Rethinking Flow and Diffusion Bridge Models for Speech Enhancement

Dahan Wang[1,2], Jun Gao[1,2], Tong Lei[3], Yuxiang Hu[2], Changbao Zhu[2], Kai Chen[1,2], and Jing Lu[*1,2*]

[1]Key Laboratory of Modern Acoustics, Institute of Acoustics, Nanjing University, Nanjing, China
[2]NJU-Horizon Intelligent Audio Lab, Horizon Robotics, Beijing, China
[3]Tencent AI Lab, Shenzhen, China
dahan.wang@smail.nju.edu.cn, lujing@nju.edu.cn

Abstract

Flow matching and diffusion bridge models have emerged as leading paradigms in generative speech enhancement, modeling stochastic processes between paired noisy and clean speech signals based on principles such as flow matching, score matching, and Schrödinger bridge. In this paper, we present a framework that unifies existing flow and diffusion bridge models by interpreting them as constructions of Gaussian probability paths with varying means and variances between paired data. Furthermore, we investigate the underlying consistency between the training/inference procedures of these generative models and conventional predictive models. Our analysis reveals that each sampling step of a well-trained flow or diffusion bridge model optimized with a data prediction loss is theoretically analogous to executing predictive speech enhancement. Motivated by this insight, we introduce an enhanced bridge model that integrates an effective probability path design with key elements from predictive paradigms, including improved network architecture, tailored loss functions, and optimized training strategies. Experiments on denoising and dereverberation tasks demonstrate that the proposed method outperforms existing flow and diffusion baselines with fewer parameters and reduced computational complexity. The results also highlight that the inherently predictive nature of this generative framework imposes limitations on its achievable upper-bound performance.

Appendix, code, and audio samples —
https://github.com/Dahan-Wang/Rethinking-Flow-and-Diffusion-Bridge-Models-for-Speech-Enhancement

1 Introduction

Deep learning-based methods have achieved remarkable success in speech enhancement (SE), which aims to recover clean speech from noisy observations. These methods can be broadly categorized into predictive (discriminative) and generative frameworks. Predictive models (Yin et al. 2020; Zheng et al. 2021) learn a direct mapping from noisy signals to clean speech, whereas generative methods model the distribution of clean speech conditioned on its noisy counterpart. Recently, various generative paradigms have been extensively explored, including generative adversarial networks (GANs) (Fu et al. 2019), variational autoencoders (VAEs) (Fang et al. 2021), self-supervised learning (SSL) models (Wang et al. 2024), and diffusion models (Tai et al. 2023a; Lei et al. 2024; Richter and Gerkmann 2024; Liu, Wang, and Plumbley 2024; Li, Sun, and Angelov 2025). These generative approaches consistently demonstrate promising performance and robust generalization across diverse unseen acoustic scenarios.

In flow and diffusion-based models, SE is naturally formulated as a conditional generation task (Tai et al. 2023b). One of the most widely adopted paradigms is to introduce noisy speech into the conditional probability path. Task-adapted score-based diffusion models (Lemercier et al. 2025) achieve this by designing the drift term of stochastic differential equations (SDEs) based on either the Ornstein-Uhlenbeck (OU) process (Richter et al. 2023) or the Brownian bridge (BB) (Lay et al. 2023), resulting in diffusion processes with means interpolating between clean and corrupted signals. More recently, the tractable Schrödinger bridge (SB) framework (Chen et al. 2023), which is also referred to as the denoising diffusion bridge model (DDBM) (He et al. 2024), has been proposed to build stochastic processes between Dirac noisy and clean data endpoints by optimizing path measures under boundary constraints. The SB model also incorporates a data prediction training strategy, achieving state-of-the-art (SOTA) performance compared to conventional diffusion models (Jukić et al. 2024). Additionally, the flow matching (FM) method has been extended to incorporate probability paths conditioned on noisy speech, enabling efficient sampling while maintaining strong performance (Korostik, Nasretdinov, and Jukić 2025; Lee et al. 2025). These works have become the foundational basis for numerous advances in generative SE (Lemercier et al. 2023; Lay et al. 2024; Richter, De Oliveira, and Gerkmann 2025).

The aforementioned methods are grounded in distinct theoretical foundations, including score-based diffusion, Schrödinger bridge, and flow matching, which have yet to be unified under a common framework in the SE field. Additionally, the use of data prediction objectives (Chen et al. 2023) suggests their resemblance to predictive methods, which similarly estimate clean speech by implicitly learning distributional mappings between paired data. This connection, however, remains underexplored in prior work.

*Jing Lu is the corresponding author.

In this paper, we present a unified framework for flow and diffusion bridge models in SE, interpreting them as constructing different Gaussian probability paths between paired noisy and clean data. Then the sampling ordinary differential equations (ODEs) are derived through conditional flow matching, and extended to SDEs for both forward and backward processes. Notably, we show that all such models can be trained using a data prediction strategy. The fundamental difference among them lies in the design of mean and variance trajectories. Our analysis further reveals that each sampling step in a well-trained flow matching or diffusion bridge model is theoretically equivalent to predictive SE, and the final output is a weighted sum of these step-wise predictions. This suggests that these models, while generative in form, fundamentally operate as predictive models—explaining their effectiveness in single-step sampling and highlighting opportunity for improvement via predictive techniques.

Motivated by the above insights, we propose an enhanced bridge model that combines an effective probability path design with key strengths of the predictive paradigm. Specifically, we adopt a high-performance backbone (Wang et al. 2023), and introduce a time embedding mechanism to effectively leverage the information encoded in the diffusion time. Moreover, we refine the data prediction loss to optimize the model training, and integrate a fine-tuning strategy (Lay et al. 2024) for further performance gain. Experimental results reveal that the proposed model outperforms SOTA flow matching- and diffusion-based baselines while incurring markedly fewer parameters and reduced computational overhead. Furthermore, our findings highlight an upper-bound performance constraint imposed by the predictive nature of such generative frameworks.

