SCTF: SELF-CORRECTING TIME-FREQUENCY DIFFUSION FOR MULTIVARIATE TIME-SERIES ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, diffusion models have shown considerable promise for multivariate time-series anomaly detection by reconstructing masked observations from the remaining context and using reconstruction errors as anomaly evidence. However, anomalies in the visible context can bias the reconstruction toward abnormal patterns, making the reconstruction appear more similar to the anomalous observations and thereby making the anomalies harder to distinguish. Existing strategies attempt to limit the influence of unreliable conditioning contexts, but such control relies on reliability estimates that may themselves be inaccurate and are often treated as fixed during reconstruction. This reveals a broader bottleneck: effective anomaly reconstruction depends not only on reconstruction capability, but also on whether conditioning reliability can be accurately estimated, controlled, and corrected. To address this bottleneck, we introduce SCTF (Self-Correcting Time-Frequency Diffusion), built on the principle that conditioning reliability should be treated as a dynamic quantity rather than a fixed judgment. SCTF estimates context risks from complementary views, uses them to adaptively reduce the influence of potentially unreliable contexts, and progressively refines the risks using reconstruction feedback, forming a self-correcting loop between reliability estimation and generative reconstruction. Synchronized temporal and frequency diffusion further allows frequency-domain reconstruction to provide auxiliary information for temporal denoising while offering complementary feedback for risk refinement. Extensive experiments on benchmark and real-world datasets demonstrate strong F1 performance against representative state-of-the-art methods.

1 INTRODUCTION

Multivariate time-series anomaly detection aims to identify abnormal observations that may indicate faults or unexpected operating conditions in monitored systems (Su et al., 2019; Deng & Hooi, 2021; Xu et al., 2022). However, anomalous events are often rare and difficult to annotate accurately because their identification may require substantial expert knowledge (Zamanzadeh Darban et al., 2024; Cho et al., 2025). Consequently, many methods learn normal patterns from available observations and detect anomalies based on deviations from these patterns (Su et al., 2019; Audibert et al., 2020; Tuli et al., 2022).

Reconstruction-based methods detect anomalies from discrepancies between reconstructed and observed values (Su et al., 2019; Audibert et al., 2020; Tuli et al., 2022). Their effectiveness depends on the learned reconstruction patterns: limited models may fail to capture complex normal behavior, whereas overly expressive models may also reconstruct anomalies and distort detection residuals (Song et al., 2023). Diffusion models have recently shown strong potential for time-series reconstruction (Tashiro et al., 2021; Chen et al., 2023), and many diffusion-based detectors reconstruct masked observations from the remaining visible context using reconstruction errors as anomaly evidence. ImDiffusion, for example, formulates this process as diffusion-based imputation and introduces noisy references to reduce anomaly influence (Chen et al., 2023). However, the visible context may itself contain anomalous patterns. As illustrated in Fig. 1, unreliable contexts can bias the reconstruction process toward abnormal patterns, distorting the resulting reconstruction errors. Such biased errors may underestimate anomalies or introduce false alarms in normal regions, mak-

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.

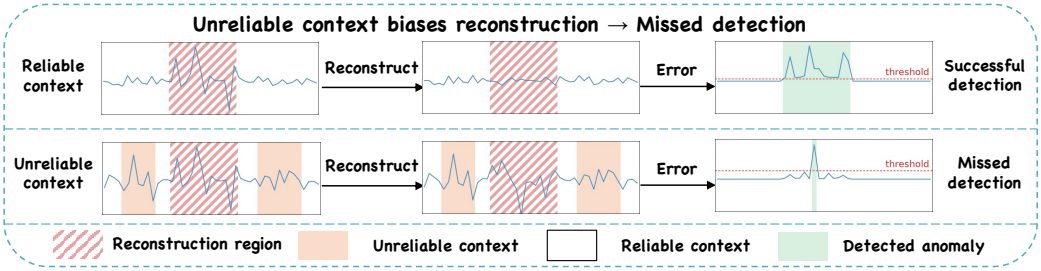


Figure 1: Illustration of how unreliable contexts bias diffusion-based anomaly reconstruction. Reliable contexts preserve reconstruction errors for anomalous regions, enabling successful detection. In contrast, unreliable contexts containing anomalous patterns bias the reconstruction process and distort anomaly-related reconstruction errors, resulting in missed detection.

ing reliable anomaly discrimination difficult. Therefore, improving reconstruction capability alone cannot resolve this problem; controlling the reliability of conditioning contexts is equally important. The challenge consequently extends to how conditioning reliability can be accurately estimated, effectively controlled, and progressively corrected when the context may be contaminated.

Recent diffusion-based studies address this issue by controlling the information used for reconstruction. DiffAD selects likely normal observations using a density-ratio criterion and gradually increases their contribution during denoising (Xiao et al., 2023), while DDMT adaptively masks suspicious observations and their surrounding regions (Yang et al., 2023a), demonstrating the importance of regulating unreliable conditions. However, these reliability judgments may themselves be inaccurate, and incorrect assessments can continue to affect reconstruction if they are not revised using reconstruction feedback. Moreover, evidence from a single representation may be insufficient for reliably identifying contaminated conditions, motivating the use of complementary information from multiple views (Jeong & Yoon, 2026; Mao et al., 2025). CATCH combines frequency-domain representations with cross-channel dependencies for anomaly discrimination (Wu et al., 2025), while MODEM shows that multi-resolution and frequency information can improve diffusion-based reconstruction (Zhong et al., 2025). These observations suggest that conditioning reliability should not be treated as a fixed judgment, but as a quantity that can be estimated from complementary views, used to regulate unreliable conditions, and progressively corrected using reconstruction feedback.

To this end, we propose SCTF, a Self-Correcting Time–Frequency Diffusion framework for multivariate time-series anomaly detection, built on this principle of dynamic conditioning reliability. SCTF establishes a closed-loop interaction between reliability estimation and generative reconstruction. Specifically, complementary temporal, frequency, and cross-channel views are integrated to estimate context risks. These risks determine the corruption applied to conditioning contexts, such that higher-risk contexts exert less influence on reconstruction. Since the initial risks may be inaccurate, temporal and frequency reconstruction errors are periodically fed back to refine them during training, forming a self-correcting loop between reliability estimation and generative reconstruction. Meanwhile, synchronized temporal and frequency diffusion allows frequency-domain reconstruction to provide auxiliary information for temporal denoising while supplying complementary feedback for risk refinement.

Our main contributions are summarized as follows:

- We identify conditioning reliability as a key bottleneck in diffusion-based anomaly reconstruction and develop a self-correcting framework that treats it as a quantity to be estimated, regulated, and progressively corrected rather than a fixed judgment. Multi-view context risks control unreliable conditions through risk-adaptive corruption.
- We develop reconstruction-driven risk refinement together with synchronized temporal and frequency diffusion. Reconstruction feedback periodically corrects context risks, while frequency-domain reconstruction provides auxiliary information for temporal denoising and complementary feedback for risk refinement.
- We conduct extensive experiments on benchmark and real-world datasets, demonstrating the effectiveness and generalization ability of SCTF.

2 Related Work

A widely used approach to multivariate time-series anomaly detection is to reconstruct observed sequences and identify anomalies from reconstruction errors (Zamanzadeh Darban et al., 2024). Reconstruction-based methods include OmniAnomaly, USAD, and TranAD (Su et al., 2019; Audibert et al., 2020; Tuli et al., 2022), while other methods such as Anomaly Transformer and DCdetector detect anomalies by modeling deviations from normal temporal patterns (Xu et al., 2022; Yang et al., 2023c). More recently, diffusion models have been introduced for robust time-series reconstruction. ImDiffusion performs masked diffusion imputation (Chen et al., 2023), DDMT applies adaptive masking (Yang et al., 2023a), DiffAD regulates conditional contexts during denoising (Xiao et al., 2023), MODEM incorporates multi-resolution and frequency modeling (Zhong et al., 2025), and ReCDiff improves conditional guidance under anomaly-contaminated inputs (Xu et al., 2026). These studies highlight the importance of controlling unreliable reconstruction conditions, but the reliability judgments themselves may remain inaccurate when anomalies exhibit diverse characteristics.

Complementary temporal, frequency, and cross-channel representations have also been explored for anomaly detection. TFAD and CATCH exploit frequency-domain information, with CATCH further modeling cross-channel dependencies (Zhang et al., 2022; Wu et al., 2025), while PGRF-Net combines multiple anomaly cues from prediction, variable relations, contextual deviation, and local spikes (Jeong & Yoon, 2026). These studies show that complementary information can improve anomaly characterization, but it is primarily used for anomaly discrimination or representation learning. In contrast, SCTF uses complementary views to estimate conditioning reliability and treats it as a dynamic quantity rather than a fixed judgment: context risks reduce the influence of potentially unreliable conditions and are progressively refined using reconstruction feedback. Synchronized temporal and frequency diffusion further allows frequency-domain reconstruction to provide auxiliary information for temporal denoising while supplying complementary feedback for risk refinement, forming a closed loop between conditioning-reliability estimation and generative reconstruction.

3 Proposed Method

3.1 Problem Statement

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]^\top \in \mathbb{R}^{T \times C}$ denote a multivariate time series with T timestamps and C variables. The goal of multivariate time-series anomaly detection is to estimate an anomaly score s_t for each timestamp t, where higher values indicate greater anomaly likelihood. We consider a reconstruction-based setting, where normal patterns are learned from available training observations and anomalies are identified through reconstruction errors. Since reconstruction conditions may contain anomalies, we aim to prevent unreliable contexts from misleading the generative process and obtain reliable normal references for detection.

3.2 Overview of SCTF

As illustrated in Fig. 2, SCTF establishes a closed-loop interaction between context reliability estimation and time–frequency diffusion reconstruction through the closely connected components. First, multi-view context risk estimation integrates complementary temporal, frequency, and cross-channel views to estimate the reliability of conditioning contexts. These risks determine the corruption applied to conditioning contexts, reducing the influence of potentially unreliable information on reconstruction. Second, synchronized temporal and frequency diffusion branches reconstruct masked targets in parallel. During reverse diffusion, the frequency-domain clean-state estimate is projected into the temporal space and coupled with the temporal clean-state estimate, providing auxiliary information for temporal reconstruction. Finally, temporal and frequency reconstruction errors provide complementary feedback for periodically refining the context risks during training. The refined risks are fed back to update the adaptive corruption process, progressively correcting conditioning reliability and forming a self-correcting loop between reliability estimation and generative reconstruction.

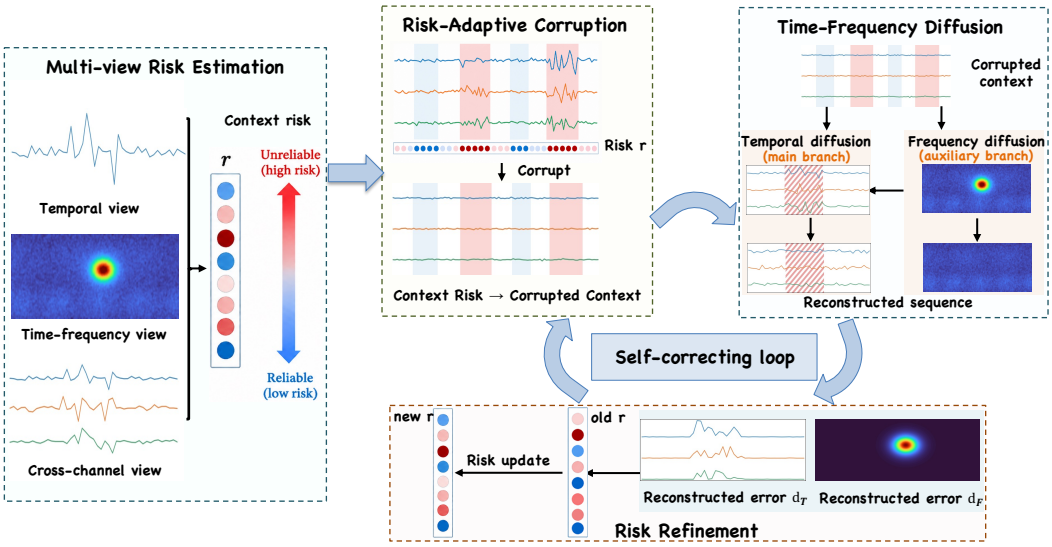


Figure 2: Overview of SCTF. Multi-view evidence estimates context risks, which are used for risk-adaptive context corruption to reduce the influence of unreliable contexts. Synchronized time–frequency diffusion reconstructs masked targets with frequency-assisted temporal reconstruction. Reconstruction errors further refine the context risks, forming a self-correcting loop.

3.3 Multi-View Context Risk Estimation

To provide complementary evidence for context-risk estimation, we construct three priors that characterize time–frequency energy, local time-domain amplitude, and cross-channel deviation, respectively. For each timestamp t, the three views produce prior scores $\{p_m(t)\}_{m=1}^{3}$, where a larger value indicates stronger evidence of abnormality. These complementary priors are then integrated to estimate the initial context risk $r^{(0)}(t)$.

