SPECTRAL SNR-T BIAS CORRECTION FOR DIFFUSION MODELS IN NON-STATIONARY TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion probabilistic models have been increasingly applied to probabilistic time series forecasting. However, the inference-time signal-to-noise ratio–timestep (SNR-t) bias previously observed in visual diffusion models remains insufficiently explored in this setting. SNR-t bias denotes an inference-time mismatch between the empirical SNR of an intermediate denoising sample and the expected SNR implied by its nominal diffusion timestep. While training establishes a fixed correspondence between timesteps and noise statistics through the forward perturbation process, accumulated reverse-process errors can break this correspondence during inference. In this work, we systematically examine SNR-t bias in diffusion-based time series forecasting across multiple models and benchmark datasets. We find that the bias exhibits a distinctive spectral signature: the estimated noise residual develops frequency-dependent power distortions along the reverse trajectory, a phenomenon we term **spectral SNR-t bias**. This spectral distortion can introduce spurious trends, periodic artifacts, and miscalibrated prediction intervals. Building on this diagnosis, we propose **Spectral-Shape SNR Correction (SSSC)**, a training-free, plug-and-play inference procedure that adaptively reshapes intermediate samples toward the spectral profile expected at their nominal timesteps. SSSC uses spectral energy to modulate frequency-wise correction strengths calibrated on held-out data and abstains from correction when reliable calibration is not possible. Experiments across seven diffusion forecasters show that SSSC is particularly effective for over-dispersed forecasts, reducing CRPS by up to 55% relative to uncorrected baselines while also improving point forecast accuracy and probabilistic calibration.

1 INTRODUCTION

Diffusion probabilistic models (DPMs) have emerged as a prominent framework for probabilistic time series forecasting Rasul et al. (2021); Tashiro et al. (2021); Shen & Kwok (2023); Yuan & Qiao (2024); Ye et al. (2025). In the standard variance-preserving formulation Ho et al. (2020), the forward process perturbs a clean target $\mathbf{x}_0$ according to $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\bar{\alpha}_t$ is determined by the noise schedule Nichol & Dhariwal (2021). This construction establishes a training-time correspondence between the conditioning timestep and the perturbation statistics, with a schedule-defined nominal signal-to-noise ratio (SNR) $\mathrm{SNR}_{\mathrm{nom}}(t) = \bar{\alpha}_t/(1-\bar{\alpha}_t)$. During inference, however, this correspondence is no longer guaranteed to hold. Each intermediate sample $\hat{\mathbf{x}}_{t-1}$ is generated from the preceding state $\hat{\mathbf{x}}_t$, rather than sampled directly from the corresponding forward marginal. Accumulated denoising errors and sampling approximations can therefore cause the empirical SNR of these samples to deviate from the nominal value associated with their conditioning timestep. This mismatch, termed SNR-t bias, has been observed in image diffusion models Ning et al. (2023; 2024); Li et al. (2024a); Lin et al. (2024); Yu et al. (2026), but remains insufficiently explored in diffusion-based time series forecasting.

Time series often exhibit non-stationary and context-dependent temporal structure, with their statistical and spectral characteristics varying across time and forecasting contexts Kim et al. (2022); Liu et al. (2022); Fan et al. (2023); Ye et al. (2024); Liu et al. (2023); Dai et al. (2024). This raises the question of whether the SNR-t bias observed in image diffusion models also arises in diffusion-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

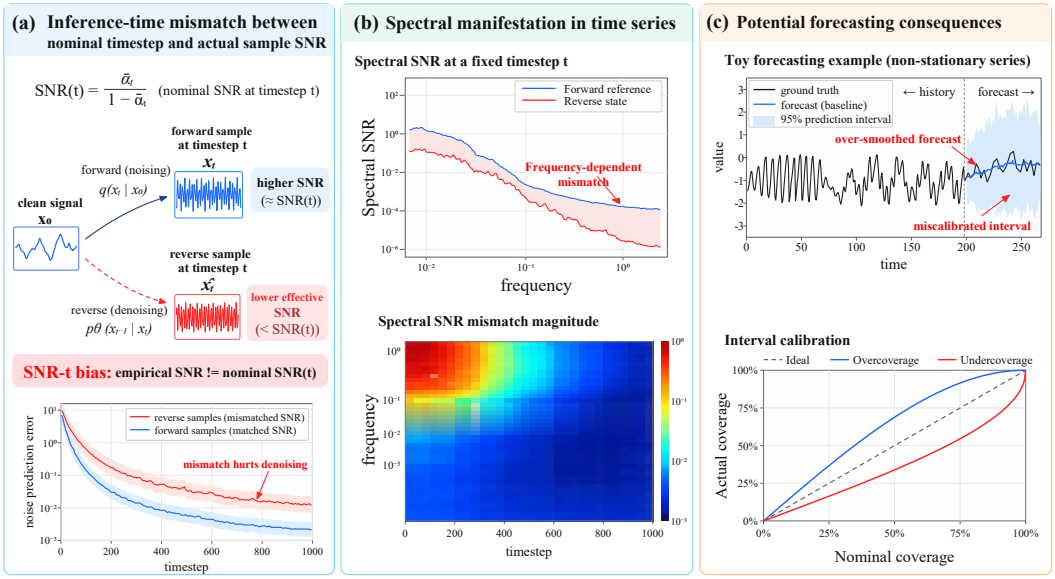


Figure 1: **Schematic illustration of spectral SNR-t bias in diffusion forecasting.** (a) Reverse samples may depart from the signal-to-noise statistics associated with their nominal timestep, potentially affecting denoising. (b) Spectral SNR profiles at a fixed timestep and a frequency–timestep heatmap illustrate the frequency-dependent manifestation of this mismatch. (c) A toy forecasting example and interval-calibration curves illustrate potential distortions in temporal structure and predictive uncertainty. In the calibration plot, the dashed diagonal denotes ideal coverage; curves above and below it indicate overcoverage and undercoverage, respectively.

based time series forecasting, and how it manifests across frequencies. Temporal signals distribute their energy unevenly across frequency components Zhou et al. (2022); Yi et al. (2023), making an aggregate SNR insufficient to characterize discrepancies within individual frequency bands. In particular, deviations from the reference signal-to-noise statistics associated with a nominal timestep need not be uniform across the spectrum. We refer to this frequency-dependent manifestation of SNR-t mismatch as **spectral SNR-t bias**. Such a mismatch may leave excess or deficient energy in particular frequency components of the final forecasts, distorting temporal structure and predictive uncertainty. Figure 1 summarizes SNR-t mismatch, its spectral manifestation, and its potential forecasting consequences.

Motivated by the potential spectral consequences of SNR-t bias, we propose **Spectral-Shape SNR Correction (SSSC)**, a training-free, plug-and-play inference-time correction method. Rather than intervening at individual reverse steps, SSSC applies frequency-wise correction to the final forecast samples. The method comprises three modules. *Spectral Profile Estimation (SPE)* estimates the power spectra of training targets and forecasts from the frozen model to identify frequency-dependent energy discrepancies. *Spectral Correction (SC)* constructs bounded gains through spectral matching and energy-gated shrinkage, with optional low-frequency protection. *Adaptive Selection (AS)* selects a configuration using the continuous ranked probability score (CRPS) on training data, retaining the original forecasts when the required improvement is not achieved or the spectral stand-down rule is triggered. The selected gain is fixed before testing and applied once after sampling, leaving the reverse sampling trajectory unchanged. SSSC requires neither model parameter updates nor additional diffusion sampling at test time.

Our main contributions are summarized as follows:

- We examine SNR-t bias in diffusion-based time series forecasting from a spectral perspective, focusing on its frequency-dependent manifestation and potential consequences for temporal structure and predictive uncertainty.
- We propose **SSSC**, a training-free, plug-and-play method that combines spectral profile estimation, bounded frequency-wise correction, and training-based configuration selection

to compensate for spectral discrepancies in final forecast samples without modifying the reverse sampling trajectory.

- Extensive experiments across multiple diffusion forecasters and benchmark datasets demonstrate the effectiveness of SSSC, with CRPS reductions of up to approximately 55% relative to uncorrected baselines.

2 Related Work

Diffusion-based Time Series Forecasting. Diffusion models have emerged as a flexible framework for probabilistic time series forecasting, with existing methods differing in temporal representation, conditioning mechanisms, and uncertainty modeling. TimeGrad Rasul et al. (2021) combines recurrent history encoding with autoregressive diffusion, while ScoreGrad Yan et al. (2021) adopts conditional score-based modeling in continuous time. CSDI Tashiro et al. (2021) conditions diffusion on observed entries for imputation and forecasting, and SSSD Lopez Alcaraz & Strodthoff (2022) incorporates structured state space layers to capture long-range temporal dependencies. TimeDiff Shen & Kwok (2023) introduces future mixup and autoregressive initialization for non-autoregressive prediction. TMDM Li et al. (2024b) incorporates Transformer-derived conditional information into both the forward and reverse diffusion processes. Diffusion-TS Yuan & Qiao (2024) combines temporal decomposition with a Fourier-based reconstruction objective, whereas MG-TSD Fan et al. (2024) uses multi-granularity temporal targets to guide diffusion training.

Recent studies further distinguish different sources and structures of predictive uncertainty. D3U Li et al. (2025) combines a pretrained point forecaster with a conditional diffusion model of its prediction residuals, using a patch-based denoising network to model the uncertain component. NsDiff Ye et al. (2025) incorporates conditional mean and variance estimates through a location–scale noise formulation and an uncertainty-aware noise schedule. Complementary to these conditional formulations, TSDiff Kollovieh et al. (2023) trains an unconditional diffusion model and uses self-guidance for forecasting at inference time; it also refines predictions from existing forecasters through iterative updates based on the learned density. Our work focuses on SNR-t mismatch and its potential spectral consequences in forecasts generated by pretrained diffusion models, rather than introducing a new forecasting architecture or uncertainty parameterization.

Inference-time Bias in Diffusion Models. A related line of research studies discrepancies between diffusion training and sampling. Exposure bias arises because denoisers are trained on samples from the forward process but operate on recursively generated states during inference, allowing prediction errors to accumulate across reverse steps Ning et al. (2023; 2024). Input Perturbation Ning et al. (2023) addresses this discrepancy by adding perturbations to denoiser inputs during training. At inference time, Epsilon Scaling Ning et al. (2024) rescales noise predictions, while the Time-Shift Sampler Li et al. (2024a) adjusts sampling timesteps to improve their alignment with generated states. Lin et al. Lin et al. (2024) identify additional training–inference inconsistencies caused by nonzero terminal SNR and mismatched sampling start points.