Our main contributions are summarized as follows:

- **Unified Generative Framework:** We present a unified theoretical framework that encompasses existing flow and diffusion bridge models between paired data, including score-based diffusion, Schrödinger bridge, and flow matching, which are widely used generative approaches in SE.
- **Predictive Equivalence Insight:** We investigate the inherent equivalence between flow matching/diffusion bridge models and predictive methods, showing that these generative models share key mechanisms with predictive models. This insight provides practical guidance for model improvement and suggests that the predictive nature of such generative models may impose a ceiling on their performance.
- **Enhanced Bridge Model:** Our proposed enhanced bridge model incorporates advanced predictive strategies. Our model achieves significantly better performance and efficiency compared to existing flow and diffusion baselines.

2 Related Work

2.1 Score-based Diffusion Models

Score-based generative models (Welker, Richter, and Gerkmann 2022; Richter et al. 2023) describe the forward diffusion process through the forward SDE:

$$\mathrm{d}\mathbf{x}_t = \mathbf{f}_t(\mathbf{x}_t, \mathbf{y})\mathrm{d}t + g_t\mathrm{d}\mathbf{w}_t, \tag{1}$$

where $t \in [0,1]$ denotes a continuous time variable, $\mathbf{x}_t \in \mathbb{C}^{F\times L}$ represents the state of the process, i.e. the reshaped spectral coefficient vector with F frequency bins and L frames, $\mathbf{y}$ is the noisy speech vector, $\mathbf{w}_t$ is a standard Wiener process, $\mathbf{f}_t(\cdot,\cdot)$ is the drift term, and g_t is a scalar-valued diffusion coefficient. The initial condition of $\mathbf{x}_t$ is the clean speech $\mathbf{s}$. For the OU process, the drift term is defined as $\mathbf{f}_t(\mathbf{x}_t, \mathbf{y}) = \gamma(\mathbf{y} - \mathbf{x}_t)$, where γ is the stiffness coefficient. For the BB process, the drift term is $\mathbf{f}_t(\mathbf{x}_t, \mathbf{y}) = \frac{\mathbf{y}-\mathbf{x}_t}{1-t}$. For the diffusion coefficient, the variance-exploding (VE) schedule is commonly adopted, i.e., $g_t = \sqrt{ck^t}$. The combinations of these drift and diffusion terms are referred to as Ornstein-Uhlenbeck with variance exploding (OUVE) (Richter et al. 2023) and Brownian bridge with exponential diffusion coefficient (BBED) (Lay et al. 2023), respectively. The reverse SDE and its corresponding probability flow ODE (PFODE) are respectively given by

$$\mathrm{d}\mathbf{x}_t = \left[\mathbf{f}_t(\mathbf{x}_t, \mathbf{y}) - g_t^2\nabla_{\mathbf{x}_t}\log p_t(\mathbf{x}_t|\mathbf{s},\mathbf{y})\right]\mathrm{d}t + g_t\mathrm{d}\bar{\mathbf{w}}_t, \tag{2}$$

$$\mathrm{d}\mathbf{x}_t = \left[\mathbf{f}_t(\mathbf{x}_t, \mathbf{y}) - \frac{1}{2}g_t^2\nabla_{\mathbf{x}_t}\log p_t(\mathbf{x}_t|\mathbf{s},\mathbf{y})\right]\mathrm{d}t, \tag{3}$$

where $\mathrm{d}t$ represents a negative infinitesimal time step, $\bar{\mathbf{w}}_t$ is the reverse-time Wiener process, $p_t(\mathbf{x}_t|\mathbf{s},\mathbf{y})$ denotes the conditional probability path (or perturbation kernel), and $\nabla_{\mathbf{x}_t}\log p_t(\mathbf{x}_t|\mathbf{s},\mathbf{y})$ is the corresponding score function. The probability path has a Gaussian form defined by

$$p_t(\mathbf{x}_t|\mathbf{s},\mathbf{y}) = \mathcal{N}\left(\mathbf{x}_t; \boldsymbol{\mu}_t(\mathbf{s},\mathbf{y}), \sigma_t^2\mathbf{I}\right), \tag{4}$$

with its mean and variance determined by $\mathbf{f}_t$ and g_t. The score function can be obtained via denoising score matching:

$$\nabla_{\mathbf{x}_t}\log p_t(\mathbf{x}_t|\mathbf{s},\mathbf{y}) = -\frac{\mathbf{x}_t - \boldsymbol{\mu}_t}{\sigma_t^2}, \tag{5}$$

which is the training objective of the backbone network.

2.2 Schrödinger Bridge

The SB problem originates from the optimization of path measures with constrained boundaries. For dual Dirac distribution boundaries centered on paired clean and noisy speech, the SB solution can be expressed as a couple of forward-backward SDEs (Chen et al. 2023):

$$\mathrm{d}\mathbf{x}_t = \left[f_t\mathbf{x}_t - g_t^2\frac{\mathbf{x}_t - \bar{\alpha}_t\mathbf{y}}{\alpha_t^2\bar{\rho}_t^2}\right]\mathrm{d}t + g_t\mathrm{d}\mathbf{w}_t, \tag{6}$$

$$\mathrm{d}\mathbf{x}_t = \left[f_t\mathbf{x}_t + g_t^2\frac{\mathbf{x}_t - \alpha_t\mathbf{s}}{\alpha_t^2\rho_t^2}\right]\mathrm{d}t + g_t\mathrm{d}\bar{\mathbf{w}}_t, \tag{7}$$

with the corresponding probability path defined as

$$p_t(\mathbf{x}_t|\mathbf{s},\mathbf{y}) = \mathcal{N}\left(\frac{\alpha_t\bar{\rho}_t^2\mathbf{s} + \bar{\alpha}_t\rho_t^2\mathbf{y}}{\rho_1^2}, \frac{\alpha_t^2\bar{\rho}_t^2\rho_t^2}{\rho_1^2}\mathbf{I}\right), \tag{8}$$

and the PFODE formulated as

$$\mathrm{d}\mathbf{x}_t = \left[f_t\mathbf{x}_t - \frac{1}{2}g_t^2\frac{\mathbf{x}_t - \bar{\alpha}_t\mathbf{y}}{\alpha_t^2\bar{\rho}_t^2} + \frac{1}{2}g_t^2\frac{\mathbf{x}_t - \alpha_t\mathbf{s}}{\alpha_t^2\rho_t^2}\right]\mathrm{d}t, \tag{9}$$