Time–frequency energy. We use a wavelet transform to characterize localized energy variations in both time and frequency (Mallat, 1989; Daubechies, 1990). For each channel, wavelet coefficients are computed over multiple scales covering different frequency bands, and their squared magnitudes are aggregated to obtain the local time–frequency energy $E_c(t)$. The prior is then defined as

$$p_1(t) = \max_c E_c(t), \tag{1}$$

which retains the strongest time–frequency energy response across channels. A large $p_1(t)$ therefore indicates a pronounced local energy variation around timestamp t.

Local amplitude. We further characterize abnormal variations in the time domain. After z-score standardizing each channel, we compute the joint amplitude across channels and smooth it over a local temporal window to suppress isolated fluctuations. The resulting prior is

$$p_2(t) = \frac{1}{|\mathcal{A}(t)|} \sum_{\tau \in \mathcal{A}(t)} \sqrt{\sum_{c=1}^{C} X_{\tau,c}^2}, \tag{2}$$

where $\mathcal{A}(t)$ denotes the local temporal window centered around timestamp t. A large $p_2(t)$ indicates a strong local amplitude response (Allen, 1978; Withers et al., 1998), providing complementary evidence from the time domain.

Cross-channel deviation. Anomalies may occur only in a subset of channels, causing their local behavior to deviate from the remaining channels. We therefore measure cross-channel consistency within a local temporal window. Let $\mathbf{R}(t)$ denote the local inter-channel correlation matrix, where $|R_{ij}(t)|$ measures the correlation strength between channels i and j around timestamp t. The prior

is defined as

$$p_3(t) = 1 - \frac{1}{C(C-1)} \sum_{i \neq j} |R_{ij}(t)|, \tag{3}$$

where a larger $p_3(t)$ indicates stronger cross-channel deviation and thus stronger evidence of channel-specific abnormality (Gibbons & Ringdal, 2006; Schimmel & Paulssen, 1997).

Multi-view fusion. The three prior scores are first standardized using z-score normalization to eliminate scale differences across views (Bishop, 2006). Before reconstruction feedback becomes available, we initialize the context risk by equally combining the standardized priors as $r^{(0)}(t) = \sigma\left(\frac{1}{3}\sum_{m=1}^{3} p_m(t)\right)$, where $\sigma(\cdot)$ maps the fused score to $[0, 1]$. After the first refinement round, we employ a lightweight temporal convolutional network (TCN) $\mathcal{G}_\psi$ to adaptively fuse the three prior sequences, yielding $\mathbf{r}^{(l)} = \sigma(\mathcal{G}_\psi(p_1, p_2, p_3))$, $l \geq 1$. Here, p_m denotes the standardized sequence of the m-th prior, and $r^{(l)}(t)$ is the estimated context risk at timestamp t. The TCN is progressively optimized using reconstruction-derived supervision through the self-correcting refinement process described in Sec. 3.5.

3.4 Risk-Adaptive Time–Frequency Diffusion

The context risk estimated in Sec. 3.3 indicates which parts of the context may provide unreliable conditioning information. We incorporate this risk into conditional diffusion by adaptively corrupting the observed context. Meanwhile, synchronized diffusion processes are constructed in the temporal and frequency domains, whose clean-state estimates are coupled during reverse generation.

During reconstruction, four complementary random masks are used so that different parts of the sequence are reconstructed in turn. Specifically, the four masks are mutually exclusive and collectively cover the entire sequence, i.e., $\sum_{i=1}^{4} \mathbf{M}_i = \mathbf{1}$ and $\mathbf{M}_i \odot \mathbf{M}_j = 0$ for $i \neq j$. This design allows reconstruction errors to be obtained for all timestamps and aggregated to update anomaly scores over the entire sequence.

Given a reconstruction mask $\mathbf{M}$, we divide the input into masked targets and observed context:

$$\mathbf{X}^m = \mathbf{M} \odot \mathbf{X}, \qquad \mathbf{X}^o = (\mathbf{1} - \mathbf{M}) \odot \mathbf{X}. \tag{4}$$

Risk-Adaptive Context Corruption. Let $\{\beta_k\}_{k=1}^{K}$ denote the diffusion noise schedule, $\alpha_k = 1 - \beta_k$, and $\bar{\alpha}_k = \prod_{j=1}^{k} \alpha_j$. For an observed position i with risk $r_i \in [0, 1]$, we define $\bar{\alpha}_{k,i}^{r} = \bar{\alpha}_k^{1+\rho r_i}$, where $\rho \geq 0$ controls the strength of risk adaptation. Since $0 < \bar{\alpha}_k < 1$, a higher risk produces a smaller $\bar{\alpha}_{k,i}^{r}$ and therefore stronger corruption.

The observed context is corrupted as

$$\widetilde{x}_{k,i}^{o} = \sqrt{\bar{\alpha}_{k,i}^{r}}\, x_i^o + \sqrt{1 - \bar{\alpha}_{k,i}^{r}}\, \xi_i, \qquad \xi_i \sim \mathcal{N}(0, 1). \tag{5}$$

Thus, low-risk observations preserve more of their original information, whereas high-risk contexts exert weaker conditional influence.

The masked reconstruction targets follow the standard forward diffusion:

$$\mathbf{X}_k^m = \sqrt{\bar{\alpha}_k}\mathbf{X}_0^m + \sqrt{1 - \bar{\alpha}_k}\boldsymbol{\epsilon}_T, \qquad \boldsymbol{\epsilon}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{6}$$

Synchronized Time–Frequency Diffusion. We further introduce a frequency-domain diffusion branch to complement the temporal reconstruction. Applying the wavelet transform $\mathcal{W}(\cdot)$ to each channel and aggregating the coefficients into B frequency bands gives $\mathbf{F}_0 = \mathcal{W}(\mathbf{X}_0) \in \mathbb{R}^{T \times C \times B}$. Since the temporal axis is preserved, the same reconstruction mask is applied to obtain $\mathbf{F}_0^m$ and $\mathbf{F}_0^o$.

The masked frequency target follows the same diffusion schedule:

$$\mathbf{F}_k^m = \sqrt{\bar{\alpha}_k}\mathbf{F}_0^m + \sqrt{1 - \bar{\alpha}_k}\boldsymbol{\epsilon}_F, \qquad \boldsymbol{\epsilon}_F \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{7}$$

while its observed context is corrupted using the same risk-adaptive coefficient:

$$\widetilde{\mathbf{f}}_{k,i}^{o} = \sqrt{\bar{\alpha}_{k,i}^{r}}\, \mathbf{f}_i^o + \sqrt{1 - \bar{\alpha}_{k,i}^{r}}\, \boldsymbol{\xi}_{F,i}, \qquad \boldsymbol{\xi}_{F,i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{8}$$

At diffusion step k, the temporal and frequency denoisers independently predict the injected noise as $\widehat{\boldsymbol{\epsilon}}_{T,k} = \epsilon_{\theta_T}\left(\mathbf{X}_k^m, \widetilde{\mathbf{X}}_k^o, k\right)$ and $\widehat{\boldsymbol{\epsilon}}_{F,k} = \epsilon_{\theta_F}\left(\mathbf{F}_k^m, \widetilde{\mathbf{F}}_k^o, k\right)$, respectively. The two denoisers are trained using the corresponding noise-prediction objectives:

$$\mathcal{L}_T = \mathbb{E}_{k,\boldsymbol{\epsilon}_T}\left[\|\boldsymbol{\epsilon}_T - \widehat{\boldsymbol{\epsilon}}_{T,k}\|_2^2\right], \qquad \mathcal{L}_F = \mathbb{E}_{k,\boldsymbol{\epsilon}_F}\left[\|\boldsymbol{\epsilon}_F - \widehat{\boldsymbol{\epsilon}}_{F,k}\|_2^2\right]. \tag{9}$$

Following DDIM (Song et al., 2021), the predicted noise gives the clean-state estimates

$$\widehat{\mathbf{X}}_0^{(k)} = \frac{\mathbf{X}_k^m - \sqrt{1-\bar{\alpha}_k}\,\widehat{\boldsymbol{\epsilon}}_{T,k}}{\sqrt{\bar{\alpha}_k}}, \qquad \widehat{\mathbf{F}}_0^{(k)} = \frac{\mathbf{F}_k^m - \sqrt{1-\bar{\alpha}_k}\,\widehat{\boldsymbol{\epsilon}}_{F,k}}{\sqrt{\bar{\alpha}_k}}. \tag{10}$$

Cross-Domain Clean-State Coupling. To transfer frequency information to temporal reconstruction, we map the frequency-domain clean-state estimate into the temporal space. Since the frequency representation consists of aggregated bands and is not directly invertible, we use a lightweight three-layer feed-forward projection with ReLU activations, $\mathcal{P}_\phi$, to obtain $\mathbf{H}_k = \mathcal{P}_\phi\big(\widehat{\mathbf{F}}_0^{(k)}\big)$.

The temporal clean-state estimate is refined by incorporating frequency-domain information:

$$\overline{\mathbf{X}}_0^{(k)} = (1-\lambda_k)\widehat{\mathbf{X}}_0^{(k)} + \lambda_k \mathbf{H}_k, \tag{11}$$

where $\lambda_k \in [0,1]$ controls the contribution of frequency information. Previous studies show that reverse diffusion exhibits a coarse-to-fine spectral evolution, where low-frequency structures are recovered earlier and high-frequency details are progressively restored at later denoising steps (Yang et al., 2023b; Qian et al., 2024). Motivated by this behavior, we progressively increase λ_k during reverse denoising. The schedule is given in Appendix A.4.

To jointly optimize the cross-domain projection and the two diffusion branches, we introduce the clean-state coupling loss

$$\mathcal{L}_C = \mathbb{E}_k\left[\left\|\overline{\mathbf{X}}_0^{(k)} - \mathbf{X}_0^m\right\|_2^2\right]. \tag{12}$$

The overall objective is therefore

$$\mathcal{L}_{\text{diff}} = \omega_T \mathcal{L}_T + \omega_F \mathcal{L}_F + \omega_C \mathcal{L}_C, \tag{13}$$

where ω_T, ω_F, and ω_C control the contributions of temporal denoising, frequency denoising, and cross-domain coupling, respectively. $\mathcal{L}_T$ and $\mathcal{L}_F$ supervise the temporal and frequency denoisers, while $\mathcal{L}_C$ jointly optimizes the two denoisers and projection parameters ϕ. All modules are optimized using $\mathcal{L}_{\text{diff}}$ with losses evaluated on the corresponding reconstruction targets.

After obtaining the frequency-refined clean estimate $\overline{\mathbf{X}}_0^{(k)}$, we recompute the temporal noise as $\overline{\boldsymbol{\epsilon}}_{T,k} = (\mathbf{X}_k^m - \sqrt{\bar{\alpha}_k}\overline{\mathbf{X}}_0^{(k)})/\sqrt{1-\bar{\alpha}_k}$ to maintain consistency with the current diffusion state.

Finally, the synchronized reverse updates are

$$\mathbf{X}_{k-1}^m = \sqrt{\bar{\alpha}_{k-1}}\,\overline{\mathbf{X}}_0^{(k)} + \sqrt{1-\bar{\alpha}_{k-1}}\,\overline{\boldsymbol{\epsilon}}_{T,k}, \qquad \mathbf{F}_{k-1}^m = \sqrt{\bar{\alpha}_{k-1}}\,\widehat{\mathbf{F}}_0^{(k)} + \sqrt{1-\bar{\alpha}_{k-1}}\,\widehat{\boldsymbol{\epsilon}}_{F,k}. \tag{14}$$

The two branches therefore evolve synchronously, while the frequency-domain clean estimate progressively assists the temporal reverse trajectory.

3.5 Self-Correcting Risk Refinement

The initial context risk $\mathbf{r}^{(0)}$ is estimated from the multi-view priors and may be inaccurate for some observations. The refinement process itself does not require anomaly labels; instead, we periodically use reconstruction information from the risk-adaptive time–frequency diffusion during training to update the context risks. Let $\mathbf{X}_{\text{rec}}^{(l)}$ and $\mathbf{F}_{\text{rec}}^{(l)}$ denote the temporal- and frequency-domain reconstructions at the l-th refinement round, respectively.

We first compute the temporal and frequency-domain reconstruction error as

$$d_T^{(l)}(t) = \frac{1}{C}\sum_{c=1}^{C}\left|X_{t,c} - X_{\text{rec},t,c}^{(l)}\right|, \qquad d_F^{(l)}(t) = \frac{1}{CB}\sum_{c=1}^{C}\sum_{b=1}^{B}\left|F_{t,c,b} - F_{\text{rec},t,c,b}^{(l)}\right|, \tag{15}$$

where B denotes the number of frequency bands. The two reconstruction errors provide complementary feedback from the temporal and frequency domains.

