TED-ITS: TAIL-ENHANCED CONDITIONAL DIFFUSION FINE-TUNING FOR LABEL-FREE INDUSTRIAL TIME-SERIES GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models have achieved strong performance in time-series forecasting, missing-value imputation, and generation. However, real-world industrial time series often exhibit long-tailed distributions, where abundant normal samples dominate rare fault patterns and degrade the modeling of minority faults. Although class-conditional diffusion models alleviate this issue through semantic class guidance, two challenges remain. First, rare fault samples provide insufficient supervision for learning class-specific denoising dynamics, while increasing diffusion noise progressively obscures fault-specific characteristics. Second, class-conditional diffusion relies on explicit labels, which are generally unavailable from historical or incomplete observations in forecasting and imputation, leading to a mismatch between training and inference. To address these challenges, we propose TED-ITS, a tail-enhanced fine-tuning framework built upon a pre-trained class-conditional diffusion model. TED-ITS introduces Noise-Adaptive Target Reconstruction (NATR), which adaptively constructs tail-class reconstruction targets according to noise-aware trajectory similarity and class frequency to strengthen minority fault supervision. It further incorporates Balanced-Condition Regularization (BCR), which samples auxiliary labels from a balanced distribution and enforces time- and frequency-domain consistency across different class conditions, thereby reducing reliance on explicit labels. Extensive experiments on multiple industrial benchmarks demonstrate that TED-ITS improves minority fault modeling and achieves superior performance on downstream forecasting and imputation tasks compared with existing methods.

1 INTRODUCTION

Diffusion models (DMs) have achieved impressive performance in a variety of general time-series analysis tasks, including forecasting and missing-value imputation (Rasul et al., 2021; Tashiro et al., 2021; Yuan & Qiao, 2024; Yang et al., 2026). However, real-world industrial data often follows a long-tailed distribution, where normal states dominate while fault events are rare. Under such an imbalance, diffusion models tend to overfit dominant normal patterns and struggle to generate realistic minority fault samples, thereby limiting their ability to capture critical tail-class behaviors (Cui et al., 2019; Kang et al., 2019; Menon et al., 2020).

Class conditioning provides informative semantic supervision that helps models to distinguish class-specific modes under long-tailed distributions. From a Bayesian perspective, class labels act as informative semantic priors (Zhang et al., 2024), providing persistent class guidance throughout the denoising process. Therefore, existing methods typically improve generation quality of minority patterns by strengthening the connection between each sample, its class label, and its clean target (Dhariwal & Nichol, 2021; Ho & Salimans, 2022). Despite this advantage, class conditioning alone is insufficient to address the long-tailed problem of industrial time-series modeling, where two closely related challenges remain. As shown in Fig. 1, minority fault classes remain severely underrepresented, resulting in insufficient learning of their class-specific denoising dynamics. As the noise level increases, these fault-specific characteristics become progressively obscured, amplifying the adverse effect of insufficient minority-class supervision caused by long-tailed data. Second, class-conditional diffusion models rely on explicit class labels during denoising. However,

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.

in downstream tasks such as forecasting and missing-value imputation, the ground-truth operating state of future or missing intervals is not observed at inference time. This creates a fundamental mismatch: class-conditional diffusion models rely on explicit class labels to guide the denoising process, whereas such labels cannot be assumed to be available during inference.

To address these challenges, we develop TED-ITS, a tail-enhanced fine-tuning framework built upon a pretrained class-conditional diffusion model. Considering that diffusion noise progressively obscures the state-specific differences between normal and fault samples at high noise levels, while their shared operating dynamics remain more prominent, we design a Noise-Adaptive Target Reconstruction (NATR) strategy. Specifically, the strategy selects compatible clean targets according to the similarity between noisy samples and candidate references at different noise levels, and reconstructs the supervision target only when a compatible tail-class target is identified, thereby enriching minority-class supervision without introducing excessive cross-class bias. To further reduce the model's dependence on explicit class labels, we introduce a balanced condition regularization strategy that samples labels from a balanced class distribution and combines them with noisy samples drawn from the original long-tailed distribution. By enforcing time- and frequency-domain consistency between denoising outputs under the original and balanced conditions, BCR reduces the sensitivity of the denoiser to class conditions, enabling downstream inference without requiring ground-truth class labels.

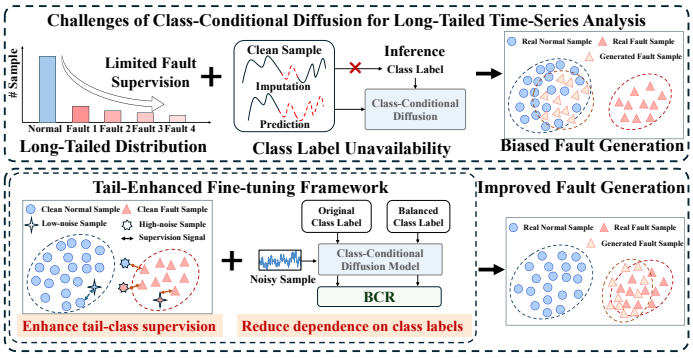


Figure 1: **Motivation for Tail-enhanced Diffusion Fine-tuning.** We address limited tail-class supervision and unavailable inference labels by enhancing tail supervision and reducing label dependence via BCR.