where f_t is the drift coefficient, $\alpha_t = \exp\left(\int_0^t f_\tau \mathrm{d}\tau\right)$, $\rho_t^2 = \int_0^t g_\tau^2 \alpha_\tau^{-2} \mathrm{d}\tau$, $\bar{\alpha}_t = \alpha_t \alpha_1^{-1}$, and $\bar{\rho}_t^2 = \rho_1^2 - \rho_t^2$. This set of formulations can serve as a unified framework for all DDBMs between paired data (He et al. 2024). For SE, a data prediction training strategy is widely adopted due to its performance advantages over score matching, that is, the network directly predicts the clean speech $\mathbf{s}$. Moreover, VE is the most commonly used schedule in SE, defined by setting $f_t = 0$ and $g_t = \sqrt{c}k^t$, which is referred to as SBVE (Jukić et al. 2024). In this paper, the SB model and the score-based diffusion models introduced in the previous subsection are collectively referred to as diffusion bridge models.

2.3 Flow Matching

A flow matching method for SE is defined by an ODE:

$$\mathrm{d}\mathbf{x}_t = \mathbf{u}_t(\mathbf{x}_t|\mathbf{s},\mathbf{y})\mathrm{d}t, \tag{10}$$

where $\mathbf{u}_t(\mathbf{x}_t|\mathbf{s},\mathbf{y})$ denotes the conditional vector field. Unlike diffusion models, the sampling process in flow models proceeds forward in time, with $t = 1$ corresponding to the target data distribution. Restricting $\mathbf{x}_t$ to follow a Gaussian probability path, the conditional vector can be derived as

$$\mathbf{u}_t(\mathbf{x}_t|\mathbf{s},\mathbf{y}) = \frac{\sigma_t'}{\sigma_t}(\mathbf{x}_t - \boldsymbol{\mu}_t) + \boldsymbol{\mu}_t'. \tag{11}$$

For SE tasks with paired clean and noisy data, following the optimal transport conditional FM (OT-CFM), the mean and variance of the probability path are set to $\boldsymbol{\mu}_t(\mathbf{s},\mathbf{y}) = (1-t)\mathbf{y} + t\mathbf{s}$ and $\sigma_t = (1-t)\sigma_{\max} + t\sigma_{\min}$, respectively (Korostik, Nasretdinov, and Jukić 2025; Lee et al. 2025).

3 Methodology

3.1 A Unified Framework for Flow and Diffusion Bridge Models

Framework We define the probability path in Gaussian form, as given in Eq. (4), with the mean specified as

$$\boldsymbol{\mu}_t(\mathbf{x}_t|\mathbf{s},\mathbf{y}) = a_t\mathbf{s} + b_t\mathbf{y}. \tag{12}$$

which interpolates between the clean and noisy speech. Based on Eq. (11), the corresponding ODE is derived as

$$\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = \frac{\sigma_t'}{\sigma_t}\mathbf{x}_t + \left(a_t' - a_t\frac{\sigma_t'}{\sigma_t}\right)\mathbf{s} + \left(b_t' - b_t\frac{\sigma_t'}{\sigma_t}\right)\mathbf{y}. \tag{13}$$

Following the SDE extension trick based on the Fokker-Planck equation (Holderrieth and Erives 2025), the associated forward-backward SDEs are formulated as

$$\mathrm{d}\mathbf{x}_t = [\kappa_t^+\mathbf{x}_t + (a_t' - a_t\kappa_t^+)\,\mathbf{s} + (b_t' - b_t\kappa_t^+)\,\mathbf{y}]\mathrm{d}t + g_t\mathrm{d}\mathbf{w}_t, \tag{14}$$

$$\mathrm{d}\mathbf{x}_t = [\kappa_t^-\mathbf{x}_t + (a_t' - a_t\kappa_t^-)\,\mathbf{s} + (b_t' - b_t\kappa_t^-)\,\mathbf{y}]\mathrm{d}t + g_t\mathrm{d}\bar{\mathbf{w}}_t, \tag{15}$$

where

$$\kappa_t^\pm = \frac{\sigma_t'}{\sigma_t} \mp \frac{g_t^2}{2\sigma_t^2}. \tag{16}$$

The detailed derivation is provided in Appendix A.1.

Based on the above framework, we interpret the core design principle of flow and diffusion bridge models as the construction of conditional probability paths between paired data, specifically through the design of a_t, b_t, and σ_t. Once the probability path is specified, the corresponding sampling equations can be directly obtained via Eqs. (13)-(15). This set of unified formulations enables a consistent description of various SE generative models without the need to start from the design of forward SDEs, as in score-based diffusion models, or to solve Kullback-Leibler-divergence optimization and partial differential equations, as required by the SB method. The parameters defining the probability paths in representative models are summarized in Table 1. Detailed proofs of how these models are derived from our framework are provided in Appendix A.2.

Method	a_t	b_t	σ_t
OUVE	$e^{-\gamma t}$	$1 - e^{-\gamma t}$	$\frac{c\left(k^{2t} - \mathrm{e}^{-2\gamma t}\right)}{2(\gamma + \log k)}$
BBED	$1 - t$	t	$c(1-t)E_t$ *
SB	$\alpha_t\bar{\rho}_t^2/\rho_1^2$	$\bar{\alpha}_t\rho_t^2/\rho_1^2$	$\alpha_t^2\bar{\rho}_t^2\rho_t^2/\rho_1^2$
OT-CFM	t	$1 - t$	$(1-t)\sigma_{\max} + t\sigma_{\min}$

* $E_t = (k^{2t} - 1 + t) + \log(k^{2k^2})\{\mathrm{Ei}\left[2(t-1)\log k\right] - \mathrm{Ei}\left[-2\log k\right]\}(1-t)$, where $\mathrm{Ei}[\cdot]$ denotes the exponential integral function (Bender and Orszag 2013).

Table 1: Probability path parameters of representative flow and diffusion bridge models for SE.