Frequency-domain analyses provide a complementary perspective on these discrepancies. DPM-FR Yu & Zhan (2025) regulates wavelet-subband energy during reverse sampling to mitigate exposure bias. DCW Yu et al. (2026) explicitly studies SNR-t bias and applies differential correction to intermediate samples across wavelet subbands. Motivated by the spectral consequences of SNR-t mismatch, SSSC instead applies bounded frequency-wise gains to completed forecast samples. Its correction reference is estimated from training targets and model forecasts, rather than from the nominal statistics of intermediate diffusion states. The correction configuration is selected on training data and fixed before testing, leaving both model parameters and the reverse sampling trajectory unchanged.

3 Method

Problem Formulation and Overview. Given historical observations $\mathbf{X} \in \mathbb{R}^{L \times C}$, probabilistic time series forecasting aims to estimate the conditional distribution $p(\mathbf{Y} \mid \mathbf{X})$ of a future sequence $\mathbf{Y} \in \mathbb{R}^{H \times C}$, where L, H, and C denote the context length, forecasting horizon, and number of variables, respectively. A pretrained diffusion forecaster generates M future trajectories $\{\widehat{\mathbf{Y}}^{(m)}\}_{m=1}^{M}$

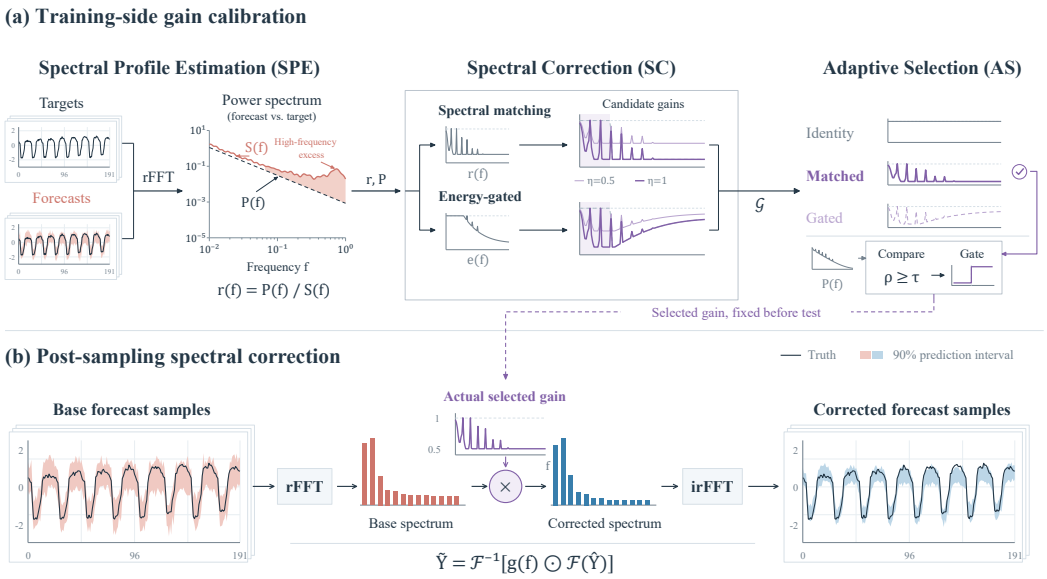


Figure 2: **Overview of Spectral-Shape SNR Correction (SSSC).** (a) Training-side spectral estimation, gain construction, and adaptive selection determine a frequency-wise gain, fixed before testing. (b) The selected gain is applied once to completed forecast samples in the Fourier domain, leaving reverse sampling unchanged.

through reverse sampling, whose empirical distribution represents its predictions and associated uncertainty. We consider correcting spectral distortions in these trajectories using statistics estimated from the training split, while keeping the forecasting model fixed.

To describe SNR–timestep mismatch and its potential spectral consequences, we use the standard variance-preserving formulation as a reference. The correction introduced below does not require all forecasting backbones to share this parameterization. Let $\mathbf{y}_0 = \mathbf{Y}$ denote the forecasting target, $\mathbf{c} = h(\mathbf{X})$ the historical condition, and t the diffusion timestep. Given a variance schedule $\{\beta_t\}_{t=1}^{T}$, the forward process progressively perturbs the target through Gaussian transitions:

$$q(\mathbf{y}_{1:T} \mid \mathbf{y}_0) = \prod_{t=1}^{T} q(\mathbf{y}_t \mid \mathbf{y}_{t-1}), \qquad q(\mathbf{y}_t \mid \mathbf{y}_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{y}_{t-1}, \beta_t \mathbf{I}), \tag{1}$$

where $\alpha_t = 1 - \beta_t$. Besides constructing training inputs, this process specifies the reference perturbation statistics at each timestep. Defining $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, its marginal distribution and direct sampling expression are

$$q(\mathbf{y}_t \mid \mathbf{y}_0) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t}\mathbf{y}_0, (1-\bar{\alpha}_t)\mathbf{I}\big), \qquad \mathbf{y}_t = \sqrt{\bar{\alpha}_t}\mathbf{y}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The timestep determines the relative strengths of signal and injected noise through $\bar{\alpha}_t$. Under the unit-average-signal-power convention, the nominal SNR is $\mathrm{SNR}_{\mathrm{nom}}(t) = \bar{\alpha}_t/(1-\bar{\alpha}_t)$. This scalar characterizes the overall perturbation level. However, the injected white noise has a flat expected power spectrum, whereas time-series signals generally distribute their energy unevenly across frequencies. Consequently, frequency-specific SNRs can differ even at the same timestep. This intrinsic variation is distinct from spectral SNR–timestep bias, which concerns departures from the corresponding forward reference at a given timestep and frequency.

The reverse process starts from $p(\mathbf{y}_T) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and generates future trajectories conditioned on the history:

$$p_\theta(\mathbf{y}_{0:T} \mid \mathbf{c}) = p(\mathbf{y}_T) \prod_{t=1}^{T} p_\theta(\mathbf{y}_{t-1} \mid \mathbf{y}_t, \mathbf{c}), \qquad p_\theta(\mathbf{y}_{t-1} \mid \mathbf{y}_t, \mathbf{c}) = \mathcal{N}\big(\boldsymbol{\mu}_\theta(\mathbf{y}_t, t, \mathbf{c}), \sigma_t^2 \mathbf{I}\big). \tag{3}$$

Each reverse state depends on preceding model predictions. A denoising error therefore affects both the current reconstruction and the inputs encountered in subsequent steps. Under the noise-

prediction parameterization, the reverse mean is

$$\boldsymbol{\mu}_\theta(\mathbf{y}_t, t, \mathbf{c}) = \frac{1}{\sqrt{\alpha_t}} \left[\mathbf{y}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{y}_t, t, \mathbf{c}) \right]. \tag{4}$$

The denoiser is trained by minimizing $\mathcal{L}_{\text{diff}} = \mathbb{E}\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{y}_t, t, \mathbf{c})\|_F^2$, with $\mathbf{y}_t$ constructed using Eq. equation 2. Training inputs thus follow prescribed perturbation statistics, whereas inference inputs are recursively generated. Accumulated errors can shift reverse states away from their corresponding forward marginals. Such deviations can manifest as SNR–timestep mismatch, in which the realized signal-to-noise statistics differ from those associated with the conditioning timestep. Frequency-dependent errors may also leave excess or deficient energy at particular frequencies in the final forecasts.

Motivated by this connection, we propose *Spectral-Shape SNR Correction* (SSSC), which uses the spectral discrepancy between training targets and generated forecasts to correct the final forecast samples. SSSC applies a single spectral transformation after sampling, targeting output spectral distortion without modifying intermediate reverse states. SSSC consists of three modules: Spectral Profile Estimation (SPE) identifies the frequency-dependent discrepancy; Spectral Correction (SC) translates it into bounded adjustments; and Adaptive Selection (AS) determines which configuration, including no correction, to apply. Together, these modules establish the correction reference, control its scope and strength, and assess its utility for probabilistic forecasting. Figure 2 summarizes the procedure.

3.1 Spectral Profile Estimation

SPE identifies frequency-dependent energy discrepancies by comparing the spectra of training targets and completed forecasts generated by the frozen model from training histories. To reduce the influence of individual sampling fluctuations, we aggregate over training windows, variables, and generated trajectories:

$$P(f) = \widehat{\mathbb{E}}_{i,c}\left[|\mathcal{F}(\mathbf{Y}_{i,c})(f)|^2\right], \qquad S(f) = \widehat{\mathbb{E}}_{i,c,m}\left[|\mathcal{F}(\widehat{\mathbf{Y}}_{i,c}^{(m)})(f)|^2\right], \tag{5}$$

where $\mathcal{F}$ denotes the discrete Fourier transform for real-valued inputs along the forecasting horizon. We retain the nonredundant frequency bins $f = 0, \ldots, F-1$, where $F = \lfloor H/2 \rfloor + 1$. The indices i, c, and m denote training windows, variables, and forecast samples, respectively, and $\widehat{\mathbb{E}}$ denotes empirical averaging. Both spectra use the same data scale, forecast length, and transform convention.

The power ratio $r(f) = P(f)/\max\{S(f), \varepsilon\}$ provides the correction reference, where $\varepsilon > 0$ is a numerical floor. When the floor is inactive, values below one indicate excess forecast energy, while values above one indicate deficient energy. This reference determines the direction of adjustment at each frequency; its scope and strength are controlled by SC to avoid unnecessarily modifying weak components or sensitive low-frequency structure.

3.2 Spectral Correction

Multiplying a Fourier amplitude by $g(f)$ scales its power by $g(f)^2$, making $\sqrt{r(f)}$ the natural gain for matching the estimated target spectrum. We introduce a strength parameter η and gain bounds $0 < g_{\min} \leq 1 \leq g_{\max}$ to control the adjustment:

$$g_\eta(f) = \text{clip}\Big(r(f)^{\eta/2}, g_{\min}, g_{\max}\Big), \qquad \mathcal{T}_g(\widehat{\mathbf{Y}}^{(m)}) = \mathcal{F}^{-1}\Big[g \odot \mathcal{F}(\widehat{\mathbf{Y}}^{(m)})\Big]. \tag{6}$$

Here, $\odot$ denotes multiplication along the frequency dimension, and $\mathcal{F}^{-1}$ is the corresponding inverse transform with output length H. The positive real-valued gain is shared across variables and applied to each forecast sample, preserving Fourier phases. For the basic matching gain, the calibration spectra satisfy $S_{\text{corr}}(f) = P(f)^\eta S(f)^{1-\eta}$ when both powers are positive and neither flooring nor clipping is active. For $0 \leq \eta \leq 1$, this interpolates between the calibration spectra on a logarithmic scale; it does not imply the same matching relation on test data.