Before risk refinement, $d_T^{(l)}$ and $d_F^{(l)}$ are normalized using min–max normalization to map the reconstruction errors into the range of $[0, 1]$. We then construct a reconstruction-refined target risk by combining the current context risk with the normalized errors:

$$r_{\text{tar}}^{(l+1)}(t) = (1 - \gamma_1 - \gamma_2) r^{(l)}(t) + \gamma_1 d_T^{(l)}(t) + \gamma_2 d_F^{(l)}(t), \tag{16}$$

where $d_T^{(l)}$ and $d_F^{(l)}$ denote the normalized temporal and frequency reconstruction errors, respectively, and $\gamma_1, \gamma_2 \geq 0$ with $\gamma_1 + \gamma_2 \leq 1$ control the contributions of temporal and frequency feedback. The target risk $\mathbf{r}_{\text{tar}}^{(l+1)}$ is then used to self-supervise the TCN $\mathcal{G}_\psi$ introduced in Sec. 3.3. Specifically, after each refinement round, we optimize ψ by minimizing

$$\mathcal{L}_R^{(l)} = \frac{1}{T} \left\| \sigma(\mathcal{G}_\psi(p_1, p_2, p_3)) - \mathbf{r}_{\text{tar}}^{(l+1)} \right\|_2^2. \tag{17}$$

During this optimization, the diffusion networks are fixed and $\mathbf{r}_{\text{tar}}^{(l+1)}$ serves as the reconstruction-derived supervision target. After updating the TCN, the refined context risk is obtained as $\mathbf{r}^{(l+1)} = \sigma(\mathcal{G}_\psi(p_1, p_2, p_3))$. Through iterative refinement, the TCN learns to map multi-view priors to reconstruction-refined context risks.

3.6 Training and Inference

During training, SCTF estimates context risks from the multi-view priors and performs risk-adaptive corruption on the conditioning contexts. The corrupted contexts are then used for diffusion reconstruction, whose reconstruction errors provide feedback for periodically refining the risk estimator and form the self-correcting training loop.

During inference, the risk estimator is fixed without further refinement or parameter updates. Given a test window, SCTF estimates context risks, applies risk-adaptive corruption to the conditioning contexts, and reconstructs the sequence using DDIM sampling. For each reconstruction mask, the condition-corruption noises are sampled once before reverse diffusion and kept fixed throughout sampling, while the corrupted conditions are recomputed at each reverse step according to the step-dependent risk-adaptive diffusion coefficient. The four complementary masks are then applied in turn, and their reconstructed regions are assembled into the final temporal reconstruction as $\mathbf{X}_{\text{rec}} = \sum_{j=1}^{4} \mathbf{M}_j \odot \mathbf{X}_{\text{rec}}^{(j)}$, where $\mathbf{X}_{\text{rec}}^{(j)}$ denotes the reconstruction obtained under mask $\mathbf{M}_j$. Since the masks are mutually exclusive and collectively cover the entire sequence, each timestamp receives exactly one reconstructed value. The anomaly score is defined as

$$s_t = \frac{1}{C} \sum_{c=1}^{C} \left| X_{t,c} - X_{\text{rec},t,c} \right|, \tag{18}$$

where a larger s_t indicates a stronger deviation from the reconstructed normal pattern. The detection threshold is computed from validation anomaly scores and applied to the test set.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate SCTF on five widely used multivariate time-series anomaly detection benchmarks, including SMD (Su et al., 2019), MSL (Hundman et al., 2018), SMAP (Hundman et al., 2018), PSM (Abdulaal et al., 2021), and SWaT (Goh et al., 2017), covering server monitoring, spacecraft telemetry, and industrial control scenarios. To further assess cross-domain generalization, we additionally evaluate SCTF on four real-world signal datasets, including STEAD (Mousavi et al., 2019), OBST2024 (Niksejel & Zhang, 2024), ShipsEar (Santos-Domínguez et al., 2016), and DAS-MedSea (Bouffaut et al., 2025), which represent seismic and underwater acoustic monitoring scenarios.

Table 1: Precision, Recall, and F1-score (%) on five multivariate time-series anomaly detection benchmarks under the PA evaluation protocol. Avg. F1 denotes the unweighted average F1-score across all benchmarks. Baseline results are collected from published studies under the corresponding PA settings; details are provided in Appendix B.5. The best results are highlighted in bold.

METHOD	SMAP			SMD			PSM			MSL			SWaT			AVG. F1
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	
SCTF	95.14	99.02	**97.04**	91.77	96.50	94.08	**98.23**	98.71	**98.47**	**96.31**	94.19	95.24	**97.42**	96.98	97.20	**96.41**
DiffAD	91.08	87.39	89.19	90.32	95.71	93.40	96.22	97.70	96.95	85.59	88.32	86.93	86.64	88.35	87.49	90.79
ImDiffusion	87.71	96.18	91.75	94.53	94.48	94.50	97.72	96.83	97.27	87.74	84.65	86.16	89.88	84.65	87.09	91.35
DDMT	94.08	99.08	96.51	88.07	81.50	84.66	97.92	98.40	98.16	92.67	98.07	95.30	96.56	**99.78**	**97.24**	94.37
Anomaly Transformer	85.36	88.03	86.67	87.26	89.70	88.46	92.31	93.29	92.80	86.34	91.08	88.64	80.57	82.34	81.45	87.60
MEMTO	93.76	**99.63**	96.61	89.13	**98.40**	93.54	97.46	**99.23**	98.34	92.07	96.76	94.36	94.18	97.54	95.83	95.74
DCdetector	**95.63**	98.92	97.02	83.59	91.10	87.18	97.14	98.74	97.94	93.69	**99.69**	**96.60**	93.11	99.77	96.33	95.01
MODEM	88.50	97.22	92.66	**95.70**	96.32	**96.01**	96.97	98.35	97.65	91.28	88.32	89.77	89.42	93.08	91.21	93.46
D3R	86.80	91.76	89.21	86.50	95.34	90.71	88.17	93.25	90.63	78.48	94.33	85.67	80.17	87.40	83.63	87.97
TFAD	83.66	86.72	85.16	92.42	91.36	91.89	90.27	97.21	93.83	87.15	89.78	88.44	73.26	84.59	78.52	87.57
TranAD	82.24	85.02	83.60	89.06	89.82	87.85	95.06	91.47	92.20	89.51	92.97	91.15	70.25	72.66	68.86	84.73

Following previous studies, we use Precision (P), Recall (R), and F1-score (F1) for anomaly detection evaluation, with F1 as the primary metric. We additionally report AUPRC to measure threshold-independent detection performance. Higher values indicate better performance for all metrics. Detailed dataset descriptions, statistics, and evaluation protocols are provided in Appendix B.1 and Appendix B.2.

4.2 Implementation Details

SCTF is implemented using PyTorch and trained on two NVIDIA RTX 3090 GPUs. We optimize the model with Adam using a learning rate of 3×10^{-4}, weight decay of 1×10^{-6}, and a batch size of 16 for 50 epochs. The diffusion process contains $K = 50$ steps, with DDIM sampling using 10 steps for reconstruction. The loss weights are set to $\omega_T = 1$, $\omega_F = 0.5$, and $\omega_C = 0.05$. We use $B = 3$ frequency bands and refine the context risk every 10 epochs. Hyperparameters are selected on the labeled validation split according to the validation F1-score, and additional sensitivity analyses are provided in Appendix B.3. After hyperparameter selection, the detection threshold is independently calibrated as $\tau_d = \mu_d + \sigma_d$ from the distribution of validation anomaly scores, without using validation labels or a predefined anomaly ratio (Xiao et al., 2023). Unless otherwise specified, SCTF results are averaged over three independent runs with different random seeds. Further implementation details are provided in Appendix B.4.

4.3 Detection Results

Table 1 compares SCTF with representative time-series anomaly detection methods on five widely used benchmarks under the PA evaluation protocol. The baseline values are collected from published studies under the corresponding PA settings, as detailed in Appendix B.5. SCTF achieves the highest reported average F1-score across the five benchmarks, with the best F1-score on SMAP and PSM and competitive performance on SMD, MSL, and SWaT. The average F1-score is computed as the unweighted arithmetic mean across the five datasets, and the mean and standard deviation of SCTF over three independent runs are reported in Appendix B.6. These results are consistent with the roles of multi-view risk estimation and risk-adaptive context corruption in reducing the influence of unreliable contexts during reconstruction.

To complement the PA comparison with a unified evaluation setting, we further conduct experiments under a no-point-adjustment protocol. As shown in Table 2, all compared methods are evaluated using the same protocol, input configuration, and window length. SCTF achieves the highest average F1-score and the best F1-score on three of the four datasets, while maintaining competitive AUPRC performance across diverse sensing scenarios. The consistent performance across heterogeneous domains further supports the effectiveness of self-correcting risk refinement and synchronized time–frequency diffusion in adapting risk estimation and reconstruction to diverse anomaly characteristics. Detailed comparisons are provided in Appendix B.7.

4.4 Ablation Study

We conduct ablation studies on STEAD and PSM to examine the main components of SCTF. Table 3 summarizes the representative variants, with detailed comparisons provided in Appendix B.8–B.11.

Table 2: F1-score and AUPRC (%) on generalization datasets. Best results are highlighted in bold.

Method	STEAD		ShipsEar		OBST		DAS-MedSea		Avg. F1
	F1	AUPRC	F1	AUPRC	F1	AUPRC	F1	AUPRC	
SCTF	**98.17** $\pm$ 0.38	**99.37** $\pm$ 0.31	**97.80** $\pm$ 0.55	99.73 $\pm$ 0.24	92.79 $\pm$ 0.57	**96.29** $\pm$ 0.35	**88.59** $\pm$ 0.62	**94.42** $\pm$ 0.28	**94.34**
DiffAD	82.88 $\pm$ 0.88	93.04 $\pm$ 0.66	94.90 $\pm$ 0.76	**99.93** $\pm$ 0.05	81.36 $\pm$ 0.86	89.65 $\pm$ 0.71	74.77 $\pm$ 0.82	76.62 $\pm$ 0.57	83.48
ImDiffusion	90.97 $\pm$ 0.74	96.02 $\pm$ 0.58	96.87 $\pm$ 0.61	99.91 $\pm$ 0.08	**94.47** $\pm$ 0.67	96.06 $\pm$ 0.55	82.98 $\pm$ 0.79	74.59 $\pm$ 0.62	91.32
Anomaly Transformer	87.54 $\pm$ 0.80	91.01 $\pm$ 0.64	93.71 $\pm$ 0.68	98.56 $\pm$ 0.55	88.24 $\pm$ 0.76	88.25 $\pm$ 0.68	68.77 $\pm$ 0.92	76.18 $\pm$ 0.60	84.56
CATCH	93.94 $\pm$ 0.64	96.69 $\pm$ 0.56	96.85 $\pm$ 0.61	93.98 $\pm$ 0.54	89.44 $\pm$ 0.72	90.85 $\pm$ 0.63	79.34 $\pm$ 0.81	81.29 $\pm$ 0.59	89.89

Table 3: Ablation study of the main components of SCTF on STEAD and PSM.

Component	Variant	STEAD F1	PSM F1
Context Corruption	No Isolation	82.93 ± 1.87	86.68 ± 1.73
	Hard Isolation ($q = 30\%$)	95.23 ± 0.64	94.15 ± 0.69
Risk Refinement	w/o Self-Correcting	83.70 ± 2.03	87.77 ± 1.96
TF Diffusion	Temporal Only	77.10 ± 1.83	83.05 ± 1.67
	TF Feature Conditioning	80.37 ± 1.42	84.96 ± 1.28
	Parallel TF Diffusion	93.87 ± 0.71	90.53 ± 0.83
Multi-View Priors	Full $-p_1$	86.21 ± 1.06	93.89 ± 0.91
	Full $-p_2$	88.59 ± 0.97	94.90 ± 0.88
	Full $-p_3$	89.33 ± 1.14	95.98 ± 0.83
Full SCTF (Ours)		$\mathbf{98.17 \pm 0.38}$	$\mathbf{98.47 \pm 0.29}$

Risk-Adaptive Context Corruption. We compare *No Isolation*, which uses all contexts as conditions, *Hard Isolation*, which removes the top-q high-risk contexts, and the full adaptive corruption that continuously adjusts the corruption strength according to each context risk. Suppressing unreliable conditions improves detection, while adaptive corruption performs best by weakening suspicious contexts without discarding potentially useful context.

Self-Correcting Risk Refinement. The *w/o Self-Correcting* variant keeps the initial multi-view context risks fixed throughout training, whereas the full model periodically updates the risk estimator using temporal and frequency reconstruction errors as feedback. The clear improvement of the full model on both datasets shows that reconstruction feedback can correct inaccurate initial risk estimates and provide more reliable conditioning information.