Diffusion Coefficient and Sampling Direction Note that there are two important issues regarding the forward-backward SDEs that require further clarification. First, to derive the SDEs, the form of the diffusion coefficient g_t must be specified. Theoretically, g_t can be arbitrary, meaning that a single probability path may correspond to a family of SDEs with different diffusion coefficients. This is because, according to the Fokker-Planck equation, the effects of g_t on the drift and diffusion terms cancel out, preserving the same underlying probability path (Holderrieth and Erives 2025). In previous diffusion bridge models, the designed SDEs represent a specific, tractable case within this broader family with arbitrary g_t. The g_t defined in these models is related to σ_t, and this relationship can be used to simplify the form of the resulting ODE and SDEs.

Second, our framework does not impose a fixed temporal direction for sampling. Instead, the direction is determined by the definitions of the path parameters. Typically, the sampling process starts at a point with mean $\mathbf{y}$ and ends at a point with mean $\mathbf{s}$ and zero variance. However, the assignment of these conditions to $t = 0$ or $t = 1$ is not fixed in advance, which is governed by the definitions of a_t, b_t, and σ_t. For diffusion bridge models, the sampling process proceeds in reverse time, meaning that the backward SDE (Eq. (14)) is used for sampling. In contrast, for flow matching models, the sampling proceeds in forward time; even when extended to the SDE form, the forward SDE is used for sampling.

Training and Sampling According to Eqs. (13)-(15), the only unknown term during the sampling process is the clean speech $\mathbf{s}$. Therefore, the network can be trained using a data prediction strategy, where the clean speech $\mathbf{s}$ serves as the training target. $\mathbf{s}$ in these equations is replaced by the net-

work's output during sampling. This strategy is particularly advantageous for SE tasks, as it allows the incorporation of auxiliary losses tailored to the characteristics of speech signals (Chen et al. 2023; Richter, De Oliveira, and Gerkmann 2025). Moreover, our framework enables the application of data prediction training to OUVE and BBED, which originally rely on score matching for optimization.

We recommend using a discretization method based on exponential integrators for sampling, as it introduces minimal discretization error (Chen et al. 2023; He et al. 2024). For simplicity, we rewrite the ODE presented in Eq. (13) as $\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = \frac{\sigma_t'}{\sigma_t}\mathbf{x}_t + m_t\mathbf{s} + n_t\mathbf{y}$, which enables the corresponding discretized sampling equation to be expressed as

$$\mathbf{x}_t = \frac{\sigma_t}{\sigma_r}\mathbf{x}_r + \sigma_t\left(\int_r^t \frac{m_\tau}{\sigma_\tau}\mathrm{d}\tau\right)\mathbf{s} + \sigma_t\left(\int_r^t \frac{n_\tau}{\sigma_\tau}\mathrm{d}\tau\right)\mathbf{y}. \quad (17)$$

However, for certain models with complex parameterizations (such as OUVE and BBED), the integral in this expression may not yield a tractable closed-form solution, making the exponential integrator method difficult to apply to these methods. The detailed derivation and further discussion are provided in Appendix A.3.

A Simple and Effective Parameterization Based on our framework, we show a simple and effective parameter configuration: $a_t = 1 - t, b_t = t, \sigma_t^2 = \sigma^2 t(1-t)$. Its corresponding sampling ODE can be derived from Eq. (13) as

$$\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = \frac{1-2t}{2t(1-t)}\mathbf{x}_t - \frac{1}{2t}\mathbf{s} + \frac{1}{2(1-t)}\mathbf{y}. \quad (18)$$

This formulation is known as Brownian bridge (BB) (He et al. 2024) or Schrödinger bridge-conditional flow matching (SB-CFM) (Tong et al. 2023), a special case of the SB parameterization listed in Table 1 with $\alpha_t = 1, \rho_t^2 = \sigma^2 t$.

3.2 Predictive Properties of Flow and Diffusion Bridge Models

Predictive Behavior in the Network's Functioning Flow matching and diffusion bridge models construct probability paths between data pairs. This contrasts with conventional flow and diffusion models, which typically learn mappings between entire distributions, transforming random samples from a source distribution into samples from a target distribution. Predictive models for SE, by comparison, can be interpreted as implicitly modeling a single-step transition between Dirac distributions centered on the paired data. This perspective aligns with the core objective of the generative models discussed in this paper, highlighting a similarity between these generative approaches and predictive models in terms of their overall processing framework.

Figure 1 illustrates the working mechanism of the backbone network of flow and diffusion bridge models during training and sampling under the data prediction strategy. The network takes as input the state $\mathbf{x}_t$, the noisy signal $\mathbf{y}$, and the time variable t, and outputs the enhanced speech. Compared to a standard predictive SE model, two additional inputs, $\mathbf{x}_t$ and t, are introduced. The state $\mathbf{x}_t$ follows the Gaussian distribution with mean $\boldsymbol{\mu}_t$ and variance σ_t^2, where $\boldsymbol{\mu}_t$ is

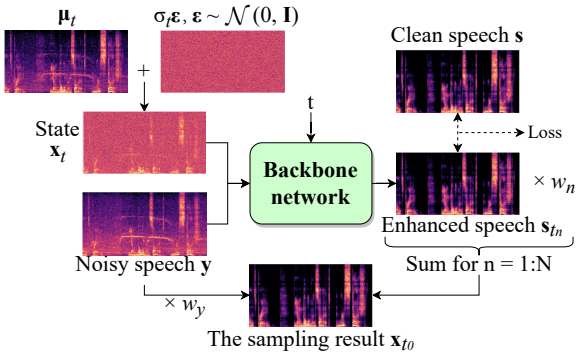

Figure 1: Illustration of the backbone network's working mechanism during training and ODE-based sampling (as expressed in Eq. (20)) under the data prediction strategy.

an interpolation between the clean and noisy speech. This makes the mean of $\mathbf{x}_t$ equivalent to a noisy signal with a relatively higher signal-to-noise ratio (SNR). As sampling proceeds, the SNR of $\boldsymbol{\mu}_t$ increases gradually and eventually approaches that of the clean speech. The diffusion time t encodes this SNR progression as well as the level of the variance. Therefore, the backbone network can be viewed as a predictive SE model augmented with auxiliary information. This reveals a strong alignment between the working mechanism of these generative model and conventional predictive models.