Full matching can be sensitive to spectral estimation errors, particularly at frequencies with weak target energy. We therefore include an energy-gated shrinkage variant:

$$w(f) = \mathrm{clip}\left(\sqrt{\frac{P(f)}{\max\{\bar{P}, \varepsilon\}}}, 0, 1\right), \qquad g_\eta^{\mathrm{sh}}(f) = \min\Big\{1, \mathrm{clip}\Big(r(f)^{\eta w(f)/2}, g_{\min}, g_{\max}\Big)\Big\}, \tag{7}$$

where $\bar{P} = (F-1)^{-1}\sum_{f=1}^{F-1} P(f)$ is the mean non-DC target power. The gate is used only in the shrinkage family and reduces adjustments to weak target components; the upper bound of one prevents amplification. To optionally preserve the forecast level and slowly varying structure, we also construct low-frequency-protected versions of both gain families by setting $g(f) = 1$ for $f < k_{\mathrm{lo}}$.

Together, these variants provide different choices of correction direction, strength, and frequency range. We use $\eta \in \{0.5, 1.0\}$ for each of the four configurations and include the identity gain, yielding nine candidate configurations. AS determines which configuration to apply.

3.3 Adaptive Selection

Closer spectral agreement does not necessarily imply a better predictive distribution. We therefore select corrections using the continuous ranked probability score (CRPS) on training data. Let $\mathcal{G}$ be the candidate gain set, including the identity gain $\mathbf{1}$, and let $\widehat{\mathcal{L}}_{\mathrm{CRPS}}(g)$ denote the empirical CRPS after applying $\mathcal{T}_g$ to the training-side forecasts. The best-scoring candidate is

$$g^\star = \arg\min_{g \in \mathcal{G}} \widehat{\mathcal{L}}_{\mathrm{CRPS}}(g). \tag{8}$$

A non-identity correction is accepted only if it achieves a minimum relative CRPS improvement $\delta \in [0, 1)$ over the identity gain; otherwise, the original forecasts are retained.

We additionally apply a spectral stand-down rule when target power is strongly concentrated at DC and low frequencies. Let $\bar{P}_{\mathrm{lo}}$, $\bar{P}_{\mathrm{mid}}$, and $\bar{P}_{\mathrm{hi}}$ denote the mean target power within the corresponding non-DC bands. We define the concentration statistic $\rho = (\bar{P}_{\mathrm{mid}} + \bar{P}_{\mathrm{hi}})/(P(0) + \bar{P}_{\mathrm{lo}} + \bar{P}_{\mathrm{mid}} + \bar{P}_{\mathrm{hi}})$, with a numerical floor on the denominator. The final gain is

$$g_{\mathrm{final}} = \begin{cases} \mathbf{1}, & \rho < \tau \text{ or } \widehat{\mathcal{L}}_{\mathrm{CRPS}}(g^\star) > (1-\delta)\widehat{\mathcal{L}}_{\mathrm{CRPS}}(\mathbf{1}), \\ g^\star, & \text{otherwise.} \end{cases} \tag{9}$$

Ties in candidate selection favor the identity gain. The improvement criterion excludes corrections with insufficient training-score gains, while the stand-down rule retains the original forecasts for strongly concentrated target spectra. Neither rule guarantees an improvement on test data.

All spectral estimates and selection decisions use only the training split. The selected gain is then fixed and applied to test forecasts as $\widetilde{\mathbf{Y}}^{(m)} = \mathcal{T}_{g_{\mathrm{final}}}(\widehat{\mathbf{Y}}^{(m)})$. At test time, SSSC requires no parameter updates or additional diffusion sampling; its overhead consists of a batched Fourier transform, frequency-wise multiplication, and an inverse transform. Offline calibration requires forecasts from the frozen model on training-side data.

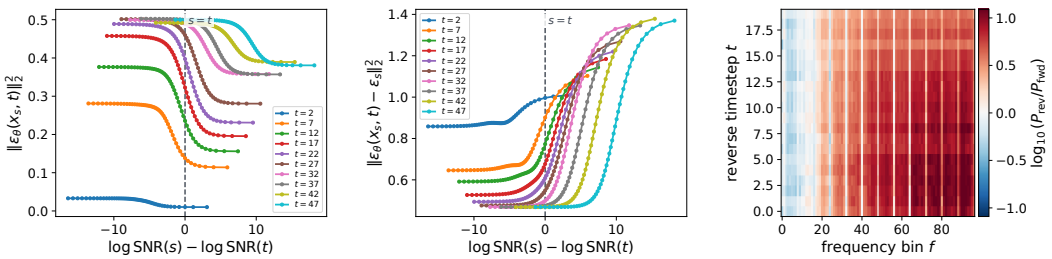
4 Experiments

In this section, we first examine the sensitivity of diffusion forecasters to SNR–timestep mismatch and characterize spectral deviations along reverse trajectories. We then evaluate the forecasting performance of SSSC across multiple backbones and benchmark datasets. Finally, we conduct ablation studies to assess the contributions of individual components and examine whether frequency-dependent correction provides benefits beyond scalar calibration. The analysis further characterizes the spectral and distributional effects of correction, with particular attention to changes in predictive intervals and calibration. Details of the experimental setup and evaluation protocol are provided in appendix.

4.1 Diagnosing SNR–Timestep Mismatch

We first examine the sensitivity of diffusion forecasters to SNR–timestep mismatch by fixing the conditioning timestep t while varying the actual perturbation timestep s. The matched condition $s = t$ corresponds to zero log-SNR difference. Figures 3a and 3b show that, as the log-SNR difference increases, the predicted-noise energy generally decreases, whereas the squared error against the injected noise increases. These responses reveal a systematic dependence on input perturbation statistics despite an unchanged timestep condition. The cross-model probes in Figure 6 in the appendix further characterize this sensitivity across backbones, with NsDiff included as a near-noiseless control.

We next examine spectral differences encountered during reverse sampling by comparing noise-prediction power on reverse states and matched forward states at the same timestep. Figure 3c shows predominantly positive log-power ratios at higher frequencies, whereas lower-frequency ratios are smaller or negative. The discrepancy also varies across reverse timesteps, revealing frequency-dependent deviations in denoiser outputs rather than a uniform spectral offset. This comparison characterizes noise-prediction spectra, rather than directly estimating the SNR of reverse states. Together, the controlled probe and trajectory analysis establish mismatch sensitivity and frequency-structured forward–reverse differences, motivating the investigation of frequency-dependent output correction.



(a) Controlled SNR–t response. (b) Error against injected noise. (c) Forward–reverse spectral gap.

Figure 3: **Controlled mismatch sensitivity and reverse-trajectory spectral diagnostics.** (a,b) Predicted-noise energy and squared error against injected noise as the actual perturbation timestep s varies under a fixed conditioning timestep t. Colors denote different fixed timesteps, and the dashed line marks the matched condition $s = t$. (c) Frequency-resolved log-power ratio $\log_{10}(P_{\text{rev}}/P_{\text{fwd}})$ between noise predictions on reverse states and matched forward states at the same timestep. Positive values indicate higher noise-prediction power on reverse states.

4.2 Main Results

Table 1 presents paired comparisons of six diffusion forecasters before and after applying SSSC. SSSC reduces CRPS in most reported settings and improves both CRPS and QICE for all six backbones on Electricity. In particular, CRPS decreases from 7.272 to 3.239 for CSDI and from 0.435 to 0.269 for TMDM, corresponding to relative reductions of 55.5% and 38.2%, respectively. Improvements in both metrics are also observed for CSDI, TimeDiff, TMDM, and NsDiff on Traffic, indicating that the benefits extend beyond a single backbone or dataset.

The gains nevertheless vary across settings, and forecasting and calibration improvements do not always coincide. For TMDM on ETTm2, CRPS decreases from 0.289 to 0.219, whereas QICE increases from 2.610 to 4.202. TimeDiff exhibits the opposite trade-off on this dataset: QICE decreases from 13.995 to 11.811, but CRPS increases from 0.201 to 0.558. These results demonstrate the benefits of SSSC while highlighting the need to assess forecasting performance and calibration jointly. Additional benchmark results and predictive-interval comparisons are provided in the appendix.

Table 1: Paired comparison of six diffusion forecasters before and after applying SSSC. Lower values indicate better performance, with better results highlighted in red. Additional results are provided in the appendix.

Dataset	Setting	TimeGrad		CSDI		TimeDiff		DiffusionTS		TMDM		NsDiff	
		CRPS ↓	QICE ↓	CRPS ↓	QICE ↓	CRPS ↓	QICE ↓	CRPS ↓	QICE ↓	CRPS ↓	QICE ↓	CRPS ↓	QICE ↓
ETTh1	Base	0.606	6.731	0.550	4.538	0.471	14.765	0.559	4.598	0.452	2.821	0.392	1.470
	+SSSC	**0.540**	**3.196**	**0.517**	**3.956**	**0.428**	**12.528**	**0.513**	**2.266**	**0.404**	**1.260**	**0.353**	2.833
ETTh2	Base	1.212	9.488	0.683	7.123	0.471	14.813	1.168	9.577	0.383	4.471	0.358	**2.074**
	+SSSC	**0.836**	**6.426**	**0.523**	**6.255**	**0.344**	**14.451**	**1.142**	**7.970**	**0.273**	**1.563**	**0.266**	5.462
ETTm1	Base	0.744	7.608	0.647	5.017	0.589	14.997	**0.627**	**5.481**	0.371	4.905	0.318	2.805
	+SSSC	0.744	7.608	**0.567**	**3.645**	**0.494**	**13.288**	0.684	8.047	0.371	4.905	0.318	2.805
ETTm2	Base	**0.743**	**6.530**	0.814	9.035	**0.201**	13.995	1.035	9.959	0.289	2.610	0.256	**2.030**
	+SSSC	0.803	9.144	**0.463**	**5.430**	0.558	**11.811**	**0.891**	11.507	**0.219**	4.202	**0.175**	2.645
Electricity	Base	0.599	5.829	7.272	17.987	0.687	15.475	0.504	8.872	0.435	9.558	0.249	6.548
	+SSSC	**0.484**	**5.673**	**3.239**	**16.341**	**0.659**	**14.706**	**0.495**	**7.098**	**0.269**	**6.457**	**0.198**	**2.459**
ILI	Base	1.104	6.519	1.244	7.693	1.152	14.942	1.612	10.090	0.967	**6.217**	**1.001**	**5.464**
	+SSSC	**0.960**	**4.994**	**0.958**	**5.938**	**1.101**	**14.969**	**1.311**	**8.809**	**0.814**	8.239	1.006	5.949
Solar	Base	0.335	6.985	**0.499**	11.285	0.449	13.996	0.536	9.315	0.416	6.575	0.327	5.217
	+SSSC	0.335	6.985	0.528	**7.930**	0.449	13.996	0.536	9.315	**0.348**	**5.537**	**0.314**	**4.670**
Traffic	Base	0.407	4.581	0.949	7.582	0.748	15.481	0.655	2.911	0.587	8.684	0.362	7.031
	+SSSC	0.407	4.581	**0.683**	**4.885**	**0.719**	**14.384**	0.655	2.911	**0.379**	**6.578**	**0.297**	**4.595**
Weather	Base	0.280	4.703	0.471	5.673	0.382	14.207	0.355	6.465	0.284	**6.355**	0.289	**3.479**
	+SSSC	0.280	4.703	**0.325**	**3.597**	**0.317**	**11.650**	0.355	6.465	**0.278**	6.882	**0.270**	4.332