Synchronized TF Diffusion. We compare four ways of incorporating frequency information: *Temporal Only* removes the frequency branch; *TF Feature Conditioning* uses frequency information only as auxiliary features without frequency-domain diffusion; *Parallel TF Diffusion* performs temporal and frequency diffusion independently without cross-domain coupling; and the full model synchronizes the two branches by using the frequency-domain clean-state estimate to assist temporal reverse diffusion. The results show progressive improvements from feature conditioning to explicit dual-domain diffusion and further to synchronized reconstruction, supporting the benefit of cross-domain coupling.

Multi-View Anomaly Priors. We remove the time–frequency energy prior p_1, local amplitude prior p_2, and cross-channel deviation prior p_3 individually from the full risk estimator while keeping the remaining components unchanged. Removing any prior degrades performance, showing that the three views provide complementary evidence for context-risk estimation.

5 Conclusion

This paper presents SCTF for multivariate time-series anomaly detection, built on the principle that conditioning reliability should be treated as a dynamic quantity rather than a fixed judgment. SCTF estimates context risks from complementary views, reduces the influence of unreliable conditioning contexts, and progressively refines the risks through reconstruction feedback, forming a self-correcting loop between reliability estimation and generative reconstruction. Synchronized temporal and frequency diffusion further provides auxiliary frequency-domain information for temporal reconstruction and feedback for risk refinement. Experiments on benchmark and real-world datasets demonstrate the effectiveness and generalization ability of SCTF under conventional and no-point-adjustment protocols. These results suggest that conditioning reliability is an important bottleneck beyond reconstruction capability alone.

AI USE STATEMENT

We used ChatGPT and Codex to assist with data cleaning, repeated experimental runs, manuscript editing, and literature search. The tools also provided limited feedback on experimental setup and revisions, as well as checks of mathematical derivations, notation, and technical descriptions. Codex was additionally used to assist with repeated experiments under different random seeds. Generative AI was not used to independently originate the core methodology, determine the final experimental design, or draw research conclusions. All AI-assisted suggestions were critically reviewed by the authors. Mathematical suggestions were independently re-derived and checked for theoretical consistency, while experimental procedures and results were manually examined by the research team. The repeated experiments conducted with different random seeds showed consistent performance trends without substantial deviations from the initial results. The authors take full responsibility for the final methodology, experiments, analyses, claims, and manuscript.

REFERENCES

Ahmed Abdulaal, Zhuanghua Liu, and Tomer Lancewicki. Practical approach to asynchronous multivariate time series anomaly detection and localization. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021.

Rex V. Allen. Automatic earthquake recognition and timing from single traces. *Bulletin of the Seismological Society of America*, 68(5):1521–1532, 1978. doi: 10.1785/BSSA0680051521.

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A. Zuluaga. USAD: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020. doi: 10.1145/3394486.3403392.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Léa Bouffaut, Quentin Goestchel, Anthony Sladen, Arthur Hartog, and Holger Klinck. Distributed acoustic sensing fin whale labeled dataset for level estimations, 2025.

Yuhang Chen, Chaoyun Zhang, Minghua Ma, Yudong Liu, Ruomeng Ding, Bowen Li, Shilin He, Saravan Rajmohan, Qingwei Lin, and Dongmei Zhang. ImDiffusion: Imputed diffusion models for multivariate time series anomaly detection. *Proceedings of the VLDB Endowment*, 17(3):359–372, 2023. doi: 10.14778/3632093.3632101.

Dongchan Cho, Jiho Han, Keumyeong Kang, Minsang Kim, Honggyu Ryu, and Namsoon Jung. Structured temporal causality for interpretable multivariate time series anomaly detection. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Ingrid Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, 36(5):961–1005, 1990. doi: 10.1109/18.57199.

Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4027–4035, 2021. doi: 10.1609/aaai.v35i5.16523.

Steven J. Gibbons and Frode Ringdal. The detection of low magnitude seismic events using array-based waveform correlation. *Geophysical Journal International*, 165(1):149–166, 2006. doi: 10.1111/j.1365-246X.2006.02865.x.

Jonathan Goh, Sridhar Adepu, Khurum Nazir Junejo, and Aditya Mathur. A dataset to support research in the design of secure water treatment systems. In *Critical Information Infrastructures Security*, pp. 88–99. Springer International Publishing, 2017. doi: 10.1007/978-3-319-71368-7_8.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 387–395, 2018.

Jahoon Jeong and Hyunsoo Yoon. Pgrf-net: A prototype-guided relational fusion network for diagnostic multivariate time-series anomaly detection. In *International Conference on Learning Representations*, 2026.

Beibu Li, Qichao Shentu, Yang Shu, Hui Zhang, Ming Li, Ning Jin, Bin Yang, and Chenjuan Guo. Crossad: Time series anomaly detection with cross-scale associations and cross-window modeling. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Stephane G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989. doi: 10.1109/34.192463.

Kai Mao, Yiyang Lian, Yangyang Wang, Meiqin Liu, Nanning Zheng, and Ping Wei. Unveiling multi-view anomaly detection: Intra-view decoupling and inter-view fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12381–12389, 2025. doi: 10.1609/aaai.v39i12.33349.

S. Mostafa Mousavi, Yixiao Sheng, Weiqiang Zhu, and Gregory C. Beroza. STanford EArthquake Dataset (STEAD): A global data set of seismic signals for AI. *IEEE Access*, 7:179464–179476, 2019. doi: 10.1109/ACCESS.2019.2947848.

Alireza Niksejel and Miao Zhang. OBSTransformer: A deep-learning seismic phase picker for OBS data using automated labelling and transfer learning. *Geophysical Journal International*, 237(1): 485–505, 2024. doi: 10.1093/gji/ggae049.

Yurui Qian, Qi Cai, Yingwei Pan, Yehao Li, Ting Yao, Qibin Sun, and Tao Mei. Boosting diffusion models with moving average sampling in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8911–8920, 2024.

David Santos-Domínguez, Soledad Torres-Guijarro, Antonio Cardenal-López, and Antonio Pena-Gimenez. ShipsEar: An underwater vessel noise database. *Applied Acoustics*, 113:64–69, 2016. doi: 10.1016/j.apacoust.2016.06.008.

Martin Schimmel and Hanneke Paulssen. Noise reduction and detection of weak, coherent signals through phase-weighted stacks. *Geophysical Journal International*, 130(2):497–505, 1997. doi: 10.1111/j.1365-246X.1997.tb05664.x.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Junho Song, Keonwoo Kim, Jeonglyul Oh, and Sungzoon Cho. MEMTO: Memory-guided transformer for multivariate time series anomaly detection. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2828–2837, 2019. doi: 10.1145/3292500.3330672.

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*, volume 34, 2021.

Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. TranAD: Deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment*, 15(6): 1201–1214, 2022. doi: 10.14778/3514061.3514067.

Chengsen Wang, Zirui Zhuang, Qi Qi, Jingyu Wang, Xingyu Wang, Haifeng Sun, and Jianxin Liao. Drift doesn't matter: Dynamic decomposition with diffusion reconstruction for unstable multivariate time series anomaly detection. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0473.

M. Withers, R. Aster, C. Young, J. Beiriger, M. Harris, S. Moore, and J. Trujillo. A comparison of select trigger algorithms for automated global seismic phase and event detection. *Bulletin of the Seismological Society of America*, 88(1):95–106, 1998. doi: 10.1785/BSSA0880010095.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Xingjian Wu, Xiangfei Qiu, Zhengyu Li, Yihang Wang, Jilin Hu, Chenjuan Guo, Hui Xiong, and Bin Yang. Catch: Channel-aware multivariate time series anomaly detection via frequency patching. In *International Conference on Learning Representations*, 2025.

Chunjing Xiao, Zehua Gou, Wenxin Tai, Kunpeng Zhang, and Fan Zhou. Imputation-based time-series anomaly detection with conditional weight-incremental diffusion models. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2742–2751, 2023. doi: 10.1145/3580305.3599391.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations*, 2022.

Jinghong Xu, Shengdong Du, Haochen Shi, Jie Hu, Yan Yang, Fengmao Lv, and Tianrui Li. Pre-trained conditional encoding guided diffusion for time series anomaly detection. *Expert Systems with Applications*, 325:132680, 2026. doi: 10.1016/j.eswa.2026.132680.

Chaocheng Yang, Tingyin Wang, and Xuanhui Yan. DDMT: Denoising diffusion mask transformer models for multivariate time series anomaly detection. *arXiv preprint arXiv:2310.08800*, 2023a.

Xingyi Yang, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Diffusion probabilistic model made slim. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22552–22562, 2023b.

Yiyuan Yang, Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. DCdetector: Dual attention contrastive representation learning for time series anomaly detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3033–3045, 2023c. doi: 10.1145/3580305.3599295.

Zahra Zamanzadeh Darban, Geoffrey I. Webb, Shirui Pan, Charu C. Aggarwal, and Mahsa Salehi. Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*, 57(1):1–42, 2024. doi: 10.1145/3691338.

Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. Tfad: A decomposition time series anomaly detection architecture with time-frequency analysis. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, pp. 2497–2507, 2022. doi: 10.1145/3511808.3557470.

Guojin Zhong, Pan Wang, Jin Yuan, Zhiyong Li, and Long Chen. Multi-resolution decomposable diffusion model for non-stationary time series anomaly detection. In *International Conference on Learning Representations*, 2025.

A Additional Method Details

A.1 Construction of the Time–Frequency Energy Prior

For the c-th channel, we compute a continuous wavelet transform (CWT) using the Morlet wavelet (`morl`) to characterize localized variations across different frequency ranges. The wavelet coefficient at timestamp t and scale s is written as

$$w_{c,s}(t) = \sum_{\tau} X_{\tau,c} \psi_s(t - \tau), \tag{19}$$

where $\psi_s(\cdot)$ denotes the scaled wavelet function. The coefficient $w_{c,s}(t)$ captures the local response of channel c around timestamp t at the frequency associated with scale s, and $|w_{c,s}(t)|^2$ represents the corresponding local energy.

To aggregate frequency information at a coarser level, we organize the wavelet coefficients according to their associated frequencies and uniformly partition the overall frequency range into $B = 3$ non-overlapping bands of equal bandwidth, corresponding to low-, mid-, and high-frequency components. Let $\mathcal{B}_b$ denote the set of scales whose associated frequencies fall within band b. The local energy within each band is computed as

$$E_{c,b}(t) = \frac{1}{|\mathcal{B}_b|} \sum_{s \in \mathcal{B}_b} |w_{c,s}(t)|^2. \tag{20}$$

The band-wise energies are then aggregated as

$$E_c(t) = \sum_{b=1}^{B} E_{c,b}(t), \tag{21}$$

and the strongest response across channels is retained to construct the time–frequency energy prior:

$$p_1(t) = \max_c E_c(t). \tag{22}$$

This construction summarizes localized energy variations across low-, mid-, and high-frequency ranges while preserving their temporal locations, providing frequency-sensitive evidence for context-risk estimation.

A.2 Construction of the Cross-Channel Deviation Prior

For each timestamp t, we construct a local temporal window $\mathcal{A}(t)$ and calculate the Pearson correlation coefficient between each pair of channels:

$$R_{ij}(t) = \frac{\sum_{\tau \in \mathcal{A}(t)} \left(X_{\tau,i} - \mu_i(t)\right)\left(X_{\tau,j} - \mu_j(t)\right)}{\sqrt{\sum_{\tau \in \mathcal{A}(t)} \left(X_{\tau,i} - \mu_i(t)\right)^2} \sqrt{\sum_{\tau \in \mathcal{A}(t)} \left(X_{\tau,j} - \mu_j(t)\right)^2}}, \tag{23}$$

where

$$\mu_i(t) = \frac{1}{|\mathcal{A}(t)|} \sum_{\tau \in \mathcal{A}(t)} X_{\tau,i} \tag{24}$$

denotes the local mean of channel i. The resulting matrix $\mathbf{R}(t) \in \mathbb{R}^{C \times C}$ characterizes local inter-channel relationships. Since anomaly detection is performed at the timestamp level, we aggregate the absolute off-diagonal correlations into a shared cross-channel consistency measure and convert it into the deviation prior $p_3(t)$ according to Eq. 3. This allows channel-specific deviations to contribute to the reliability estimate of the corresponding timestamp, with larger deviations producing higher prior scores. For single-channel inputs, p_3 is omitted. As energy-based priors may emphasize high-amplitude events, the cross-channel prior provides complementary relational evidence, and all priors are normalized before fusion to reduce view-specific bias.