Analysis of Sampling Result Composition We analyze the composition of the final sampling result by examining the first-order discretized sampling equation based on the exponential integrator. Specifically, we adopt the first-order discretization of the ODE given in Eq. (17) and, following the diffusion bridge models, perform sampling in the reverse time direction. For clarity, we rewrite Eq. (17) as $\mathbf{x}_t = \xi(t,r)\mathbf{x}_r + \eta(t,r)\mathbf{s} + \zeta(t,r)\mathbf{y}$. Substituting the discretized time steps $t = t_n, r = t_{n+1}$, denoting that $\theta(t_n, t_{n+1}) = \theta_n, \theta = \xi, \eta, \zeta$, and replacing the clean speech $\mathbf{s}$ with the network output $\mathbf{s}_{t_{n+1}}$ at each step, the sampling equation can be rewritten as

$$\mathbf{x}_{t_n} = \xi_n\mathbf{x}_{t_{n+1}} + \eta_n\mathbf{s}_{t_{n+1}} + \zeta_n\mathbf{y}, \mathbf{x}_{t_N} = \mathbf{y}. \quad (19)$$

The final sampling result can then be expressed as

$$\mathbf{x}_{t_0} = \sum_{n=1}^{N}(w_n\mathbf{s}_{t_n}) + w_y\mathbf{y}, \quad (20)$$

where

$$w_n = \tilde{\xi}_{n-2}\eta_{n-1}, w_y = \sum_{n=1}^{N+1}\tilde{\xi}_{n-2}\zeta_{n-1}, \quad (21)$$

with $\tilde{\xi}_n = \prod_{k=1}^{n}\xi_k, n \geq 0, \tilde{\xi}_{-1} = 0$, and $\zeta_N = 1$. It is important to note that the sampling endpoint t_0 is typically set to a small positive value (e.g., 10^{-4}) to avoid numerical singularities. To obtain more specific results, we consider the parameterization of the SB model and apply discretization,

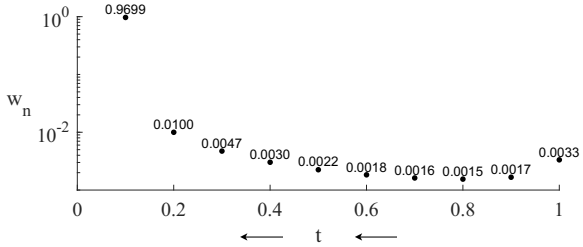

Figure 2: Weight distribution of network outputs at each step in ODE-based sampling result (SB-CFM parameterization, and $N = 10$). The arrows indicate that sampling proceeds in the reverse time direction.

obtaining

$$w_n = \frac{\alpha_0 \rho_0 \bar{\rho}_0}{\rho_N^2} \left(\frac{\bar{\rho}_{n-1}}{\rho_{n-1}} - \frac{\bar{\rho}_n}{\rho_n} \right), w_y = \frac{\alpha_0 \rho_0^2}{\alpha_N \rho_N^2}. \quad (22)$$

By substituting the specific parameter values of α_t, ρ_t, and $\bar{\rho}_t$, the exact values of these weights can be explicitly calculated. Detailed derivations and analyses of the above formulas are provided in Appendix A.4.

Eq. (20) reveals that the final sampling result is a weighted combination of the network's clean speech estimates at each step and the noisy signal $\mathbf{y}$, with the weights determining their respective contributions. Fig. 1 provides an intuitive illustration of this weighted combination. Using the SB-CFM parameterization (set $\sigma = 1$) described above, we perform numerical simulations on the weights defined in Eq. (22). Specifically, we set the number of sampling steps $N = 10$, obtaining $w_y = 10^{-4}$, and the weights w_n at each step are plotted in Fig. 2. The simulation results indicate that the final output is largely dominated by the network's estimate at the last step, while the contributions from earlier steps and the noisy input $\mathbf{y}$ are negligible. Note that if the network's outputs at each step do not outperform those of traditional predictive models, the SE tasks may not gain a substantial advantage from adopting this generative framework.

Furthermore, it is important to emphasize that one-step sampling is nearly equivalent to a predictive model. Its output relies entirely on a single model call based on data prediction, without leveraging information from intermediate states $\mathbf{x}_t$. In this case, training is only meaningful at $t = 1$, while training at other time steps becomes redundant and offers no meaningful contribution to performance.

3.3 Improved Bridge Model for Speech Enhancement Incorporating Predictive Paradigms

In the previous section, we analyze the underlying consistency between flow/diffusion bridge models and predictive SE methods. Motivated by this insight, we propose a series of improvements applicable to the flow and diffusion bridge models described by our unified framework. Given the demonstrated advantages of SB parameterization in prior studies, we integrate these enhancements with the SB model to construct an improved bridge model.

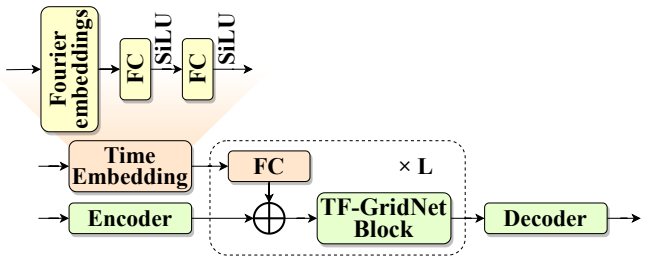

Figure 3: Schematic illustration of the time-embedding-assisted TF-GridNet.

Improved Backbone Network We integrate TF-GridNet (Wang et al. 2023), a SOTA predictive SE model, as the backbone network in the generative framework, replacing the commonly used U-Net architectures such as Noise Conditional Score Network (NCSN++) (Song et al. 2021). TF-GridNet is highly effective for speech estimation due to its ability to capture correlations between subbands and frames. However, the original TF-GridNet architecture cannot directly accept diffusion time as an input.