4.3 Ablation Study

We assess the contributions of the three SSSC modules through ablation experiments on Electricity using CSDI, TimeDiff, TMDM, and NsDiff. The *w/o SPE* variant examines the role of Spectral Profile Estimation in constructing the correction reference from target and forecast spectra. The *w/o SC* variant evaluates the contribution of Spectral Correction, which translates this reference into bounded frequency-wise gains. The *w/o AS* variant examines Adaptive Selection, which uses training-side CRPS to select a correction configuration. These variants are compared with full SSSC and the uncorrected Base forecasts to distinguish component contributions from the overall effect of correction.

Table 2: Ablation study of SSSC on the Electricity dataset using CSDI, TimeDiff, TMDM, and NsDiff as backbones. Lower values are better, and the best results for each backbone are highlighted in bold. QICE is multiplied by 100.

Methods	CSDI		TimeDiff		TMDM		NsDiff	
Variant	CRPS ↓	QICE ↓	CRPS ↓	QICE ↓	CRPS ↓	QICE ↓	CRPS ↓	QICE ↓
Base	7.272	17.987	0.687	15.475	0.435	9.558	0.249	6.548
w/o SPE	3.718	17.910	0.673	14.874	0.376	2.973	0.347	**1.793**
w/o SC	3.718	17.910	1.019	15.159	0.356	**2.210**	0.265	2.106
w/o AS	3.675	17.869	0.662	14.788	0.282	6.173	0.217	2.212
SSSC	**3.239**	**16.341**	**0.659**	**14.706**	**0.269**	6.457	**0.198**	2.459

As shown in Table 2, full SSSC achieves the lowest CRPS for all four backbones. The effects of individual ablations vary across models: for NsDiff, w/o SPE increases CRPS from 0.198 to 0.347, whereas for TimeDiff, w/o SC increases it from 0.659 to 1.019. The w/o AS variant also yields higher CRPS across all four backbones, indicating a benefit from configuration selection in these comparisons. Calibration does not follow the same ranking: w/o SC achieves the lowest QICE for TMDM, while w/o SPE performs best on this metric for NsDiff. Thus, the complete configuration consistently favors CRPS within the reported ablation, but does not uniformly minimize calibration error.

Comparison with scalar calibration controls. To determine whether the improvement of SSSC can be explained by a simple rescaling of the predictive distribution, we compare it with global scaling and mean-preserving dispersion scaling on Electricity.

Figure 4 compares SSSC with two scalar calibration controls while holding the dataset fixed. For CSDI, SSSC reduces relative CRPS to 0.45, outperforming global scaling (0.51) and dispersion scaling (0.75). For TMDM, SSSC achieves 0.62, compared with 0.80 and 0.67 for the two scalar controls, respectively.

These results indicate that the improvement cannot be explained solely by uniformly shrinking the forecast or its ensemble dispersion. Instead, the frequency-dependent gain provides additional benefits by correcting different spectral components separately. The comparison is restricted to Electricity so that the observed differences reflect the correction mechanism rather than variations between datasets.

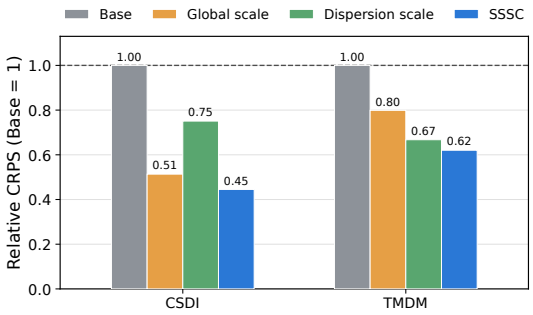
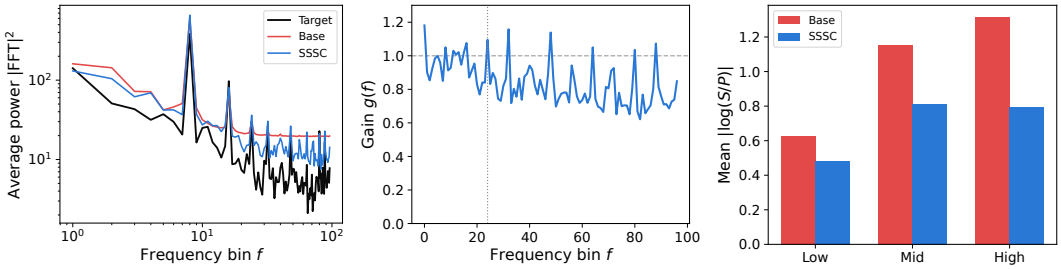


Figure 4: **Scalar-control comparison on Electricity.** CRPS is normalized by the corresponding Base score; lower values are better.

Spectral analysis. Figure 5 examines the effect of frequency-wise correction on NsDiff on ETTh2. The nonconstant gain adjusts individual spectral components, bringing the forecast spectrum closer to the target and reducing discrepancies across the low-, mid-, and high-frequency bands. This improvement in spectral agreement is accompanied by a decrease in CRPS from 0.358 to 0.266. These results support the effectiveness of output-level spectral correction in this setting, demonstrating improvements in both spectral consistency and probabilistic forecasting performance as measured by CRPS.



(a) Test-set power spectra. (b) Frequency-wise correction gain. (c) Band-wise spectral discrepancy.

Figure 5: **Output spectral correction for NsDiff on ETTh2.** SSSC applies a frequency-dependent gain that brings the predictive spectrum closer to the target and reduces spectral discrepancies across the low-, mid-, and high-frequency bands.

5 Conclusion

In this work, we proposed Spectral-Shape SNR Correction (SSSC), a training-free, plug-and-play method motivated by the potential spectral consequences of SNR-t bias in diffusion-based time series forecasting. SSSC integrates spectral profile estimation, spectral correction, and adaptive selection to compensate for frequency-dependent discrepancies in final forecast samples using spectral references estimated from training data. The selected correction is fixed before testing and applied after sampling, leaving model parameters and the reverse sampling trajectory unchanged and requiring no additional diffusion sampling at test time. Experiments across multiple diffusion forecasters and benchmark datasets demonstrate CRPS reductions of up to approximately 55% relative to uncorrected baselines, with calibration effects varying across settings. These findings suggest that the spectral structure of generated samples provides an additional basis for improving probabilistic forecasts, complementing advances in diffusion model design and sampling.

REFERENCES

Tao Dai, Beiliang Wu, Peiyuan Liu, Naiqi Li, Yuerong Xue, Shu-Tao Xia, and Zexuan Zhu. Ddn: Dual-domain dynamic normalization for non-stationary time series forecasting. In *Advances in Neural Information Processing Systems*, 2024.

Wei Fan, Pengyang Wang, Dongkun Wang, Dongjie Wang, Yuanchun Zhou, and Yanjie Fu. Dish-TS: A general paradigm for alleviating distribution shift in time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7522–7529, 2023. doi: 10.1609/aaai.v37i6.25914. URL https://ojs.aaai.org/index.php/AAAI/article/view/25914.

Xinyao Fan, Yueying Wu, Chang Xu, Yuhao Huang, Weiqing Liu, and Jiang Bian. MG-TSD: Multi-granularity time series diffusion models with guided learning process. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=CZiY6OLktd.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007. doi: 10.1111/j.1467-9868.2007.00587.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2007.00587.x.

Xizewen Han, Huangjie Zheng, and Mingyuan Zhou. CARD: Classification and regression diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022.

Marcel Kollovieh, Abdul Fatir Ansari, Michael Bohlke-Schneider, Jasper Zschiegner, Hao Wang, and Yuyang Wang. Predict, refine, synthesize: Self-guiding diffusion models for probabilistic time series forecasting. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 2018. doi: 10.1145/3209978.3210006. URL https://dl.acm.org/doi/10.1145/3209978.3210006.

Mingxiao Li, Tingyu Qu, Ruicong Yao, Wei Sun, and Marie-Francine Moens. Alleviating exposure bias in diffusion models through sampling with shifted time steps. In *The Twelfth International Conference on Learning Representations*, 2024a.

Qi Li, Zhenyu Zhang, Lei Yao, Zhaoxia Li, Tianyi Zhong, and Yong Zhang. Diffusion-based decoupled deterministic and uncertain framework for probabilistic multivariate time series forecasting. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yuxin Li, Wenchao Chen, Xinyue Hu, Bo Chen, Baolin Sun, and Mingyuan Zhou. Transformer-modulated diffusion models for probabilistic multivariate time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=qae04YACHs.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5404–5411, 2024.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Zhiding Liu, Mingyue Cheng, Zhi Li, Zhenya Huang, Qi Liu, Yanhu Xie, and Enhong Chen. Adaptive normalization for non-stationary time series forecasting: A temporal slice perspective. In *Advances in Neural Information Processing Systems*, 2023.

Juan Miguel Lopez Alcaraz and Nils Strodthoff. Diffusion-based time series imputation and forecasting with structured state space models, 2022. URL https://arxiv.org/abs/2208.09399.

James E. Matheson and Robert L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096, 1976. doi: 10.1287/mnsc.22.10.1087. URL https://doi.org/10.1287/mnsc.22.10.1087.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 2021.

Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Input perturbation reduces exposure bias in diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26245–26265. PMLR, 2023.

Mang Ning, Mingxiao Li, Jianlin Su, Albert Ali Salah, and Itir Onal Ertugrul. Elucidating the exposure bias in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning (ICML)*, 2021.

Lifeng Shen and James Kwok. Non-autoregressive conditional diffusion models for time series prediction. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31016–31029. PMLR, 2023. URL https://proceedings.mlr.press/v202/shen23d.html.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Tijin Yan, Hongwei Zhang, Tong Zhou, Yufeng Zhan, and Yuanqing Xia. Scoregrad: Multivariate probabilistic time series forecasting with continuous energy-based generative models. *arXiv preprint arXiv:2106.10121*, 2021.

Weiwei Ye, Songgaojun Deng, Qiaosha Zou, and Ning Gui. Frequency adaptive normalization for non-stationary time series forecasting. In *Advances in Neural Information Processing Systems*, volume 37, pp. 31350–31379, 2024.

Weiwei Ye, Zhuopeng Xu, and Ning Gui. Non-stationary diffusion for probabilistic time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, Pengyang Wang, Hui He, Ning An, Defu Lian, Longbing Cao, and Zhendong Niu. Frequency-domain MLPs are more effective learners in time series forecasting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 76656–76679, 2023.

Meng Yu and Kun Zhan. Frequency regulation for exposure bias mitigation in diffusion models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.

Meng Yu, Lei Sun, Jianhao Zeng, Xiangxiang Chu, and Kun Zhan. Elucidating the SNR-t bias of diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Xinyu Yuan and Yan Qiao. Diffusion-TS: Interpretable diffusion for general time series generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2403.01742.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325. URL https://ojs.aaai.org/index.php/AAAI/article/view/17325.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27268–27286. PMLR, 2022.

A Theoretical Analysis

We analyze how reverse-process errors can induce SNR–timestep mismatch, how this mismatch acquires a spectral structure, and why SSSC can compensate for the resulting output distortion. We use the variance-preserving formulation in Section 3, assume finite second moments, and suppress conditioning and channel indices.

1. Formation of SNR–timestep mismatch. Let $\mathbf{y}_t^{\mathrm{ref}}$ be a reference trajectory from the forward process, with $\mathbf{y}_0^{\mathrm{ref}} = \mathbf{y}_0$, and let $\widehat{\mathbf{y}}_t$ be the generated reverse state. For this analysis, couple their reverse transitions using the same Gaussian draws and the exact forward-posterior variance $\sigma_t^2 = \beta_t(1-\bar{\alpha}_{t-1})/(1-\bar{\alpha}_t)$. Define the state discrepancy $\mathbf{e}_t = \widehat{\mathbf{y}}_t - \mathbf{y}_t^{\mathrm{ref}}$ and the noise-prediction discrepancy $\boldsymbol{\delta}_t = \boldsymbol{\epsilon}_\theta(\widehat{\mathbf{y}}_t, t, \mathbf{c}) - \boldsymbol{\epsilon}_t^{\mathrm{ref}}$, where $\boldsymbol{\epsilon}_t^{\mathrm{ref}}$ is the aggregate forward noise of the reference state. Subtracting the two reverse updates gives

$$\mathbf{e}_{t-1} = \frac{\mathbf{e}_t}{\sqrt{\alpha_t}} - \kappa_t \boldsymbol{\delta}_t, \qquad \kappa_t = \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}. \tag{10}$$

Thus, a prediction discrepancy changes both the current state and the input to subsequent denoising steps.

Let $Y_0(f)$, $N_t(f)$, and $E_t(f)$ be the unitary Fourier coefficients of $\mathbf{y}_0$, $\boldsymbol{\epsilon}_t^{\mathrm{ref}}$, and $\mathbf{e}_t$. Write $\mathcal{P}(f) = \mathbb{E}|Y_0(f)|^2 > 0$ and $Q_t(f) = \mathbb{E}|E_t(f)|^2$. An explicit sufficient condition for noise contamination is $\mathbb{E}[E_t(f)Y_0(f)^*] = \mathbb{E}[E_t(f)N_t(f)^*] = 0$. Under this orthogonality assumption,

$$\mathrm{SNR}_{\mathrm{ref}}(t,f) = \frac{\bar{\alpha}_t \mathcal{P}(f)}{1-\bar{\alpha}_t}, \qquad \frac{\mathrm{SNR}_{\mathrm{rev}}(t,f)}{\mathrm{SNR}_{\mathrm{ref}}(t,f)} = \left(1 + \frac{Q_t(f)}{1-\bar{\alpha}_t}\right)^{-1}. \tag{11}$$

Indeed, the reverse residual has power $1 - \bar{\alpha}_t + Q_t(f)$. Consequently, $Q_t(f) > 0$ produces SNR–timestep mismatch in this regime, and a nonconstant $Q_t(f)$ produces frequency-dependent relative departures. The intrinsic variation of $\mathcal{P}(f)$ alone is not SNR–timestep bias.

2. From trajectory errors to output spectral bias. Unrolling Eq. equation 10, with $\bar{\alpha}_0 = 1$, yields

$$\mathbf{e}_0 = \frac{\mathbf{e}_T}{\sqrt{\bar{\alpha}_T}} - \sum_{t=1}^{T} \frac{\kappa_t}{\sqrt{\bar{\alpha}_{t-1}}} \boldsymbol{\delta}_t. \tag{12}$$

This connects the final discrepancy to initialization and denoising discrepancies throughout sampling, without assuming independence across steps. Let $\mathcal{S}(f) = \mathbb{E}|\mathcal{F}(\widehat{\mathbf{y}}_0)(f)|^2$. Since $\widehat{\mathbf{y}}_0 = \mathbf{y}_0 + \mathbf{e}_0$, the following identity holds without output orthogonality:

$$\mathcal{S}(f) - \mathcal{P}(f) = Q_0(f) + 2\,\mathrm{Re}\,\mathbb{E}[Y_0(f)E_0(f)^*]\,. \tag{13}$$

The error power and its correlation with the signal can therefore produce frequency-dependent excess or deficient energy. For orthogonal output errors, $\mathcal{S}(f)/\mathcal{P}(f) = 1 + Q_0(f)/\mathcal{P}(f)$: even spectrally flat additive errors produce unequal relative distortions when target power varies across frequencies. A single scalar gain cannot remove this distortion unless $\mathcal{S}(f)/\mathcal{P}(f)$ is constant.

3. Why frequency-wise correction is effective. SSSC estimates the output discrepancy using the calibration spectra $P(f)$ and $S(f)$ in Eq. equation 5. For positive powers and inactive flooring and clipping, its basic gain is $g_\eta(f) = [P(f)/S(f)]^{\eta/2}$, where $0 \le \eta \le 1$. Because Fourier multiplication scales power by $g_\eta(f)^2$,

$$\begin{aligned} S_{\mathrm{corr}}(f) &= g_\eta(f)^2 S(f) = P(f)^\eta S(f)^{1-\eta}, \\ \left|\log \frac{S_{\mathrm{corr}}(f)}{P(f)}\right| &= (1-\eta)\left|\log \frac{S(f)}{P(f)}\right|. \end{aligned} \tag{14}$$

Thus, the gain contracts the calibration log-power discrepancy at each frequency rather than imposing uniform rescaling. Clipping with $0 < g_{\min} \le 1 \le g_{\max}$ preserves non-increase because the clipped gain remains between the identity and full matching. Energy-gated shrinkage instead uses strength $\eta w(f)$ on excess-power bins, while leaving deficient-power bins unchanged; low-frequency protection likewise retains the original protected components.

These results explain a sufficient mechanism for SNR–t mismatch and its output-level compensation. They do not imply restoration of intermediate-state SNR or guaranteed improvements in test CRPS, motivating the adaptive selection in Section 3.3.

B More Experiment Details

B.1 Benchmark Datasets

We evaluate SSSC on 14 publicly available datasets for multivariate probabilistic time series forecasting: ETTh1, ETTh2, ETTm1, ETTm2, Weather, Solar-Energy, Electricity, Traffic, ExchangeRate, ILI, and PEMS03, PEMS04, PEMS07, and PEMS08. Dataset statistics and forecasting configurations are summarized in Table 3. All datasets are partitioned chronologically into training, validation, and test sets. The four ETT datasets use fixed split boundaries corresponding to a $6 : 2 : 2$ ratio, whereas the remaining datasets use a $7 : 1 : 2$ ratio. Public data sources and benchmark releases are listed below.

(i) **ETTh1, ETTh2, ETTm1, and ETTm2:** Zhou et al. (2021) `https://github.com/zhouhaoyi/ETDataset`.

(ii) **Weather:** `https://www.bgc-jena.mpg.de/wetter/`.

(iii) **Solar-Energy:** Lai et al. (2018) `https://github.com/laiguokun/multivariate-time-series-data/tree/master/solar-energy`.

(iv) **Electricity:** `https://archive.ics.uci.edu/dataset/321/electricityloaddiagrams20112014`.

(v) **Traffic:** `http://pems.dot.ca.gov/`.

(vi) **ExchangeRate:** Lai et al. (2018) `https://github.com/laiguokun/multivariate-time-series-data/tree/master/exchange_rate`.

(vii) **ILI:** `https://github.com/thuml/Autoformer`.

(viii) **PEMS03, PEMS04, PEMS07, and PEMS08:** `https://github.com/guoshnBJTU/ASTGNN`.

Table 3: Statistics and forecasting configurations of the benchmark datasets. Dimension denotes the number of variables. Dataset size reports the numbers of time points in the training, validation, and test splits, respectively, excluding historical context reused at split boundaries. Context and prediction lengths are measured in time steps.

Dataset	Dimension	Frequency	Dataset size	Context length	Prediction length
ETTh1, ETTh2	7	1 hour	(8640, 2880, 2880)	168	192
ETTm1, ETTm2	7	15 min	(34560, 11520, 11520)	168	192
Weather	21	10 min	(36887, 5270, 10539)	168	192
Solar-Energy	137	10 min	(36792, 5256, 10512)	168	192
Electricity	321	1 hour	(18412, 2632, 5260)	168	192
Traffic	862	1 hour	(12280, 1756, 3508)	168	192
ExchangeRate	8	1 day	(5311, 760, 1517)	168	192
ILI	7	1 week	(676, 97, 193)	168	36
PEMS03	358	5 min	(18345, 2622, 5241)	$48^{\dagger}$	$24^{\dagger}$
PEMS04	307	5 min	(11894, 1700, 3398)	$48^{\dagger}$	$24^{\dagger}$
PEMS07	883	5 min	(19756, 2824, 5644)	$48^{\dagger}$	$24^{\dagger}$
PEMS08	170	5 min	(12499, 1786, 3571)	$48^{\dagger}$	$24^{\dagger}$

Only the traffic-flow channel is used for the PEMS datasets. †The TimeGrad implementation used in our experiments adopts a 168-step context and a 192-step prediction horizon on PEMS to accommodate its autoregressive lag configuration. Its PEMS results therefore follow a different forecasting protocol from those of the other backbones.