A.3 Analysis of Risk-Adaptive Context Corruption

We provide further analysis of the risk-adaptive corruption. For an observation with risk $r_i \in [0, 1]$, the effective cumulative diffusion coefficient is

$$\bar{\alpha}_{k,i}^{r} = \bar{\alpha}_k^{1+\rho r_i}. \tag{25}$$

Since $0 < \bar{\alpha}_k < 1$, increasing r_i reduces $\bar{\alpha}_{k,i}^{r}$. According to the corruption process in Eq. 5, this decreases the retained contribution of the original observation while increasing the injected noise. Therefore, high-risk contexts provide weaker conditional information to the denoiser, whereas low-risk contexts are preserved to a greater extent. This is consistent with the objective of suppressing potentially unreliable contexts without completely discarding the observed context.

The formulation also has intuitive boundary cases. When $r_i = 0$, $\bar{\alpha}_{k,i}^{r} = \bar{\alpha}_k$, corresponding to the original diffusion schedule. When $r_i = 1$, $\bar{\alpha}_{k,i}^{r} = \bar{\alpha}_k^{1+\rho}$, giving the strongest corruption under the specified ρ. Hence, ρ controls the overall strength of risk adaptation while preserving the original diffusion formulation.

A.4 Progressive Frequency Assists

In Sec. 3.4, the frequency-domain clean estimate is projected into the temporal space and used to refine the temporal prediction:

$$\overline{\mathbf{X}}_0^{(k)} = (1 - \lambda_k)\widehat{\mathbf{X}}_0^{(k)} + \lambda_k \mathbf{H}_k. \tag{26}$$

The reverse diffusion process proceeds from $k = K$ to $k = 0$. Previous frequency analyses of diffusion models have shown that denoising generally evolves from recovering low-frequency structures to progressively restoring higher-frequency details (Yang et al., 2023b; Qian et al., 2024). Accordingly, frequency-domain estimates at early reverse steps are relatively uncertain, whereas their recovered spectral structures become more reliable as denoising proceeds. We therefore assign weaker frequency assist at early steps and gradually increase its contribution toward the clean state.

Specifically, we adopt the linear schedule

$$\lambda_k = \lambda_{\max}\left(1 - \frac{k}{K}\right), \tag{27}$$

where $\lambda_{\max} \in [0, 1]$ controls the maximum contribution of frequency-domain guidance. Thus, frequency assist starts from zero at $k = K$ and progressively increases to $\lambda_{\max}$ as the reverse process approaches the clean state.

A.5 Network Architecture

The architecture of the denoising network is illustrated in Fig. 3. At diffusion step k, the denoiser takes the noisy target sequence $\mathbf{X}_k^m$, the corrupted observed condition $\widetilde{\mathbf{X}}_k^o$, and the diffusion step k as inputs, and predicts the diffusion noise ϵ_θ. The backbone consists of a stack of Condition Residual Networks, which progressively refine the latent representation while incorporating the observed context.

Within each Condition Residual Network, $\mathbf{X}_k^m$ is first processed by a temporal Transformer to capture dependencies along the temporal dimension, followed by a spatial Transformer to model inter-variable correlations. The diffusion step k is mapped to a step embedding and injected into the intermediate representation to condition the denoising process on the current noise level. In parallel, the corrupted context $\widetilde{\mathbf{X}}_k^o$ is embedded as conditional information and fused with the latent representation through cross-attention, allowing the denoiser to exploit reliable contexts without directly mixing them with the noisy target representation. The resulting representation is passed through the stacked residual blocks and finally mapped by a linear projection to obtain the predicted noise ϵ_θ. The temporal and frequency denoisers instantiate the same architecture independently, using separate parameter sets θ_T and θ_F without parameter sharing.

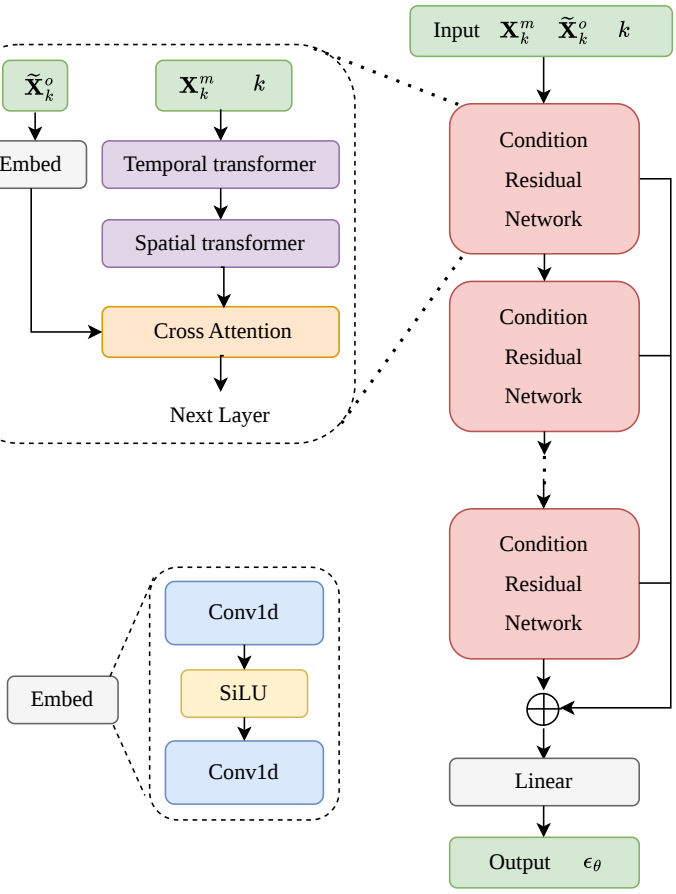


Figure 3: Architecture of the denoising network.

B Additional Experimental Details

B.1 Datasets

We use the following five widely adopted datasets for multivariate time-series anomaly detection experiments.

- **SMD** (Server Machine Dataset) (Su et al., 2019) contains monitoring data collected from 28 server machines, with 38 variables recorded for each machine.
- **PSM** (Pooled Server Metrics) (Abdulaal et al., 2021) consists of server monitoring metrics collected from multiple application server nodes at eBay.
- **MSL** (Mars Science Laboratory) and **SMAP** (Soil Moisture Active Passive) (Hundman et al., 2018) are spacecraft telemetry datasets collected by NASA, containing 55 and 25 variables, respectively.
- **SWaT** (Secure Water Treatment) (Goh et al., 2017) contains measurements collected from a water treatment testbed with 51 variables under normal operations and simulated attacks.

The dataset statistics are summarized in Table 4.

To further assess the generalization ability of SCTF, we use four real-world signal datasets covering seismic and underwater acoustic monitoring scenarios.

Table 4: Statistics of the datasets used for anomaly detection.

Dataset	Entities	Dimensions	Train #	Val. #	Test #	Anomaly Rate (%)
SMD	28	38	637564	70841	708420	4.16
PSM	1	25	119232	13249	87841	27.76
MSL	27	55	52485	5832	73729	10.48
SMAP	55	25	126742	14083	444035	12.83
SWaT	1	51	445500	49500	449919	12.14

Table 5: Statistics of the datasets used in cross-dataset generalization.

Dataset	Dim.	Train Ent.	Val. Ent.	Test Ent.	Train #	Val. #	Test #	Anom. (%)
STEAD	3	80	10	10	1,200,000	150,000	150,000	15.88
OBST2024	3	–	10	10	–	150,000	150,000	20.00
DAS-MedSea	3	–	10	10	–	150,000	150,000	30.15
ShipsEar	3	–	10	10	–	150,000	150,000	22.74

- **STEAD** (Mousavi et al., 2019) is a large-scale seismic dataset containing three-component earthquake and noise waveforms collected from seismic stations worldwide. The anomaly intervals are defined as the time periods containing earthquake arrivals, while background noise segments are regarded as normal observations.
- **OBST2024** (Niksejel & Zhang, 2024) contains three-component seismic waveforms recorded by ocean-bottom seismometers (OBSs), including earthquake and noise signals. Similar to STEAD, earthquake arrival periods are treated as anomaly intervals and non-event periods are considered normal.
- **DAS-MedSea** (Bouffaut et al., 2025) contains distributed acoustic sensing recordings collected in the Mediterranean Sea, including annotated fin-whale acoustic events. The occurrence periods of fin-whale calls are regarded as anomaly intervals, while continuous background recordings are treated as normal observations.
- **ShipsEar** (Santos-Domínguez et al., 2016) is an underwater acoustic dataset containing recordings of different vessel types and background noise. The periods containing vessel acoustic signals are considered anomaly intervals, with background noise segments treated as normal.

For underwater acoustic datasets, we select three spatially adjacent channels from the original multi-channel recordings to construct the multivariate input representation.

For the generalization evaluation, SCTF is trained on the training split of STEAD, with hyperparameters selected using the labeled STEAD validation split and subsequently fixed for all datasets. The trained model is then evaluated on the test splits of STEAD, OBST2024, DAS-MedSea, and ShipsEar without retraining or target-domain hyperparameter adjustment. As summarized in Table 5, all evaluation inputs are represented using three channels, with 80 STEAD entities used for training and 10 entities from each dataset reserved for validation and testing, respectively. For each dataset, the detection threshold is calibrated from the distribution of validation anomaly scores without using validation labels, while the test split is used only for final evaluation. The same window length and model configuration are maintained across all datasets. This protocol evaluates cross-dataset generalization under fixed model and hyperparameter settings across different signal distributions and sensing scenarios.

B.2 Evaluation Metrics

We evaluate anomaly detection performance using Precision (P), Recall (R), F1-score (F1), and the area under the precision–recall curve (AUPRC). Precision, Recall, and F1 are used as the primary point-wise detection metrics, while AUPRC provides complementary threshold-independent evaluation.

Precision, Recall, and F1-score. Let TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively. Precision measures the proportion of predicted anomalies that correspond to actual anomalies, while Recall measures the proportion of actual anomalies

successfully detected:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \qquad \mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{28}$$

The F1-score is defined as the harmonic mean of Precision and Recall:

$$\mathrm{F1} = \frac{2\,\mathrm{P}\,\mathrm{R}}{\mathrm{P} + \mathrm{R}}. \tag{29}$$

Precision reflects resistance to false alarms, whereas Recall measures the ability to identify anomalous observations. F1 provides a balanced summary of these two aspects and is therefore used as our primary evaluation metric.

AUPRC. The precision–recall (PR) curve characterizes the trade-off between Precision and Recall across different anomaly-score thresholds. AUPRC is the area under the PR curve and provides a threshold-independent measure of anomaly detection performance across different operating points.

Higher values indicate better detection performance for all reported metrics. Unless otherwise specified, F1 is used as the principal metric in the main comparison, while AUPRC is reported as a complementary threshold-independent measure of anomaly separability.

Evaluation Protocol. For the standard benchmarks, we follow the point-adjustment settings reported by the corresponding sources. For the generalization evaluation, all baselines are trained on STEAD and directly evaluated on OBST2024, DAS-MedSea, and ShipsEar without additional training or fine-tuning. All methods are evaluated under the same no-point-adjustment protocol using a unified input configuration and window length.

B.3 Additional Hyperparameter Analysis

All sensitivity results in this subsection are reported on the validation split, whereas the main detection results in Tables 2 and 3 are reported on the test sets. We provide additional sensitivity analyses of the main hyperparameters of SCTF on STEAD, including the number of frequency bands B, the maximum frequency-guidance weight $\lambda_{\max}$, the feedback weights (γ_1, γ_2), the number of Condition Residual Networks N, and the number of DDIM sampling steps S. As described in Sec. 4.2, the default hyperparameters are selected using the labeled validation split according to the validation F1-score. The following analyses examine the robustness and computational trade-offs around the selected configuration rather than using the test set for hyperparameter selection.

Frequency Decomposition and Guidance. We first analyze the number of frequency bands B and the maximum frequency-guidance weight $\lambda_{\max}$. As shown in Table 6, using multiple frequency bands improves performance, while increasing B beyond the default value yields only limited changes. The selected setting $B = 3$ therefore lies within a stable performance region. Table 7 further shows that removing frequency assist or using an excessive guidance weight degrades performance, whereas moderate guidance provides a better balance between the temporal and frequency branches. The selected setting $\lambda_{\max} = 0.5$ also exhibits stable performance.

Table 6: Sensitivity to the number of frequency bands on STEAD.

B	2	**3**	4
F1	96.10	**98.39**	98.14
AUPRC	95.83	**99.50**	99.22

Table 7: Sensitivity to the frequency-guidance weight on STEAD.

$\lambda_{\max}$	0.00	0.25	**0.50**	0.75	1.00
F1	93.87	95.63	**98.39**	96.61	94.97
AUPRC	95.42	94.88	**99.50**	96.58	95.56

Risk-Refinement Weights. We further analyze the feedback weights (γ_1, γ_2), where γ_1 and γ_2 control the temporal and frequency reconstruction feedback, respectively, while $1 - \gamma_1 - \gamma_2$ determines the retained proportion of the previous risk estimate. As shown in Table 8, the selected setting $(\gamma_1, \gamma_2) = (0.5, 0.2)$ achieves strong performance. The results indicate that assigning more weight to temporal feedback while retaining part of the previous risk estimate provides a favorable balance between refinement strength and stability.