To leverage the information in the diffusion time t, we introduce a time-embedding mechanism to make TF-GridNet time-dependent. As illustrated in Fig. 1, the diffusion time is first projected into a high-dimensional vector using a time embedding module, which consists of Fourier embeddings followed by fully connected (FC) layers with sigmoid linear unit (SiLU) activation functions (Elfwing, Uchibe, and Doya 2018). The resulting time embedding vector is then incorporated into each TF-GridNet block. Specifically, it is processed by a dedicated FC layer and added to the input features at the start of each TF-GridNet block.

Improved Loss Function In previous studies, the data prediction loss for diffusion models is generally defined as a combination of MSE loss on the complex spectrogram, time-domain L1 loss, and PESQ loss. However, these configurations may underemphasize the importance of spectral amplitude and over-optimize PESQ. Therefore, inspired by predictive SE models, we introduce the negative SI-SNR (Le Roux et al. 2019) loss and the power-compressed spectrum loss into the SB-based diffusion model, defined as

$$\mathcal{L}_{\text{SI-SNR}}(\hat{x}, x) = -\log_{10}\left(\frac{\|x_t\|^2}{\|\hat{x} - x_t\|^2} \right), x_t = \frac{\langle \hat{x}, x \rangle x}{\|x\|^2}, \quad (23)$$

$$\mathcal{L}_{\text{mag}}\left(\hat{X}, X\right) = \text{MSE}\left(|\hat{X}|^{0.3}, |X|^{0.3} \right), \quad (24)$$

$$\mathcal{L}_{\text{real/imag}}\left(\hat{X}, X\right) = \text{MSE}\left(\frac{\hat{X}_{\text{r/i}}}{|\hat{X}|^{0.7}}, \frac{X_{\text{r/i}}}{|X|^{0.7}} \right), \quad (25)$$

where x and $\hat{x}$ represent clean and enhanced waveforms, X and $\hat{X}$ are their corresponding spectrograms, the subscripts r, i represent the real and imaginary parts of the spectrograms, respectively, $\langle \cdot, \cdot \rangle$ denotes the inner product operator, and $\text{MSE}(\cdot, \cdot)$ represents the mean squared error (MSE). The overall loss function for model training is given by

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{SI-SNR}} + \lambda_2 \mathcal{L}_{\text{mag}} + \lambda_3 \left(\mathcal{L}_{\text{real}} + \mathcal{L}_{\text{imag}} \right), \quad (26)$$

where $\lambda_1, \lambda_2, \lambda_3$ are the empirical weights.

Backbone	Loss	CRP	Schedule	Para. (M)	MACs (G)	SI-SNR	ESTOI	PESQ	DNSMOS	UTMOS
Noisy	-	-	-	-	-	5.613	0.669	1.406	2.147	1.476
NCSN++	Original	✗	SBVE	65.6	66×5	14.158	0.836	2.706	3.666	2.155
NCSN++	Improved	✗	SBVE	65.6	66×5	13.481	0.842	2.802	3.726	2.160
TF-GridNet	Improved	✗	SBVE	2.2	38×5	16.646	0.871	3.068	3.761	2.246
TF-GridNet	Improved	✓	SBVE	2.2	38×5	16.424	0.874	3.213	3.752	2.253
TF-GridNet	Improved	✗	OUVE	2.2	38×60	11.302	0.778	2.129	3.385	1.874
TF-GridNet	Improved	✗	BBED	2.2	38×60	14.429	0.843	2.800	3.691	2.133
TF-GridNet	Improved	✗	OT-CFM	2.2	38×5	14.866	0.851	2.834	3.385	2.168
TF-GridNet	Improved	✗	SB-CFM	2.2	38×5	16.177	0.867	3.102	3.742	2.216
TF-GridNet	Improved	✗	SBVE	2.2	38×5	16.646	0.871	3.068	3.761	2.246

Table 2: Ablation study results on DNS3 test set.

Incorporation of a Predictive Fine-tuning Strategy A fine-tuning method called correcting the reverse process (CRP) has been introduced into BBED to mitigate errors accumulated during the sampling process (Lay et al. 2024). CRP fine-tunes the score model by minimizing an MSE loss between the clean speech and the signal generated using the Euler-Maruyama (EuM) first-order sampling method. CRP only updates the model weights during the last model call. This fine-tuning strategy can be generalized to various flow and diffusion bridge models by replacing the EuM method with preferred sampling method, such as the exponential integrator-based approach. Moreover, the original MSE loss used in CRP can be replaced with our improved data prediction loss. It is worth emphasizing that updating weights only at the final step is consistent with our earlier finding that the last sampling step has the greatest influence on the final result and plays a dominant role in the model's overall performance.

4 Experiments

4.1 Experimental setup

Datasets and Implementation Details We conduct experiments on two datasets. The first dataset is constructed for both denoising and dereverberation tasks, using clean and noise samples from the 3rd Deep Noise Suppression Challenge (DNS3) dataset (Reddy et al. 2021). The second one is the standardized VoiceBank+DEMAND dataset (Valentini-Botinhao et al. 2016), which is widely used as a benchmark for SE. All utterances are downsampled from 48 kHz to 16 kHz. Details regarding hyperparameter settings, training configuration, evaluation metrics, and other implementation specifics are provided in Appendix B.

Baselines We compare the proposed model with several predictive and generative baselines. The predictive baselines include NCSN++ and TF-GridNet, both trained using the proposed loss function. The generative baselines include SGMSE+ (OUVE) (Richter et al. 2023), StoRM (Lemercier et al. 2023), BBED (Lay et al. 2023), SBVE (Jukić et al. 2024), and FlowSE (Lee et al. 2025). NCSN++ is used as the backbone of SGMSE+, StoRM, and SBVE, following the configuration in (Richter et al. 2023), resulting in approximately 65.6M parameters. The training and sampling configurations of the baselines follow those of the original papers.

4.2 Experimental Results

Ablation Study Results We validate the effectiveness of the proposed modifications on the DNS3 test set. As shown in Table 2, the ablation study demonstrates that the time-embedding-assisted TF-GridNet along with the improved data prediction loss significantly improves the overall performance of the bridge model, while substantially reducing the number of parameters and computational complexity compared to NCSN++. Additionally, the integration of CRP fine-tuning yields further performance gains without increasing inference cost.