B.2 Baselines in Main Experiments

We consider six representative diffusion-based time series models as forecasting backbones: TimeGrad Rasul et al. (2021), CSDI Tashiro et al. (2021), TimeDiff Shen & Kwok (2023), Diffusion-TS Yuan & Qiao (2024), TMDM Li et al. (2024b), and NsDiff Ye et al. (2025). For each backbone, we compare the original forecasts with their SSSC-corrected counterparts, keeping the model parameters and reverse sampling procedure unchanged. Brief descriptions and reference code repositories are provided below.

(i) **TimeGrad**: an autoregressive model for multivariate probabilistic forecasting that combines recurrent representations of historical observations with denoising diffusion to generate future values sequentially.
Code: `https://github.com/microsoft/ProbTS`

(ii) **CSDI**: a conditional score-based diffusion model for probabilistic time series imputation and forecasting that conditions the denoising process on observed entries to model the distribution of unobserved values.
Code: `https://github.com/microsoft/ProbTS`

(iii) **TimeDiff**: a non-autoregressive diffusion model for time series prediction that introduces future mixup and autoregressive initialization to incorporate forecasting-specific information into the conditioning mechanism.
Code: *[Specify the implementation source used in our experiments.]*

(iv) **Diffusion-TS**: an interpretable diffusion framework for multivariate time series generation that combines temporal decomposition with a Fourier-based reconstruction objective and supports conditional generation for forecasting and imputation.
Code: `https://github.com/Y-debug-sys/Diffusion-TS`

(v) **TMDM**: a Transformer-modulated diffusion model for probabilistic multivariate forecasting that incorporates Transformer-derived conditioning information into both the forward and reverse diffusion processes.
Code: `https://github.com/LiYuxin321/TMDM`

(vi) **NsDiff**: a diffusion-based probabilistic forecasting framework that adopts a location–scale noise formulation, combining conditional mean and variance estimation with an uncertainty-aware noise schedule to model time-varying uncertainty.
Code: `https://github.com/wwy155/NsDiff`

B.3 Evaluation Metrics

We evaluate probabilistic forecasting performance using the Continuous Ranked Probability Score (CRPS) and Quantile Interval Coverage Error (QICE). Lower values indicate better performance for both metrics.

Continuous Ranked Probability Score (CRPS). CRPS Matheson & Winkler (1976) measures the discrepancy between the cumulative distribution function (CDF) of the predictive distribution and the step function associated with the observation. For a predictive CDF F and an observed value y, it is defined as

$$\mathrm{CRPS}(F, y) = \int_{\mathbb{R}} \left(F(z) - \mathbb{I}\{y \le z\}\right)^2 \mathrm{d}z, \tag{15}$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function. For diffusion-based forecasters, the predictive CDF can be approximated by the empirical distribution of M generated samples $\{\widehat{y}^{(m)}\}_{m=1}^{M}$:

$$\widehat{F}(z) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}\{\widehat{y}^{(m)} \le z\}. \tag{16}$$

Substituting $\widehat{F}$ into Eq. equation 15 yields a sample-based evaluation of the predictive distribution.

Quantile Interval Coverage Error (QICE). QICE Han et al. (2022) assesses predictive calibration by measuring the discrepancy between empirical coverage and nominal probability across quantile intervals. For each evaluated target y_n, the predictive distribution is partitioned into K intervals with equal nominal probability $1/K$. Let $\widehat{q}_{n,k}$ denote the estimated quantile at level k/K for $k = 1, \ldots, K-1$, with outer boundaries $\widehat{q}_{n,0} = -\infty$ and $\widehat{q}_{n,K} = +\infty$. The empirical coverage of the k-th interval is

$$r_k = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}\left\{\widehat{q}_{n,k-1} < y_n \leq \widehat{q}_{n,k}\right\}, \tag{17}$$

where N is the number of evaluated scalar target observations. QICE is then defined as

$$\mathrm{QICE} = \frac{1}{K} \sum_{k=1}^{K} \left| r_k - \frac{1}{K} \right|. \tag{18}$$

Smaller QICE values indicate closer agreement between empirical and nominal interval coverage. For reporting, QICE is multiplied by 100.

B.4 Overall Performance

Table 4 reports the full paired comparisons between the original forecasts (*Base*) and their SSSC-corrected counterparts (*+SSSC*) across six diffusion forecasters and 13 datasets. The comparisons evaluate the effect of correction within each backbone–dataset setting using CRPS and QICE.

Table 4: Paired comparison of six diffusion forecasters before and after applying SSSC. Lower values indicate better performance, with better results highlighted in red.

Dataset	Setting	TimeGrad		CSDI		TimeDiff		DiffusionTS		TMDM		NsDiff	
		CRPS ↓	QICE ↓	CRPS ↓	QICE ↓	CRPS ↓	QICE ↓	CRPS ↓	QICE ↓	CRPS ↓	QICE ↓	CRPS ↓	QICE ↓
ETTh1	Base	0.606	6.731	0.550	4.538	0.471	14.765	0.559	4.598	0.452	2.821	0.392	1.470
	+SSSC	**0.540**	**3.196**	**0.517**	**3.956**	**0.428**	**12.528**	**0.513**	**2.266**	**0.404**	**1.260**	**0.353**	2.833
ETTh2	Base	1.212	9.488	0.683	7.123	0.471	14.813	1.168	9.577	0.383	4.471	0.358	**2.074**
	+SSSC	**0.836**	**6.426**	**0.523**	**6.255**	**0.344**	**14.451**	**1.142**	**7.970**	**0.273**	**1.563**	**0.266**	5.462
ETTm1	Base	0.744	7.608	0.647	5.017	0.589	14.997	**0.627**	**5.481**	0.371	4.905	0.318	2.805
	+SSSC	0.744	7.608	**0.567**	**3.645**	**0.494**	**13.288**	0.684	8.047	0.371	4.905	0.318	2.805
ETTm2	Base	**0.743**	**6.530**	0.814	9.035	**0.201**	13.995	1.035	9.959	0.289	2.610	0.256	**2.030**
	+SSSC	0.803	9.144	**0.463**	**5.430**	0.558	**11.811**	**0.891**	11.507	**0.219**	4.202	**0.175**	2.645
Electricity	Base	0.599	5.829	7.272	17.987	0.687	15.475	0.504	8.872	0.435	9.558	0.249	6.548
	+SSSC	**0.484**	**5.673**	**3.239**	**16.341**	**0.659**	**14.706**	**0.495**	**7.098**	**0.269**	**6.457**	**0.198**	**2.459**
ILI	Base	1.104	6.519	1.244	7.693	1.152	14.942	1.612	10.090	0.967	**6.217**	**1.001**	**5.464**
	+SSSC	**0.960**	**4.994**	**0.958**	**5.938**	**1.101**	**14.969**	**1.311**	**8.809**	**0.814**	8.239	1.006	5.949
Solar	Base	0.335	6.985	**0.499**	11.285	0.449	13.996	0.536	9.315	0.416	6.575	0.327	5.217
	+SSSC	0.335	6.985	0.528	**7.930**	0.449	13.996	0.536	9.315	**0.348**	**5.537**	**0.314**	**4.670**
Traffic	Base	0.407	4.581	0.949	7.582	0.748	15.481	0.655	2.911	0.587	8.684	0.362	7.031
	+SSSC	0.407	4.581	**0.683**	**4.885**	**0.719**	**14.384**	0.655	2.911	**0.379**	**6.578**	**0.297**	**4.595**
Weather	Base	0.280	4.703	0.471	5.673	0.382	14.207	0.355	6.465	0.284	**6.355**	0.289	**3.479**
	+SSSC	0.280	4.703	**0.325**	**3.597**	**0.317**	**11.650**	0.355	6.465	**0.278**	6.882	**0.270**	4.332
PEMS03	Base	0.635	5.083	0.629	**4.358**	0.369	14.646	0.432	11.359	0.466	10.136	0.174	**1.684**
	+SSSC	0.635	5.083	**0.528**	6.146	**0.351**	**12.397**	0.432	11.359	**0.281**	**7.702**	**0.172**	2.986
PEMS04	Base	**0.297**	**1.857**	0.513	6.411	0.417	14.585	0.360	9.421	0.444	10.164	0.162	**1.295**
	+SSSC	0.312	4.321	0.513	6.411	**0.405**	**13.640**	**0.300**	**8.374**	**0.268**	**7.148**	**0.161**	2.103
PEMS07	Base	0.238	2.289	0.644	3.796	**0.417**	14.865	0.471	13.381	0.543	9.259	0.165	1.254
	+SSSC	0.238	2.289	**0.569**	**3.344**	0.510	**14.327**	**0.445**	**13.119**	**0.335**	**7.362**	0.165	1.254
PEMS08	Base	0.399	**7.220**	0.496	5.586	0.445	14.892	0.360	9.282	0.376	9.154	**0.170**	**1.037**
	+SSSC	**0.386**	7.860	0.496	5.586	**0.425**	**13.206**	0.360	9.282	**0.253**	**5.630**	0.173	3.738

At the reported numerical precision, SSSC reduces CRPS in 53 of the 78 backbone–dataset pairs, leaves 17 unchanged, and increases it in 8. For QICE, 43 pairs improve, 17 remain unchanged, and 18 deteriorate. Both metrics improve simultaneously in 40 pairs. These results show that SSSC improves forecasting performance in a majority of the evaluated settings, although the gains are not uniform across backbones and datasets.

The additional PEMS results further illustrate this variation. For TMDM, SSSC improves both metrics on all four PEMS datasets, with relative CRPS reductions of approximately 39.7%, 39.6%, 38.3%, and 32.7% on PEMS03, PEMS04, PEMS07, and PEMS08, respectively. However, improvements in CRPS do not always coincide with lower calibration error. For example, NsDiff on PEMS03 achieves a modest CRPS reduction from 0.174 to 0.172, while QICE increases from 1.684 to 2.986. The full results therefore support the utility of output-level correction while highlighting the importance of evaluating forecasting performance and calibration jointly.

