie, L.; and Zhang, D. 2021. A sparse dictionary learning-based denoising method for underwater acoustic sensors. *Applied Acoustics*, 180: 108140.

Yuan, Y.; Yang, G.; Wang, J. Z.; Zhang, H.; Shan, H.; Wang, F.-Y.; and Zhang, J. 2025. Dissecting and Mitigating Semantic Discrepancy in Stable Diffusion for Image-to-Image Translation. *IEEE/CAA Journal of Automatica Sinica*, 12(4): 1–14.

Zhang, W.; Li, X.; Zhou, A.; Ren, K.; and Song, J. 2021. Underwater acoustic source separation with deep Bi-LSTM networks. In *International Conference on Information Communication and Signal Processing*, 254–258. IEEE.

Zhou, A.; Zhang, W.; Li, X.; Xu, G.; Zhang, B.; Ma, Y.; and Song, J. 2023a. A novel noise-aware deep learning model for underwater acoustic denoising. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–13.

Zhou, A.; Zhang, W.; Xu, G.; Li, X.; Deng, K.; and Song, J. 2023b. DBSA-net: Dual branch self-attention network for underwater acoustic signal denoising. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 1851–1865.

Zhou, M.; Wang, J.; Feng, X.; Sun, H.; Li, J.; and Kuai, X. 2021. On generative-adversarial-network-based underwater acoustic noise modeling. *IEEE Transactions on Vehicular Technology*, 70(9): 9555–9559.

Zhu, P.; Gao, L.; Zhang, Y.; and Huang, Y. 2025. RGDNet: Recognition-Guided Underwater Acoustic Signal Denoising via Mask Integration and Signal Decoupling. *IEEE Transactions on Geoscience and Remote Sensing*.

Zhu, P.; Zhang, Y.; Huang, Y.; Zhao, C.; Zhao, K.; and Zhou, F. 2023. Underwater acoustic target recognition based on spectrum component analysis of ship radiated noise. *Applied acoustics*, 211: 109552.