Network Depth and Sampling Steps. Table 9 reports the effects of network depth and DDIM sampling steps. Increasing N from 2 to 4 substantially improves F1 and AUPRC, while further

Table 8: Sensitivity analysis of the risk-refinement weights on STEAD.

(γ_1, γ_2)	$(\mathbf{0.5}, \mathbf{0.2})$	$(0.2, 0.5)$	$(0.2, 0.2)$	$(0.4, 0.4)$	$(0.2, 0.1)$
F1	**98.39**	93.02	95.69	94.94	97.46
AUPRC	**99.50**	94.16	96.11	95.28	97.72

increasing the depth yields only marginal changes at higher inference cost. The selected setting $N = 4$ therefore provides a favorable accuracy–efficiency trade-off.

For DDIM sampling, increasing S from 5 to 10 improves detection performance, whereas further increasing it to 20 or 50 yields only minor changes in F1 and AUPRC while substantially increasing inference time. The selected setting $S = 10$ therefore provides a favorable balance between detection accuracy and inference efficiency.

Table 9: Sensitivity analysis of network depth and DDIM sampling steps on STEAD.

Parameter	**Value**	**F1**	**AUPRC**	**Inference Time (s)**
N	2	91.86	90.24	0.77
	3	96.48	95.75	1.14
	4	98.39	99.50	1.45
	5	98.23	99.41	2.02
DDIM sampling steps S	5	97.28	98.12	0.69
	10	98.39	99.50	1.45
	20	98.25	99.41	3.07
	50	98.28	99.43	9.12

B.4 Implementation Details

The detailed hyperparameter settings of SCTF are summarized in Table 10. Unless otherwise specified, the same configuration is used across all datasets. The temporal and frequency branches share the same diffusion schedule and are synchronized during both forward diffusion and reverse reconstruction. Context risks are periodically refined using reconstruction feedback during training. The frequency-guidance weight λ_k is progressively increased along the reverse trajectory, allowing frequency information to contribute more strongly at later denoising stages.

Threshold Selection. Hyperparameters are selected using validation labels, whereas the subsequent anomaly detection threshold is calibrated without label information. Specifically, the threshold is determined from the empirical distribution of validation anomaly scores (Xiao et al., 2023). In this work, the anomaly score is computed from the temporal reconstruction error, while the frequency reconstruction error is used as auxiliary information during diffusion reconstruction rather than directly contributing to anomaly decisions. For each dataset d, we obtain the validation anomaly scores $\mathcal{S}_d^{\mathrm{val}} = \{s_i\}_{i=1}^{N_d}$ and compute the dataset-specific threshold as

$$\tau_d = \mu_d + \sigma_d, \tag{30}$$

where

$$\mu_d = \frac{1}{N_d}\sum_{i=1}^{N_d} s_i, \qquad \sigma_d = \sqrt{\frac{1}{N_d}\sum_{i=1}^{N_d}(s_i - \mu_d)^2}. \tag{31}$$

The resulting threshold τ_d is fixed during testing on the corresponding dataset. This threshold calibration uses only the validation anomaly-score distribution and does not use validation labels or assume a predefined anomaly ratio.

B.5 Baselines

We compare SCTF with representative state-of-the-art time-series anomaly detection methods covering reconstruction-, representation-, and diffusion-based approaches:

- **Anomaly Transformer** (Xu et al., 2022) introduces an association discrepancy between learned series associations and a prior association to distinguish anomalous observations from normal temporal patterns.

Table 10: Detailed hyperparameter settings of SCTF.

Hyperparameter	Value
Learning rate	3×10^{-4}
Weight decay	1×10^{-6}
Batch size	16
Training epochs	50
Local prior window length $\lvert\mathcal{A}(t)\rvert$	256
Sliding-window length	256
Window stride	128
Temporal loss weight ω_T	1
Frequency loss weight ω_F	0.5
Coupling loss weight ω_C	0.05
Diffusion steps K	50
DDIM sampling steps S	10
Noise schedule	quadratic, $\beta_1 = 10^{-4}$, $\beta_K = 0.5$
Number of frequency bands B	3
Risk-adaptation coefficient ρ	0.3
Temporal feedback weight γ_1	0.5
Frequency feedback weight γ_2	0.2
Maximum frequency-guidance weight $\lambda_{\max}$	0.5
Mask ratio	0.25
Risk refinement interval	10 epochs
Layers of residual network	4

- **TimesNet** (Wu et al., 2023) transforms one-dimensional time series into two-dimensional representations according to multiple temporal periods and models intra- and inter-period variations using TimesBlocks.
- **DCdetector** (Yang et al., 2023c) employs multi-scale dual attention and contrastive learning to learn anomaly-discriminative representations without explicitly reconstructing the input sequence.
- **MEMTO** (Song et al., 2023) introduces a memory-guided reconstruction framework that learns representative normal patterns and constrains the reconstruction of anomalous observations.
- **TranAD** (Tuli et al., 2022) employs a Transformer-based reconstruction architecture with self-conditioning and adversarial training for multivariate time-series anomaly detection.
- **TFAD** (Zhang et al., 2022) jointly models temporal and frequency information and detects anomalies by learning representations from complementary time–frequency views.
- **ImDiffusion** (Chen et al., 2023) formulates time-series anomaly detection as diffusion-based imputation and identifies anomalies according to the error between observations and their reconstructed values.
- **DiffAD** (Xiao et al., 2023) adopts a mask-and-imputation diffusion framework and progressively controls the contribution of selected observations during the denoising process.
- **DDMT** (Yang et al., 2023a) combines adaptive dynamic masking with a denoising diffusion Transformer to reduce information leakage from potentially anomalous observations during reconstruction.
- **D3R** (Wang et al., 2023) decomposes non-stationary time series into stable and trend components and employs diffusion-based reconstruction to improve anomaly detection under distribution drift.
- **MODEM** (Zhong et al., 2025) develops a multi-resolution coarse-to-fine diffusion process and a frequency-enhanced decomposable network for anomaly detection in non-stationary time series.
- **CATCH** (Wu et al., 2025) performs frequency-domain patching to capture fine-grained spectral patterns and introduces a channel fusion module with masked attention to model adaptive inter-variable correlations for anomaly detection.

Table 11: Stability analysis of SCTF over three independent runs under the PA evaluation protocol. Results are reported as mean ± standard deviation (%) for each benchmark.

Metric	SMAP	SMD	PSM	MSL	SWaT	Average
Precision	95.14 ± 0.41	91.77 ± 0.48	98.23 ± 0.31	96.31 ± 0.35	97.42 ± 0.30	–
Recall	99.02 ± 0.28	96.50 ± 0.33	98.71 ± 0.26	94.19 ± 0.46	96.98 ± 0.32	–
F1-score	97.04 ± 0.37	94.08 ± 0.44	98.47 ± 0.29	95.24 ± 0.39	97.20 ± 0.34	96.41

- **CrossAD** (Li et al., 2025) models cross-scale associations by reconstructing fine-grained series from coarser representations and incorporates cross-window global context to capture anomaly patterns beyond fixed local windows.

For the benchmark comparison in Table 1, baseline results are collected from published studies rather than generated by our own runs. We use the results reported in the original papers when available and adopt better published results from later studies under the corresponding benchmark evaluation setting when applicable. CATCH (Wu et al., 2025) and CrossAD (Li et al., 2025) are excluded from this comparison because their reported results follow evaluation protocols different from the point-adjustment setting used in Table 1. For the cross-dataset generalization experiments, methods whose public implementations could not be reliably reproduced are excluded from the comparison.

B.6 Stability Analysis of SCTF

To further examine the influence of random initialization, we report the mean and standard deviation of SCTF over three independent runs on the five benchmark datasets. The detailed results are presented in Table 11.

B.7 Generalization Results

Table 12 reports the Precision, Recall, F1-score, and AUPRC results on STEAD, ShipsEar, OBST, and DAS-MedSea. All methods are evaluated under a unified no-point-adjustment protocol, where predicted anomalous points are directly compared with the ground-truth annotations. AUPRC is additionally reported using continuous anomaly scores to assess threshold-independent anomaly discrimination.

Across the four datasets, SCTF shows relatively stable F1 performance under different signal characteristics and anomaly patterns. It obtains the highest F1-score on three datasets, while ImDiffusion performs best on OBST. A similar pattern is observed for AUPRC, where SCTF remains competitive across all four datasets. These results suggest that the proposed reconstruction framework can generalize across heterogeneous sensing domains without relying on point adjustment.

The qualitative results in Figs. 4 and 5 further illustrate the point-wise detection behavior. On the selected segments, the compared methods exhibit different degrees of fragmented detections and false positives around anomalous regions. SCTF produces more concentrated detections around the annotated intervals, while ImDiffusion, DiffAD, Anomaly Transformer, and CATCH show different localization patterns depending on the signal. These examples provide a complementary view of the generalization results beyond the quantitative metrics.

B.8 Detailed Analysis of Risk-Adaptive Context Corruption

Table 13 compares different strategies for handling potentially unreliable conditioning contexts. *No Isolation* directly uses all contexts without condition filtering, while *Oracle Corruption* replaces the estimated risk with ground-truth anomaly information and serves as an oracle reference. *Random Isolation* removes 30% of contexts uniformly at random, whereas *Hard Isolation* removes contexts with the highest estimated risks. We consider $q \in \{10\%, 30\%, 50\%\}$ to investigate the effect of different isolation ratios. *Fixed Corruption* assigns the same risk value $r_i = \bar{r}$ to all contexts, preserving the overall corruption strength while removing context-specific adaptation. For adaptive corruption, we further evaluate different values of $\rho \in \{0.1, 0.3, 0.5\}$, which controls the influence of context risk on the corruption strength. Smaller values weaken the effect of risk-aware corruption, while larger values increase the difference between high- and low-risk contexts. We use $\rho = 0.3$ as

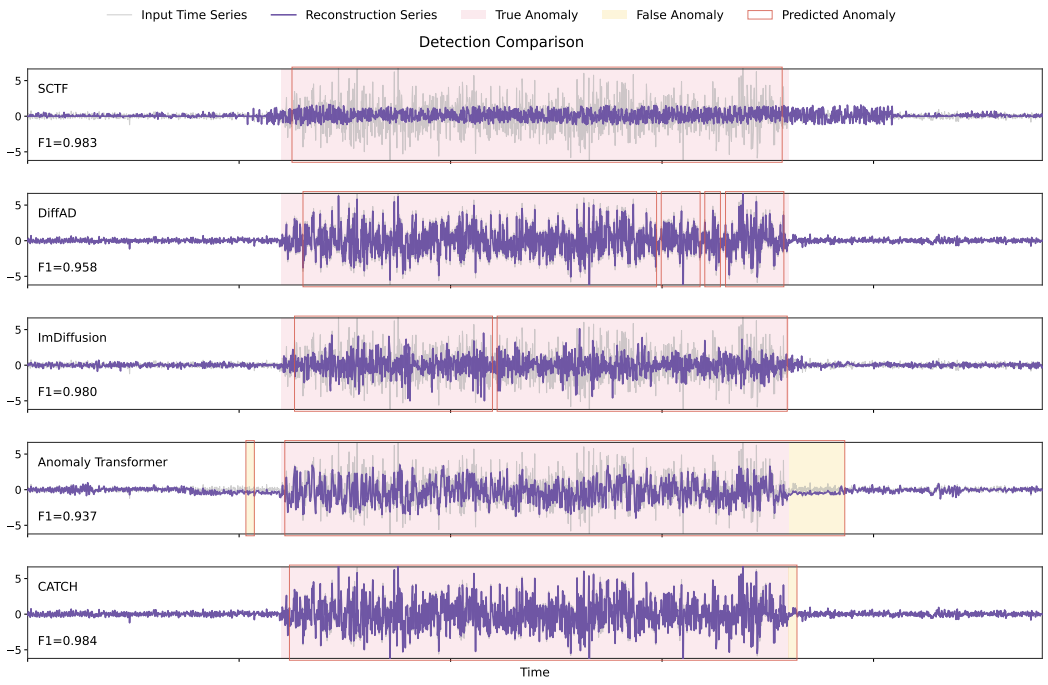


Figure 4: Comparison of anomaly detection on a representative ShipsEar segment under the no-point-adjustment setting. Gray and purple curves denote the input and reconstructed series, respectively. Pink regions indicate ground-truth anomalies, yellow regions indicate false-positive areas, and red boxes denote predicted anomalous intervals.

Table 12: Detailed Precision, Recall, F1-score, and AUPRC (%) on four generalization datasets under the no-point-adjustment evaluation protocol. Results are reported as mean ± standard deviation over three independent runs. Avg. F1 denotes the average of the mean F1-scores across the four datasets.