B.5 Cross-Model Responses to SNR–Timestep Mismatch

Figure 6 examines the responses of six diffusion forecasters to controlled SNR–timestep mismatch. For each curve, the conditioning timestep t supplied to the denoising network is fixed, while the actual perturbation timestep s of the input is varied. The case $s = t$ provides the matched reference. For the noise-prediction implementations of CSDI, TMDM, TimeGrad, and NsDiff, we report $\mathrm{RMS}(\hat{\varepsilon}_\theta)$. For the x_0-prediction implementations of Diffusion-TS and TimeDiff, we express the predictions in the nominal VP noise coordinates associated with t:

$$\hat{\varepsilon}_\theta^{\mathrm{eq}}(x_s, t) = \frac{x_s - \sqrt{\bar{\alpha}_t}\,\hat{x}_{0,\theta}(x_s, t)}{\sqrt{1-\bar{\alpha}_t}}, \tag{19}$$

and report $\mathrm{RMS}(\hat{\varepsilon}_\theta^{\mathrm{eq}})$. When $s \neq t$, this quantity is a residual relative to the nominal condition t, rather than the noise actually injected at s.

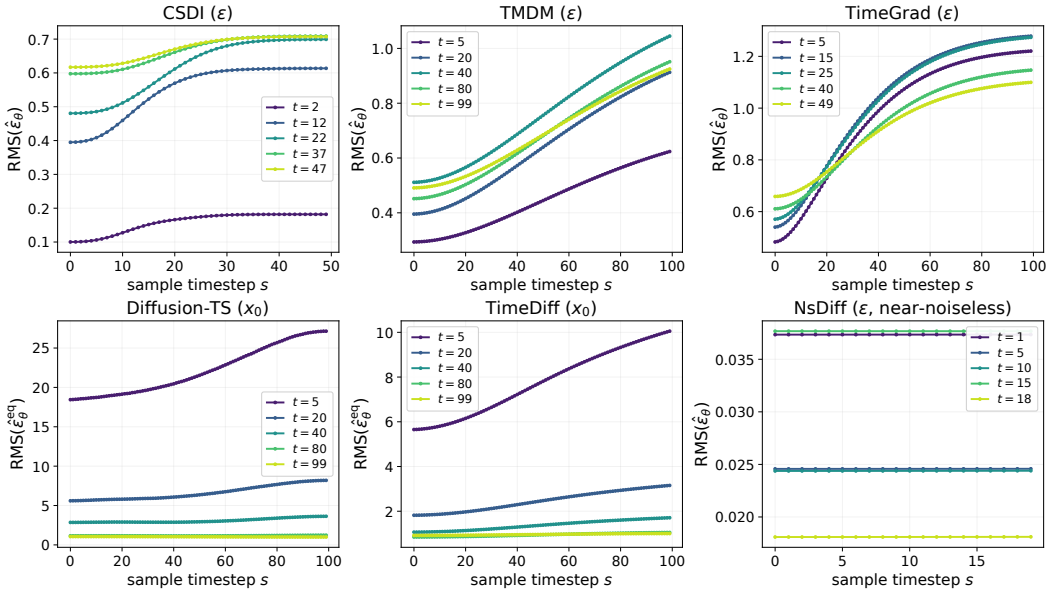


Figure 6: **Cross-model responses to controlled SNR–timestep mismatch.** Each curve fixes the conditioning timestep t while varying the input perturbation timestep s. The vertical axis reports predicted-noise RMS for CSDI, TMDM, TimeGrad, and NsDiff, and equivalent nominal-noise RMS for Diffusion-TS and TimeDiff, as defined in Eq. equation 19. The curves characterize within-model response patterns rather than forecasting accuracy.

Under a fixed conditioning timestep, CSDI, TMDM, and TimeGrad exhibit systematic changes in predicted-noise magnitude as s varies. Diffusion-TS and TimeDiff also exhibit changes in their equivalent residual magnitudes, although these responses additionally reflect the conversion in Eq. equation 19. In particular, the converted residual can vary with s even for an exact clean-signal prediction. NsDiff exhibits a comparatively weak response under the evaluated configuration. Overall, the probe characterizes model-dependent responses to deliberately mismatched input and timestep conditions. It complements the trajectory diagnostics in Section 4.1, but does not directly measure reverse-state SNR or establish forecasting degradation.

B.6 Qualitative Analysis of Predictive Intervals

Figure 7 compares the Base and SSSC-corrected 90% predictive intervals of six diffusion forecasters on the same Electricity example. The correction produces different changes across backbones and forecast positions. For NsDiff and TMDM, the corrected intervals become narrower while retaining the prominent periodic structure. In contrast, the Diffusion-TS interval expands, particularly around periodic troughs. TimeGrad exhibits more moderate and asymmetric adjustments across the forecasting horizon.

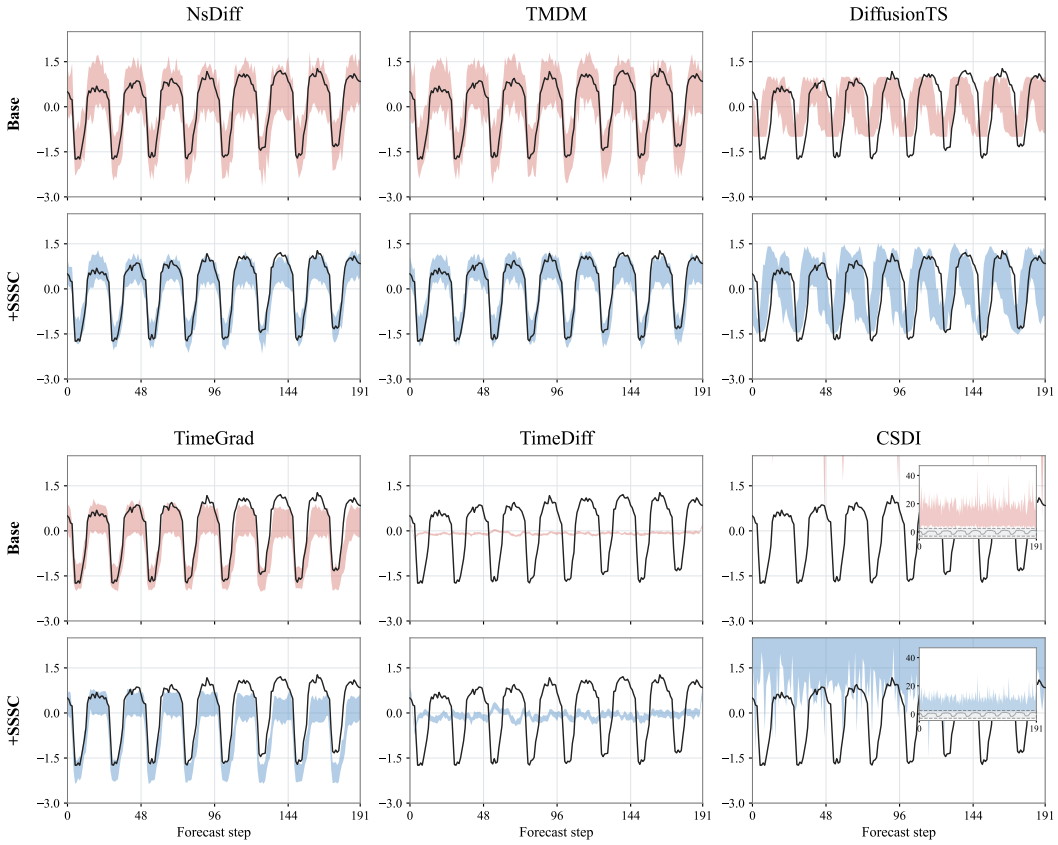
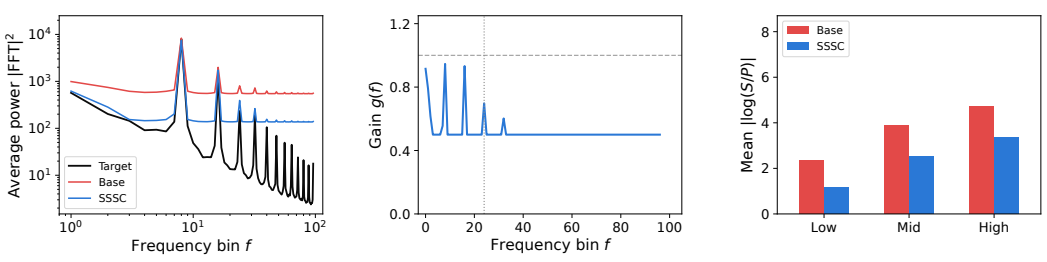


Figure 7: **Predictive intervals before and after SSSC.** Base and SSSC-corrected 90% predictive intervals are shown for six diffusion forecasters on an illustrative Electricity example (channel 187, window 0). All main panels share the same axis limits, while the CSDI insets display the full interval extent.

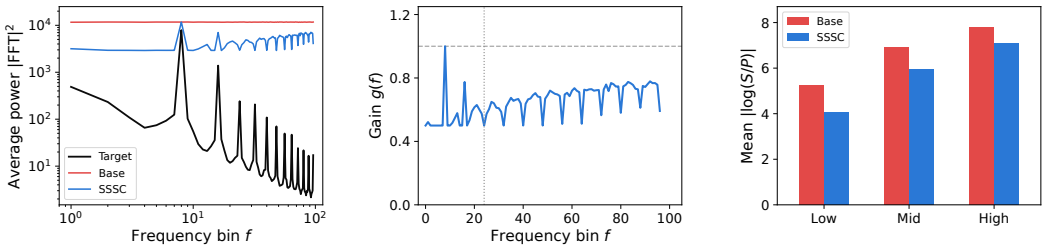
TimeDiff and CSDI illustrate contrasting effects of the correction. The Base TimeDiff interval is narrow and concentrated near zero. After correction, its interval envelope exhibits greater temporal variation, but still does not fully capture the amplitude of the observed trajectory. Conversely, the Base CSDI intervals extend over a much wider range than the observations. SSSC substantially reduces their width, as shown in the full-extent insets, although the corrected intervals remain broad relative to the observed scale. These examples illustrate backbone-dependent changes in interval width and temporal structure, rather than a consistent contraction across models. Because the example is selected for visual clarity, quantitative conclusions should be based on the aggregate CRPS and QICE results in Table 4; changes in interval width alone do not establish improved calibration Gneiting et al. (2007).