Building on the improved backbone and loss function, we conduct ablation experiments to evaluate several probability path parameterizations, including OUVE, BBED, OT-CFM, SB-CFM, and SBVE, among which only SB-CFM has not been previously applied to SE. Notably, BBED, SB-CFM, and SBVE exhibit zero variance at the starting point of sampling, which corresponds to a Dirac distribution centered on the noisy input. However, due to the complex definitions of σ_t in OUVE and BBED, it is difficult to obtain tractable solutions for the exponential integrator-based samplers. Consequently, we follow the original implementations for OUVE and BBED, employing predictor-corrector (PC) samplers, which require more sampling steps to maintain performance. Experimental results show that SB-CFM and SBVE outperform the alternatives in SE tasks. Based on these findings, we adopt the SBVE schedule as the probability path in our improved bridge model. Overall, the results support the conclusion that Gaussian probability paths with Dirac endpoints, along with exponential integrator-based sampling, provide strong performance guarantees for flow and diffusion bridge models in SE.

Comparison with the Baseline Models The comparison results on the DNS3 test set are presented in Table 3. Compared with the predictive baselines, the proposed model with one-step sampling outperforms NCSN++ and achieves performance comparable to TF-GridNet, one of the current SOTA predictive models. With additional sampling steps, the proposed model slightly outperforms TF-GridNet across most metrics. This reinforces our earlier conclusion that one-step sampling under this generative framework is essentially equivalent to a predictive model. Furthermore, it significantly surpasses the generative baselines, especially the SOTA SBVE model, in terms of both performance and ef-

Model	Para. (M)	MACs (G)	SI-SNR	ESTOI	PESQ	DNSMOS	UTMOS
Noisy	-	-	5.613	0.669	1.406	2.147	1.476
NCSN++	59.6	66	14.146	0.842	2.673	3.747	2.182
TF-GridNet	2.1	38	16.448	0.872	3.187	3.743	2.236
SGMSE+	65.6	66×60	11.873	0.796	2.336	3.647	2.007
StoRM	65.6	$66 + 66 \times 60$	12.463	0.805	2.297	3.625	2.060
SBVE	65.6	66×60	14.959	0.844	2.592	3.729	2.208
Proposed (NFEs=1)	2.2	38×1	16.245	0.870	3.185	3.740	2.237
Proposed (NFEs=5)	2.2	38×5	16.424	0.874	3.213	3.752	2.253

Table 3: Performance on DNS3 test set.

Model	SI-SNR	ESTOI	PESQ	DNSMOS
Noisy	8.4	0.79	1.97	3.09
NCSN++	18.8	0.88	3.01	3.56
TF-GridNet	19.5	0.88	3.17	3.57
SGMSE+*	17.3	0.87	2.93	3.56
StoRM*	18.8	0.88	2.93	-
BBED*	18.8	0.88	3.09	3.57
SBVE*	19.4	0.88	2.91	3.59
FlowSE*	19.0	0.88	3.12	3.58
Proposed	19.6	0.89	3.30	3.57

* Metrics are provided by their original papers.

Table 4: Performance on Voicebank+DEMAND test set.

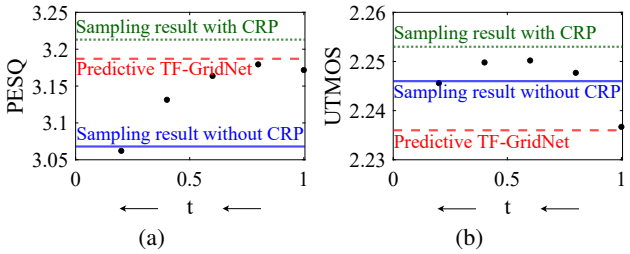

Figure 4: Average PESQ and UTMOS of network outputs at each step during sampling ($N = 5$) for the proposed bridge model (without CRP). Dots represent the scores of intermediate network outputs; lines indicate the metrics of the predictive TF-GridNet output and the final sampling results of the proposed bridge model with and without CRP fine-tuning. The arrows indicate that sampling is performed in the reverse time direction.

ficiency, requiring fewer sampling steps, fewer parameters, and lower computational complexity.

Table 4 presents results on the Voicebank+DEMAND test set, with scores for generative baselines taken from their original papers. The proposed model achieves SOTA performance across nearly all metrics, further validating the effectiveness of integrating predictive paradigms into diffusion models. These results also support the view that such generative models inherently exhibit predictive behavior.

Impact of Predictive Behavior on the Performance of Flow and Diffusion Bridge Models Based on our analysis of the inherent equivalence between flow matching/diffusion bridge models and predictive methods, we observe that the quality of the final sampling result is largely determined by the accuracy with which the network estimates the clean speech at each sampling step. Fig. 4 presents the average PESQ and UTMOS of the network outputs at each step ($N = 5$) for the proposed bridge model without CRP fine-tuning, along with the scores of the final sampling result. As previously discussed, the network output at the last step ($t = 0.2$) contributes most heavily to the final result, thus leading to nearly identical scores. Fig. 4 also includes the scores of enhanced outputs from the predictive TF-GridNet model, which closely match those of the network output at the first sampling step. This supports our earlier conclusion that at $t = 1$, where the network input consists solely of the noisy signal, the model behaves equivalently to a predictive system.

Furthermore, the scores at all sampling steps are comparable to those of the predictive model, indicating that this generative framework achieves strong denoising and dereverberation performance with each model call. However, this predictive-like behavior suggests an inherent upper bound on performance, that is, it may not significantly outperform its corresponding predictive model for SE tasks.

Additionally, we observe that during training, the final model call, which dominates the final sampling result, may be slightly under-optimized (lower PESQ than other steps). Fine-tuning this step using CRP compensates for this limitation and further enhances the overall performance of the bridge model.