(a) STEAD and ShipsEar

Method	STEAD				ShipsEar			
	P	R	F1	AUPRC	P	R	F1	AUPRC
SCTF	97.92 ± 0.52	98.42 ± 0.41	**98.17** ± 0.38	**99.37** ± 0.31	99.05 ± 0.48	96.58 ± 0.72	**97.80** ± 0.55	99.73 ± 0.24
DiffAD	77.99 ± 0.84	88.43 ± 0.91	82.88 ± 0.88	93.04 ± 0.66	98.07 ± 0.72	91.92 ± 0.89	94.90 ± 0.76	**99.93** ± 0.05
ImDiffusion	94.84 ± 0.67	87.41 ± 0.82	90.97 ± 0.74	96.02 ± 0.58	97.58 ± 0.63	96.17 ± 0.79	96.87 ± 0.61	99.91 ± 0.08
Anomaly Transformer	84.93 ± 0.88	90.32 ± 0.69	87.54 ± 0.80	91.01 ± 0.64	88.69 ± 0.71	99.33 ± 0.62	93.71 ± 0.68	98.56 ± 0.55
DCdetector	73.04 ± 0.93	66.85 ± 0.90	69.81 ± 0.92	71.97 ± 0.75	67.93 ± 0.86	68.12 ± 0.81	68.02 ± 0.84	76.15 ± 0.70
MEMTO	73.90 ± 0.77	71.49 ± 0.90	72.68 ± 0.85	22.15 ± 0.62	78.55 ± 0.73	98.25 ± 0.65	87.30 ± 0.67	79.02 ± 0.59
TimesNet	86.38 ± 0.82	63.82 ± 0.94	73.41 ± 0.91	84.59 ± 0.68	86.28 ± 0.79	56.08 ± 0.93	67.98 ± 0.88	84.30 ± 0.72
CATCH	90.03 ± 0.66	98.21 ± 0.61	93.94 ± 0.64	96.69 ± 0.56	96.13 ± 0.62	97.58 ± 0.60	96.85 ± 0.61	93.98 ± 0.54
D3R	88.21 ± 0.87	75.24 ± 0.94	81.21 ± 0.90	76.71 ± 0.73	76.28 ± 0.85	68.33 ± 0.91	72.09 ± 0.88	78.11 ± 0.69
TFAD	82.30 ± 0.81	76.35 ± 0.89	79.21 ± 0.86	81.28 ± 0.67	76.74 ± 0.78	72.00 ± 0.84	74.29 ± 0.82	73.24 ± 0.64
DDMT	83.03 ± 0.76	63.90 ± 0.95	72.22 ± 0.89	77.27 ± 0.63	86.68 ± 0.70	68.33 ± 0.91	76.42 ± 0.80	78.60 ± 0.58
CrossAD	86.91 ± 0.72	97.24 ± 0.64	91.79 ± 0.69	97.02 ± 0.57	76.48 ± 0.83	90.03 ± 0.74	82.70 ± 0.79	83.30 ± 0.65

(b) OBST, DAS-MedSea, and average F1

Method	OBST				DAS-MedSea				Avg. F1
	P	R	F1	AUPRC	P	R	F1	AUPRC	
SCTF	99.32 ± 0.46	87.06 ± 0.68	92.79 ± 0.57	**96.29** ± 0.35	89.42 ± 0.50	87.78 ± 0.74	**88.59** ± 0.62	**94.42** ± 0.28	**94.34**
DiffAD	79.61 ± 0.92	83.18 ± 0.78	81.36 ± 0.86	89.65 ± 0.71	73.64 ± 0.90	75.93 ± 0.69	74.77 ± 0.82	76.62 ± 0.57	83.48
ImDiffusion	95.31 ± 0.70	93.65 ± 0.64	**94.47** ± 0.67	96.06 ± 0.55	88.65 ± 0.83	77.99 ± 0.91	82.98 ± 0.79	74.59 ± 0.62	91.32
Anomaly Transformer	89.09 ± 0.79	87.40 ± 0.73	88.24 ± 0.76	88.25 ± 0.68	75.00 ± 0.94	63.50 ± 0.89	68.77 ± 0.92	76.18 ± 0.60	84.56
DCdetector	59.74 ± 0.90	52.43 ± 0.95	55.85 ± 0.93	65.41 ± 0.78	56.79 ± 0.87	60.92 ± 0.81	58.78 ± 0.85	69.87 ± 0.72	63.12
MEMTO	68.18 ± 0.94	85.22 ± 0.80	75.75 ± 0.88	87.35 ± 0.70	63.43 ± 0.92	66.08 ± 0.74	64.73 ± 0.84	53.53 ± 0.61	75.12
TimesNet	78.65 ± 0.75	53.07 ± 0.95	63.38 ± 0.89	75.66 ± 0.73	87.04 ± 0.66	59.09 ± 0.93	70.39 ± 0.84	69.10 ± 0.58	68.79
CATCH	87.29 ± 0.78	91.70 ± 0.69	89.44 ± 0.72	90.85 ± 0.63	76.91 ± 0.86	81.93 ± 0.77	79.34 ± 0.81	81.29 ± 0.59	89.89
D3R	64.09 ± 0.95	67.39 ± 0.82	65.70 ± 0.90	76.15 ± 0.65	80.04 ± 0.74	59.64 ± 0.95	68.35 ± 0.87	68.90 ± 0.58	71.84
TFAD	73.72 ± 0.89	56.03 ± 0.95	63.67 ± 0.92	72.90 ± 0.76	73.51 ± 0.84	63.55 ± 0.90	68.17 ± 0.87	81.70 ± 0.71	71.34
DDMT	77.21 ± 0.83	65.82 ± 0.95	71.06 ± 0.90	76.54 ± 0.74	83.28 ± 0.68	39.81 ± 0.91	53.87 ± 0.86	69.68 ± 0.60	68.39
CrossAD	77.02 ± 0.88	71.09 ± 0.84	73.94 ± 0.86	72.26 ± 0.69	91.32 ± 0.62	58.63 ± 0.94	71.41 ± 0.79	93.38 ± 0.55	79.96

the default setting. The results show that risk-guided isolation performs better than random isolation, indicating that identifying unreliable conditions is important for effective corruption. Meanwhile, fixed corruption improves over directly using all observations, demonstrating the benefit of context

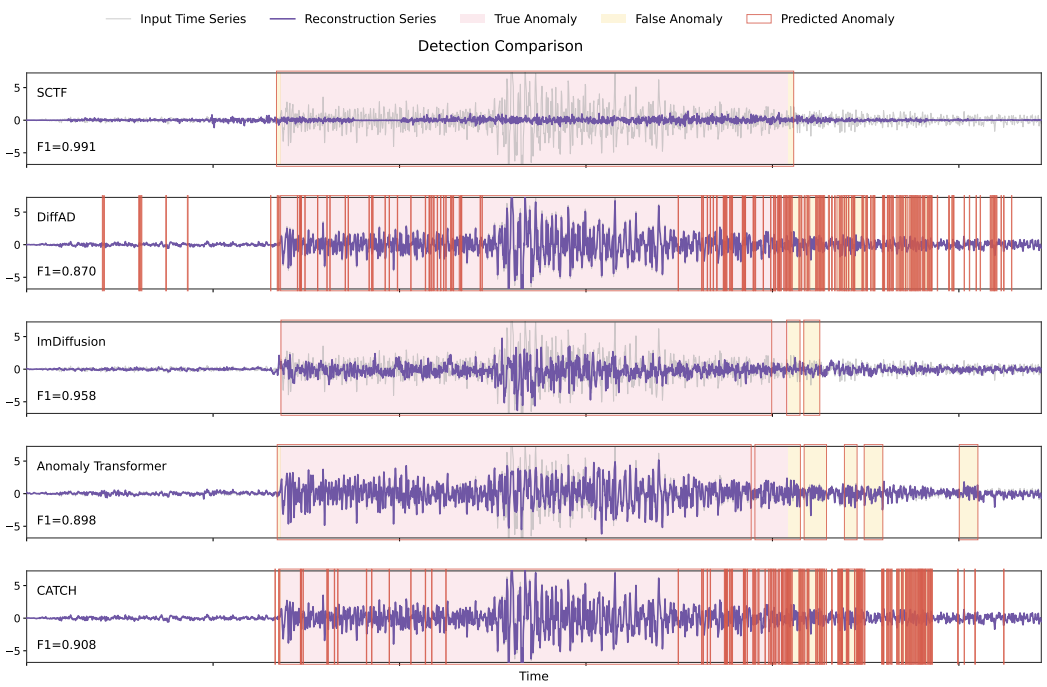


Figure 5: Comparison of anomaly detection on a representative STEAD segment under the no-point-adjustment setting.

Table 13: Detailed ablation of risk-adaptive context corruption on STEAD and PSM.

	STEAD		PSM	
Variant	**F1**	**AUPRC**	**F1**	**AUPRC**
No Isolation	82.93 ± 1.87	83.11 ± 1.52	86.68 ± 1.73	88.59 ± 1.61
Oracle Corruption (GT Risk)	98.54 ± 0.61	99.45 ± 0.38	98.65 ± 0.58	96.42 ± 0.46
Random Isolation ($q = 30\%$)	86.41 ± 0.79	91.75 ± 0.57	87.06 ± 0.82	89.74 ± 0.63
Hard Isolation ($q = 10\%$)	91.57 ± 0.71	92.93 ± 0.49	90.43 ± 0.76	90.34 ± 0.55
Hard Isolation ($q = 30\%$)	95.23 ± 0.64	95.03 ± 0.43	94.15 ± 0.69	95.68 ± 0.48
Hard Isolation ($q = 50\%$)	95.89 ± 0.62	93.48 ± 0.45	95.82 ± 0.59	94.22 ± 0.51
Fixed Corruption ($r_i = \bar{r}$)	97.52 ± 0.58	97.88 ± 0.39	96.10 ± 0.66	95.91 ± 0.47
Adaptive Corruption ($\rho = 0.1$)	90.75 ± 0.74	91.62 ± 0.54	92.83 ± 0.70	91.61 ± 0.56
Adaptive Corruption ($\rho = 0.3$)	**98.17** ± 0.38	**99.37** ± 0.31	**98.47** ± 0.29	**96.02** ± 0.44
Adaptive Corruption ($\rho = 0.5$)	93.83 ± 0.68	94.09 ± 0.50	94.71 ± 0.65	93.24 ± 0.53

corruption during reconstruction. Adaptive corruption further improves the results by assigning observation-dependent corruption strengths, with $\rho = 0.3$ achieving the best overall performance among the evaluated settings. The oracle variant provides an upper reference for evaluating the influence of risk estimation accuracy.

B.9 Detailed Analysis of Self-Correcting Risk Refinement

Table 14 further analyzes the self-correcting risk refinement process. In addition to final detection performance, we evaluate whether the estimated context risks are consistent with the ground-truth anomaly distribution using Precision, Recall, F1, and AUPRC. The labels are used only for evaluation and are not involved in risk estimation or refinement. The first setting keeps the initial context risks fixed throughout training. The remaining configurations vary the total number of training epochs and the refinement interval. For a setting with E epochs and interval I, the risks are updated every I epochs before the final epoch. The results show that self-correcting refinement improves both risk–label alignment and final detection performance on the two datasets. Increasing the training epochs provides additional improvements, while further extension leads to limited

Table 14: Detailed analysis of self-correcting risk refinement under different training epochs and refinement intervals. Our default setting uses 50 epochs with a refinement interval of 10.

Dataset	Epochs	Interval	#Ref.	Risk–GT Alignment				Detection	
				P	R	F1	AUPRC	F1	AUPRC
STEAD	50	–	0	98.87	79.52	88.14	88.08	83.70 ± 2.03	82.51 ± 1.31
	30	10	2	94.36	88.58	91.38	92.34	91.42 ± 1.62	92.95 ± 1.04
	50	10	4	99.26	97.14	98.19	99.65	98.17 ± 0.38	99.37 ± 0.31
	70	10	6	99.28	97.43	98.35	99.68	98.72 ± 0.43	99.44 ± 0.37
	50	5	9	99.18	97.10	98.13	99.51	98.38 ± 0.41	98.97 ± 0.35
	50	20	2	98.51	97.09	97.79	97.81	97.86 ± 0.47	98.54 ± 0.33
PSM	50	–	0	82.87	80.96	81.90	84.27	87.77 ± 1.96	86.69 ± 1.27
	30	10	2	91.16	92.33	91.74	93.76	92.19 ± 1.58	91.85 ± 1.06
	50	10	4	98.47	97.26	97.86	98.52	98.47 ± 0.29	96.02 ± 0.44
	70	10	6	98.52	97.27	97.89	99.02	98.49 ± 0.39	96.05 ± 0.36
	50	5	9	98.55	97.30	97.92	99.09	98.62 ± 0.45	96.11 ± 0.32
	50	20	2	97.31	96.73	97.02	98.71	98.20 ± 0.48	95.64 ± 0.41

Table 15: Detailed ablation of TF diffusion designs on STEAD and PSM.