B.7 Spectral Analysis

Figure 8 extends the main-text spectral analysis to TMDM and CSDI on Electricity. For TMDM, the Base spectrum exhibits prominent periodic peaks but also excessive power between these peaks,



(a) Test-set power spectra. (b) Frequency-wise gain. (c) Band-wise discrepancy.



(d) Test-set power spectra. (e) Frequency-wise gain. (f) Band-wise discrepancy.

Figure 8: **Output spectral correction on Electricity.** The top and bottom rows show TMDM and CSDI, respectively. Columns present test-set power spectra, frequency-wise correction gains, and band-wise absolute log-spectral discrepancies before and after SSSC.

particularly at higher frequencies. The correction gain remains near 0.5 over much of the spectrum and approaches unity at the dominant peaks, suppressing the elevated spectral background while retaining the main periodic components. In contrast, CSDI exhibits an almost flat predictive spectrum with substantially inflated power across most frequencies. SSSC applies frequency-dependent attenuation to reduce this excess, although the corrected spectrum remains well above the target over substantial portions of the frequency range. The band-wise comparisons confirm that SSSC reduces the mean absolute log-spectral discrepancy in the low-, mid-, and high-frequency bands for both backbones. These improvements in spectral agreement are accompanied by CRPS reductions from 0.435 to 0.269 for TMDM and from 7.272 to 3.239 for CSDI, as reported in Table 4. Together with the NsDiff results in Figure 5, these findings provide additional evidence that frequency-wise correction can mitigate output spectral distortion across different forecasting backbones, while the remaining discrepancies in CSDI indicate partial rather than complete spectral alignment.

B.8 Stepwise versus One-Time Correction

We compare Stepwise, which repeatedly applies spectral correction during reverse diffusion, with SSSC, which corrects each completed forecast sample once after sampling. Table 5 reports the results on ETTh1 using CSDI and TMDM, with the uncorrected forecasts (Base) serving as a reference. For CSDI, Stepwise increases CRPS from 0.5500 to 0.5813 and QICE from 4.538 to 4.744, worsening both metrics. For TMDM, it reduces CRPS from 0.4520 to 0.4266, but QICE increases from 2.821 to 5.585, indicating that the CRPS improvement comes with poorer calibration. In contrast, SSSC achieves the lowest CRPS and QICE for both backbones, with CRPS values of 0.5165 and 0.4041 for CSDI and TMDM, respectively.

A potential limitation of Stepwise is that subsequent denoising updates may counteract intermediate spectral adjustments and reintroduce discrepancies, resulting in a possible *rebound*. This provides a plausible explanation for its mixed results, although the final scores alone do not verify this trajectory-level mechanism. SSSC avoids this interaction by applying correction after the final denoising update. Stepwise also incurs repeated correction overhead within the sampling loop, with the correction workload increasing with the number of corrected diffusion steps. SSSC instead requires only one correction pass per completed forecast sample. Together, these results support

the one-time correction design in the evaluated settings, achieving lower CRPS and QICE while avoiding repeated in-loop correction.

Table 5: Comparison of Stepwise and SSSC on ETTh1. the best results for each backbone are highlighted in bold.

Model	Method	CRPS ↓	QICE ($\times 100$) ↓
CSDI	Base	0.5500	4.538
	Stepwise	0.5813	4.744
	SSSC	**0.5165**	**3.956**
TMDM	Base	0.4520	2.821
	Stepwise	0.4266	5.585
	SSSC	**0.4041**	**1.260**

B.9 SSSC Algorithm and Implementation Details

Algorithm 1 uses the discrete Fourier transform $\mathcal{F}$ for real-valued inputs along the forecasting horizon, retaining $F = \lfloor H/2 \rfloor + 1$ nonredundant frequency bins, with the corresponding inverse transform reconstructing a sequence of length H. Spectral estimation and correction use consistent data scales and Fourier normalization, and the frequency-wise gain is shared across channels and forecast samples. The operator $\widehat{\mathbb{E}}$ denotes empirical averaging over the indices indicated by its subscripts. The non-DC frequency bins $\{1, \ldots, F-1\}$ are partitioned into three predefined, nonempty bands, $\mathcal{B}_{\mathrm{lo}}$, $\mathcal{B}_{\mathrm{mid}}$, and $\mathcal{B}_{\mathrm{hi}}$, for computing the spectral concentration statistic. Low-frequency-protected candidates, denoted by the superscript lp, retain unit gain for $f < k_{\mathrm{lo}}$. The parameters satisfy $0 < g_{\min} \leq 1 \leq g_{\max} < \infty$, $\delta \in [0, 1)$, $\tau \in [0, 1]$, $\varepsilon > 0$, and $k_{\mathrm{lo}} \in \{0, \ldots, F\}$. For energy-gated shrinkage, we explicitly define $g_\eta^{\mathrm{sh}}(f) = 1$ whenever $w(f) = 0$, thereby avoiding evaluation of the undefined expression 0^0. During adaptive selection, each candidate is applied separately to the same original cached forecast ensembles and evaluated against the corresponding targets, rather than being applied sequentially to previously corrected forecasts.

Algorithm 1 Spectral-Shape SNR Correction (SSSC)

Input: Training-side calibration pairs $\mathcal{D}_{\mathrm{cal}} = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{N}$;
frozen diffusion forecaster; test history $\mathbf{X} \in \mathbb{R}^{L \times C}$; horizon H; sample count M;
gain bounds $g_{\min}, g_{\max}$; low-frequency cutoff k_{lo}; bands $\mathcal{B}_{\mathrm{lo}}, \mathcal{B}_{\mathrm{mid}}, \mathcal{B}_{\mathrm{hi}}$;
relative improvement margin δ; stand-down threshold τ; numerical floor ε.

Set $F \leftarrow \lfloor H/2 \rfloor + 1$ and $\varepsilon_\rho \leftarrow 10^{-30}$.
Define $\mathcal{T}_g(\mathbf{Z}) \leftarrow \mathcal{F}^{-1}[g \odot \mathcal{F}(\mathbf{Z})]$, with inverse output length H.
Offline calibration: Spectral Profile Estimation (SPE)
Generate and cache M completed forecasts $\{\widehat{\mathbf{Y}}_i^{(m)}\}_{m=1}^{M}$ for each $\mathbf{X}_i$ using the frozen native sampler.
Estimate the target and forecast spectra for $f = 0, \ldots, F-1$:

$$P(f) \leftarrow \widehat{\mathbb{E}}_{i,c}\big[|\mathcal{F}(\mathbf{Y}_{i,c})(f)|^2\big],$$
$$S(f) \leftarrow \widehat{\mathbb{E}}_{i,c,m}\Big[|\mathcal{F}(\widehat{\mathbf{Y}}_{i,c}^{(m)})(f)|^2\Big].$$

$r(f) \leftarrow P(f)/\max\{S(f), \varepsilon\}$; $\quad \bar{P} \leftarrow \dfrac{1}{F-1}\sum_{f=1}^{F-1} P(f)$.

$$w(f) \leftarrow \mathrm{clip}\left(\sqrt{\frac{P(f)}{\max\{\bar{P}, \varepsilon\}}}, 0, 1\right).$$

Offline calibration: Spectral Correction (SC)
Initialize $\mathcal{G} \leftarrow \{\mathbf{1}\}$.
for $\eta \in \{0.5, 1.0\}$ **do**
  $g_\eta(f) \leftarrow \mathrm{clip}\Big(r(f)^{\eta/2}, g_{\min}, g_{\max}\Big)$.
  Initialize $g_\eta^{\mathrm{sh}}(f) \leftarrow 1$ for all f; update only frequencies with $w(f) > 0$:
  $g_\eta^{\mathrm{sh}}(f) \leftarrow \min\Big\{1, \mathrm{clip}\Big(r(f)^{\eta w(f)/2}, g_{\min}, g_{\max}\Big)\Big\}$.
  Copy g_η and g_η^{sh} to g_η^{lp} and $g_\eta^{\mathrm{sh,lp}}$, respectively.
  Set $g_\eta^{\mathrm{lp}}(f) \leftarrow 1$ and $g_\eta^{\mathrm{sh,lp}}(f) \leftarrow 1$ for $f < k_{\mathrm{lo}}$.
  $\mathcal{G} \leftarrow \mathcal{G} \cup \{g_\eta, g_\eta^{\mathrm{sh}}, g_\eta^{\mathrm{lp}}, g_\eta^{\mathrm{sh,lp}}\}$.
end for
Offline calibration: Adaptive Selection (AS)

$$\bar{P}_b \leftarrow \frac{1}{|\mathcal{B}_b|}\sum_{f \in \mathcal{B}_b} P(f) \text{ for } b \in \{\mathrm{lo}, \mathrm{mid}, \mathrm{hi}\}.$$

$$\rho \leftarrow \frac{\bar{P}_{\mathrm{mid}} + \bar{P}_{\mathrm{hi}}}{\max\big\{P(0) + \bar{P}_{\mathrm{lo}} + \bar{P}_{\mathrm{mid}} + \bar{P}_{\mathrm{hi}}, \varepsilon_\rho\big\}}.$$

for each $g \in \mathcal{G}$ **do**
  Evaluate $\widehat{\mathcal{L}}_{\mathrm{CRPS}}(g)$ against the calibration targets by applying $\mathcal{T}_g$ to the original cached ensembles.
end for
$g^\star \leftarrow \arg\min_{g \in \mathcal{G}} \widehat{\mathcal{L}}_{\mathrm{CRPS}}(g)$; select $\mathbf{1}$ whenever it is a minimizer.
if $\rho < \tau$ **or** $\widehat{\mathcal{L}}_{\mathrm{CRPS}}(g^\star) > (1-\delta)\widehat{\mathcal{L}}_{\mathrm{CRPS}}(\mathbf{1})$ **then**
  $g_{\mathrm{final}} \leftarrow \mathbf{1}$.
else
  $g_{\mathrm{final}} \leftarrow g^\star$.
end if
Freeze g_{final} before testing.
Online inference
Generate $\{\widehat{\mathbf{Y}}^{(m)}\}_{m=1}^{M}$ conditioned on $\mathbf{X}$ using the unchanged native sampler.
for $m = 1, \ldots, M$ **do**
  $\widetilde{\mathbf{Y}}^{(m)} \leftarrow \mathcal{T}_{g_{\mathrm{final}}}(\widehat{\mathbf{Y}}^{(m)})$.
end for
Output: Corrected forecast ensemble $\{\widetilde{\mathbf{Y}}^{(m)}\}_{m=1}^{M}$.