5 Conclusion

In this paper, we present a unified theoretical framework that encompasses widely used generative approaches in SE, including score-based diffusion, Schrödinger bridge, and flow matching methods. We demonstrate that these flow and diffusion bridge models, although generative in form, share key mechanisms with predictive SE methods. This insight offers practical guidance for improving such models. Building on this finding, we propose an enhanced bridge model that integrates advanced predictive strategies. Our model achieves significantly better performance and efficiency than existing flow and diffusion baselines. Experimental results further suggest that the inherently predictive behavior of these generative models may impose an upper bound on their performance in denoising and dereverberation tasks.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 12274221), the Yangtze River Delta Science and Technology Innovation Community Joint Research Project (Grant No. 2024CSJGG1100), and the AI & AI for Science Project of Nanjing University.

References

Bender, C. M.; and Orszag, S. A. 2013. *Advanced mathematical methods for scientists and engineers I: Asymptotic methods and perturbation theory*. Springer Science & Business Media.

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*.

Elfwing, S.; Uchibe, E.; and Doya, K. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107: 3–11.

Fang, H.; Carbajal, G.; Wermter, S.; and Gerkmann, T. 2021. Variational autoencoder for speech enhancement with a noise-aware encoder. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 676–680. IEEE.

Fu, S.-W.; Liao, C.-F.; Tsao, Y.; and Lin, S.-D. 2019. Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *International Conference on Machine Learning*, 2031–2041. PmLR.

He, G.; Zheng, K.; Chen, J.; Bao, F.; and Zhu, J. 2024. Consistency diffusion bridge models. *Advances in Neural Information Processing Systems*, 37: 23516–23548.

Holderrieth, P.; and Erives, E. 2025. An Introduction to Flow Matching and Diffusion Models. *arXiv preprint arXiv:2506.02070*.

Jukić, A.; Korostik, R.; Balam, J.; and Ginsburg, B. 2024. Schrödinger Bridge for Generative Speech Enhancement. In *Interspeech 2024*, 1175–1179.

Korostik, R.; Nasretdinov, R.; and Jukić, A. 2025. Modifying Flow Matching for Generative Speech Enhancement. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Lay, B.; Lermercier, J.-M.; Richter, J.; and Gerkmann, T. 2024. Single and few-step diffusion for generative speech enhancement. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 626–630. IEEE.

Lay, B.; Welker, S.; Richter, J.; and Gerkmann, T. 2023. Reducing the Prior Mismatch of Stochastic Differential Equations for Diffusion-based Speech Enhancement. In *Interspeech 2023*, 3809–3813.

Le Roux, J.; Wisdom, S.; Erdogan, H.; and Hershey, J. R. 2019. SDR–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 626–630. IEEE.

Lee, S.; Cheong, S.; Han, S.; and Shin, J. W. 2025. FlowSE: Flow Matching-based Speech Enhancement. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Lei, Y.; Chen, B.; Tai, W.; Zhong, T.; and Zhou, F. 2024. Shallow diffusion for fast speech enhancement (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 23556–23558.

Lemercier, J.-M.; Richter, J.; Welker, S.; and Gerkmann, T. 2023. StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 2724–2737.

Lemercier, J.-M.; Richter, J.; Welker, S.; Moliner, E.; Välimäki, V.; and Gerkmann, T. 2025. Diffusion Models for Audio Restoration: A review. *IEEE Signal Processing Magazine*, 41(6): 72–84.

Li, Y.; Sun, Y.; and Angelov, P. P. 2025. Complex-Cycle-Consistent Diffusion Model for Monaural Speech Enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 18557–18565.

Liu, H.; Wang, W.; and Plumbley, M. D. 2024. Latent diffusion model for audio: Generation, quality enhancement, and neural audio codec. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*.

Reddy, C. K.; Dubey, H.; Koishida, K.; Nair, A.; Gopal, V.; Cutler, R.; Braun, S.; Gamper, H.; Aichner, R.; and Srinivasan, S. 2021. INTERSPEECH 2021 Deep Noise Suppression Challenge. In *Interspeech 2021*, 2796–2800.

Richter, J.; De Oliveira, D.; and Gerkmann, T. 2025. Investigating training objectives for generative speech enhancement. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Richter, J.; and Gerkmann, T. 2024. Diffusion-based speech enhancement: Demonstration of performance and generalization. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*.

Richter, J.; Welker, S.; Lemercier, J.-M.; Lay, B.; and Gerkmann, T. 2023. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 2351–2364.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *Proc. ICLR*.

Tai, W.; Lei, Y.; Zhou, F.; Trajcevski, G.; and Zhong, T. 2023a. DOSE: Diffusion dropout with adaptive prior for speech enhancement. *Advances in Neural Information Processing Systems*, 36: 40272–40293.

Tai, W.; Zhou, F.; Trajcevski, G.; and Zhong, T. 2023b. Revisiting denoising diffusion probabilistic models for speech enhancement: Condition collapse, efficiency and refinement. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 13627–13635.

Tong, A.; Fatras, K.; Malkin, N.; Huguet, G.; Zhang, Y.; Rector-Brooks, J.; Wolf, G.; and Bengio, Y. 2023. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*.

Valentini-Botinhao, C.; Wang, X.; Takaki, S.; and Yamagishi, J. 2016. Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech. In *SSW*, 146–152.

Wang, Z.; Zhu, X.; Zhang, Z.; Lv, Y.; Jiang, N.; Zhao, G.; and Xie, L. 2024. SELM: Speech enhancement using discrete tokens and language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 11561–11565. IEEE.

Wang, Z.-Q.; Cornell, S.; Choi, S.; Lee, Y.; Kim, B.-Y.; and Watanabe, S. 2023. TF-GridNet: Making time-frequency domain models great again for monaural speaker separation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Welker, S.; Richter, J.; and Gerkmann, T. 2022. Speech Enhancement with Score-Based Generative Models in the Complex STFT Domain. In *Proc. Interspeech 2022*, 2928–2932.

Yin, D.; Luo, C.; Xiong, Z.; and Zeng, W. 2020. Phasen: A phase-and-harmonics-aware speech enhancement network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 9458–9465.

Zheng, C.; Peng, X.; Zhang, Y.; Srinivasan, S.; and Lu, Y. 2021. Interactive speech and noise modeling for speech enhancement. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 14549–14557.