	STEAD		PSM	
Variant	F1	AUPRC	F1	AUPRC
Temporal Only	77.10 ± 1.83	80.77 ± 1.24	83.05 ± 1.67	84.75 ± 1.09
TF Feature Conditioning	80.37 ± 1.42	82.49 ± 0.63	84.96 ± 1.28	82.57 ± 0.58
Parallel TF Diffusion	93.87 ± 0.71	95.42 ± 0.44	90.53 ± 0.83	91.32 ± 0.51
Synchronized TF Diffusion (Ours)	**98.17** ± 0.38	**99.37** ± 0.31	**98.47** ± 0.29	**96.02** ± 0.44
Frequency Only	70.70 ± 2.41	76.23 ± 1.53	58.35 ± 2.67	64.41 ± 1.42
Temporal Feature Conditioning	79.86 ± 1.55	80.92 ± 0.66	81.77 ± 1.34	80.94 ± 0.61
Synchronized TF Diffusion (Swapped)	91.80 ± 0.76	92.34 ± 0.49	89.71 ± 0.88	90.13 ± 0.55

changes. A similar trend is observed for the refinement interval: overly sparse updates may reduce performance, whereas more frequent refinement brings limited additional gains. We therefore use 50 training epochs with a 10-epoch refinement interval as the default setting, providing a balance between performance and training cost.

B.10 Detailed Analysis of TF Diffusion

Table 15 compares several strategies for incorporating temporal and frequency information. *Temporal Only* and *Frequency Only* evaluate the contribution of each domain independently. *TF Feature Conditioning* and *Temporal Feature Conditioning* introduce cross-domain information as auxiliary features without constructing a separate diffusion process. *Parallel TF Diffusion* models the two domains independently, whereas *Synchronized TF Diffusion* uses frequency information to guide temporal reconstruction during the reverse generation process. We further evaluate *Synchronized TF Diffusion (Swapped)*, where temporal information is used to guide the frequency branch instead.

The results show that modeling temporal information alone or frequency information alone is insufficient for capturing complex anomaly patterns. Feature-level conditioning provides moderate improvement, while explicitly modeling both domains through diffusion achieves more consistent gains. Compared with parallel diffusion, synchronized diffusion further improves the performance, demonstrating the benefit of cross-domain guidance during the reconstruction process.

The swapped variant performs below the temporal-centered design on both datasets, indicating that using the temporal branch as the primary reconstruction process with frequency information as complementary guidance is more suitable for anomaly detection.

B.11 Detailed Analysis of Multi-View Anomaly Priors

Table 16 evaluates the contribution of the three anomaly priors and their combinations. The time–frequency energy prior p_1, local amplitude prior p_2, and cross-channel deviation prior p_3 characterize complementary aspects of abnormal patterns. We compare single-prior variants, equal-weight fusion of all priors, and full estimation variants that remove one prior at a time.

Table 16: Detailed ablation of the multi-view anomaly priors on STEAD and PSM.

	STEAD		PSM	
Variant	**F1**	**AUPRC**	**F1**	**AUPRC**
p_1 only	85.84 ± 1.87	89.02 ± 1.04	82.89 ± 2.03	81.57 ± 1.12
p_2 only	84.83 ± 1.62	81.67 ± 0.96	91.90 ± 1.51	92.57 ± 1.07
p_3 only	78.36 ± 2.08	80.83 ± 1.17	81.47 ± 1.94	80.50 ± 0.93
Equal-weight $p_1 + p_2 + p_3$	96.24 ± 0.43	97.43 ± 0.35	95.81 ± 0.46	95.45 ± 0.32
Full $-p_1$	86.21 ± 1.06	91.75 ± 0.84	93.89 ± 0.91	91.36 ± 0.76
Full $-p_2$	88.59 ± 0.97	89.49 ± 0.92	94.90 ± 0.88	93.17 ± 0.73
Full $-p_3$	89.33 ± 1.14	88.70 ± 0.79	95.98 ± 0.83	94.99 ± 0.87
Full (Ours)	**98.17** ± 0.38	**99.37** ± 0.31	**98.47** ± 0.29	**96.02** ± 0.44

Table 17: Prior-specific evaluation on representative signals with different anomaly characteristics.

	TF-Energy Dominant		Amplitude Dominant		Cross-Channel Dominant	
Variant	**F1**	**AUPRC**	**F1**	**AUPRC**	**F1**	**AUPRC**
Full (Ours)	98.33	97.29	98.61	97.23	87.59	88.04
Full $-p_1$	92.47	89.80	95.72	92.70	86.18	87.62
Full $-p_2$	95.29	92.45	92.50	89.17	85.73	86.91
Full $-p_3$	94.78	92.17	97.11	95.46	80.95	81.36

The results show that individual priors provide useful but incomplete anomaly evidence, and their effectiveness varies across datasets due to different anomaly characteristics. Combining multiple priors consistently improves the performance over single-view estimation, demonstrating the benefit of complementary information. Although the equal-weight fusion achieves strong results, the full estimator further improves performance by adaptively integrating different priors.

Removing any prior leads to performance degradation, while the impact varies across datasets. This indicates that each view contributes complementary information for context-risk estimation and that adaptive multi-view fusion is beneficial for obtaining more reliable risk estimates. Note that, the equal-weight variant retains the self-correcting refinement process and changes only the fusion of the three priors, whereas the w/o Self-Correcting variant keeps the initial context risks fixed throughout training.

B.12 Prior-Specific Validation

We further examine whether the three anomaly priors capture complementary anomaly characteristics by selecting representative signals according to their intrinsic properties. For the time–frequency energy prior p_1, we select a ShipsEar sequence containing a ship passage whose temporal amplitude variation is relatively weak but whose spectral structure exhibits more distinct changes. For the local amplitude prior p_2, we select an SMAP sequence containing impulsive anomalies with prominent local amplitude deviations. For the cross-channel deviation prior p_3, we use a DAS-MedSea sequence in which the whale-related responses occur asynchronously across channels due to the spatial configuration of the DAS sensors.

As shown in Table 17, removing the prior corresponding to the dominant anomaly characteristic causes the largest degradation in both F1 and AUPRC. Specifically, removing p_1 yields the largest drop for the TF-energy-dominant ShipsEar case, while removing p_2 has the strongest impact on the amplitude-dominant SMAP case. These effects are also visually illustrated in Figs. 6 and 7, respectively, where removing the corresponding prior leads to less accurate anomaly localization. Similarly, the cross-channel-dominant DAS-MedSea case is most sensitive to the removal of p_3. These results confirm that the three priors capture distinct yet complementary anomaly characteristics.

B.13 Performance without Point Adjustment.

We further evaluate representative methods under the stricter no-point-adjustment protocol on PSM, SMD, and SWaT. As shown in Table 18, the no-point-adjustment protocol removes the benefit of point-wise adjustment and therefore provides a more challenging evaluation of anomaly localization. Compared with the PA setting, this protocol places greater emphasis on accurate point-level



Figure 6: Prior-specific comparison on a representative ShipsEar sequence with weak temporal amplitude variation but distinct time–frequency characteristics. The four rows compare the full model with variants removing each anomaly prior. Gray and purple curves denote the input and reconstructed series, respectively; pink regions indicate ground-truth anomalies, yellow regions indicate false alarms, and red boxes denote predicted anomalous intervals.

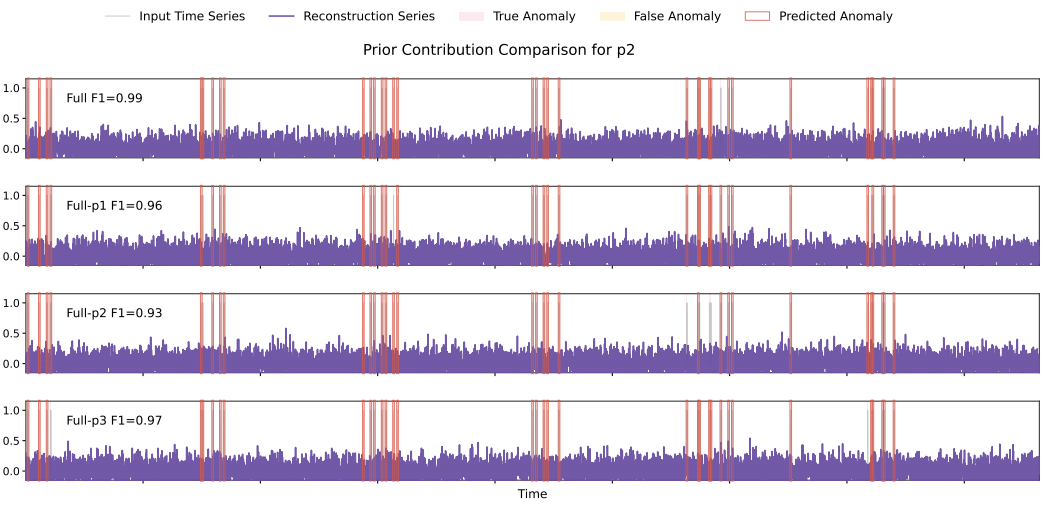


Figure 7: Prior-specific comparison on a representative SMAP sequence containing impulsive anomalies with prominent local amplitude deviations.

detection and reveals the precision–recall trade-offs of different methods. SCTF maintains balanced precision and recall across the three datasets, indicating that the proposed risk-adaptive reconstruction framework remains effective under stricter evaluation conditions.

B.14 Model Efficiency

We further compare the model complexity and inference efficiency of representative methods on STEAD. All experiments are conducted under the same experimental setting on a server equipped with two NVIDIA RTX 3090 GPUs. The number of trainable parameters is reported in millions (M). For inference efficiency, each test sample contains three channels with 6,000 time points, and the reported inference time is the average wall-clock time required to process one such sample.

Table 18: Anomaly detection performance under the no-point-adjustment protocol. Results are reported as mean ± standard deviation over three independent runs.

Dataset	Metric	Anomaly Transformer	CATCH	ImDiffusion	DiffAD	SCTF (Ours)
PSM	P	58.31 ± 3.42	69.07 ± 3.26	68.60 ± 3.19	72.66 ± 2.72	75.86 ± 2.14
	R	80.34 ± 2.87	73.64 ± 2.95	80.68 ± 2.63	71.00 ± 3.38	86.55 ± 1.87
	F1	67.57 ± 3.15	71.28 ± 3.11	74.15 ± 2.91	71.82 ± 3.06	**80.85** ± 2.03
SMD	P	91.80 ± 2.54	59.59 ± 3.68	77.13 ± 2.84	76.25 ± 2.91	83.09 ± 2.31
	R	10.69 ± 3.96	87.66 ± 2.41	72.62 ± 3.27	53.87 ± 4.12	85.41 ± 1.72
	F1	19.15 ± 4.21	70.95 ± 3.19	74.81 ± 3.02	63.14 ± 3.57	**84.23** ± 1.95
SWaT	P	59.21 ± 3.08	61.97 ± 3.52	58.74 ± 3.74	89.38 ± 2.24	72.98 ± 2.48
	R	57.84 ± 3.73	70.12 ± 2.78	87.59 ± 2.36	46.27 ± 4.35	79.59 ± 2.09
	F1	58.52 ± 3.34	65.79 ± 3.05	70.32 ± 3.13	60.97 ± 3.81	**76.14** ± 2.26
Avg. F1		48.41	69.34	73.09	65.31	**80.41**

Table 19: Comparison of model complexity and inference efficiency on STEAD.

Method	Trainable Params (M)	Inference Time (s)
Anomaly Transformer	4.74	0.12
CATCH	382.74	0.51
ImDiffusion	0.45	2.73
DiffAD	38.85	7.83
SCTF	2.92	1.45

As shown in Table 19, SCTF contains 2.92M trainable parameters and requires 1.45 s per sample for inference. The results also show that parameter count is not directly proportional to inference time, since computational cost is additionally affected by the model architecture and inference procedure. In particular, diffusion-based methods require iterative reverse sampling and therefore generally incur higher inference costs than models based on a single or limited number of forward passes. ImDiffusion and DiffAD require 2.73 s and 7.83 s per sample, respectively, whereas SCTF requires 1.45 s. Overall, SCTF maintains a moderate inference cost and model size among the evaluated methods.

C Limitations and Future Work

SCTF introduces additional computational cost due to the synchronized time–frequency diffusion process. Although DDIM sampling improves inference efficiency, further acceleration with more efficient diffusion solvers or few-step generative models is worth exploring. In addition, the current frequency decomposition and guidance schedule are fixed, and future work may investigate adaptive strategies for different time-series characteristics.

D Potential Negative Societal Impacts

SCTF may produce false alarms or missed detections when deployed in practical monitoring systems, which could affect downstream decisions. In addition, training on sensitive time-series measurements may introduce potential privacy risks.