The main contributions of this work are summarized as follows:

- We propose TED-ITS, a Tail-Enhanced Diffusion fine-tuning framework for long-tailed Industrial Time-Series generation. By adaptively reconstructing supervision targets according to the noise level and introducing balanced condition regularization, TED-ITS enhances effective supervision for minority fault classes while reducing excessive dependence on explicit class labels, without requiring additional fault data.
- We introduce Noise-Adaptive Target Reconstruction (NATR), which uses noise-aware trajectory compatibility and class-frequency constraints to selectively transfer reconstruction targets toward minority operating states, increasing the opportunities for existing fault trajectories to provide reconstruction supervision.
- We develop Balanced-Condition Regularization (BCR), which introduces uniformly sampled auxiliary class conditions and regularizes predictions of the same noisy observation in the temporal and frequency domains. BCR promotes more balanced participation of class conditions and complements the fault-related supervision provided by NATR.
- Extensive experiments on multiple industrial benchmarks demonstrate the effectiveness of TED-ITS in tail-class generation quality and downstream industrial time-series analysis tasks.

2 Related Work

2.1 Diffusion Models for Time Series

Diffusion models have recently emerged as a powerful generative paradigm due to their stable training and strong distribution modeling capability (Sohl-Dickstein et al., 2015; Ho et al., 2020). Com-

pared with VAE- and GAN-based approaches, they can better capture complex temporal dependencies and generate more realistic time-series samples. As a result, diffusion models have attracted increasing attention in multivariate time-series analysis. Early studies mainly focused on task-specific applications. TimeGrad (Rasul et al., 2021) integrates diffusion processes into probabilistic forecasting, while CSDI (Tashiro et al., 2021) formulates time-series imputation as a conditional diffusion process. These works demonstrate the effectiveness of diffusion models for complex time-series distributions. Beyond specific task applications, diffusion-TS (Yuan & Qiao, 2024) introduces an interpretable framework that jointly captures temporal dynamics and feature dependencies for general multivariate time-series generation. Subsequent studies further enhance diffusion-based time-series modeling through transformer architectures and temporal attention mechanisms (Xu et al., 2025; Yang et al., 2026).

Diffusion models have also been applied to industrial time-series generation. DiffMTS (Ren et al., 2024) and MetaInduX-TS (Wang et al., 2025) generate realistic aero-engine degradation data, while DLTTS (Hu et al., 2026) further explores long-tailed industrial time-series generation. However, the use of class supervision and head-to-tail knowledge transfer remains largely unexplored. TED-ITS addresses this gap by leveraging class supervision during the training phase and explicitly transferring knowledge from data-rich head classes to data-scarce tail classes for improved minority fault generation.

2.2 Long-Tailed Generation

Long-tailed learning has been extensively studied in discriminative tasks through re-sampling (Chawla et al., 2002), re-weighting (Cui et al., 2019), decoupled learning (Kang et al., 2019), and logit adjustment (Menon et al., 2020). To further alleviate tail-class data scarcity, recent studies have explored information augmentation and knowledge transfer paradigms, including OLTR (Liu et al., 2019), GIST (Liu et al., 2021), RIDE (Wang et al., 2021), and CReST (Wei et al., 2021), which improve tail-class learning by transferring knowledge from data-rich head classes or exploiting additional unlabeled information. Building upon these advances, long-tailed learning has recently been extended to generative modeling. Representative approaches include CBDM (Qin et al., 2023), score-oriented diffusion calibration (Zhang et al., 2024), and DiffuLT (Shao et al., 2024), which improve minority-class generation through distribution balancing or knowledge transfer. These studies demonstrate the potential of diffusion models for addressing long-tail generation problems.

However, existing long-tailed diffusion methods are primarily designed for image generation and generally assume class labels are available during both training and inference. Moreover, they rarely consider the intrinsic degradation relationships among industrial operating states. Consequently, directly applying these methods to industrial time-series generation remains challenging.

3 Methodology

In this section, we present TED-ITS, a tail-enhanced conditional diffusion fine-tuning framework for long-tailed industrial time-series generation. We first formulate the learning problem and then detail the design of its two key components: Noise-Adaptive Target Reconstruction (NATR) and Balanced-Condition Regularization (BCR).

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ denote a labeled industrial time-series dataset, where $\mathbf{x}_i \in \mathbb{R}^{D\times T}$ represents a multivariate time series with D sensor channels and T timesteps, and $y_i \in \mathcal{C} = \{C_1, \ldots, C_K\}$ denotes the corresponding operating state label (e.g., normal state or a specific fault type). Our objective is to fine-tune a pretrained class-conditional diffusion model to better capture rare fault patterns in long-tailed industrial time series and support generation without ground-truth class labels at inference.

3.1 Noise-Adaptive Target Reconstruction (NATR)

Under long-tailed training, rare fault trajectories provide insufficient reconstruction supervision, while frequent operating states dominate the learning process. To strengthen fault supervision, we consider how the forward diffusion process affects the distinguishability of different trajectories. As noise increases, the noisy distributions induced by different clean trajectories overlap more

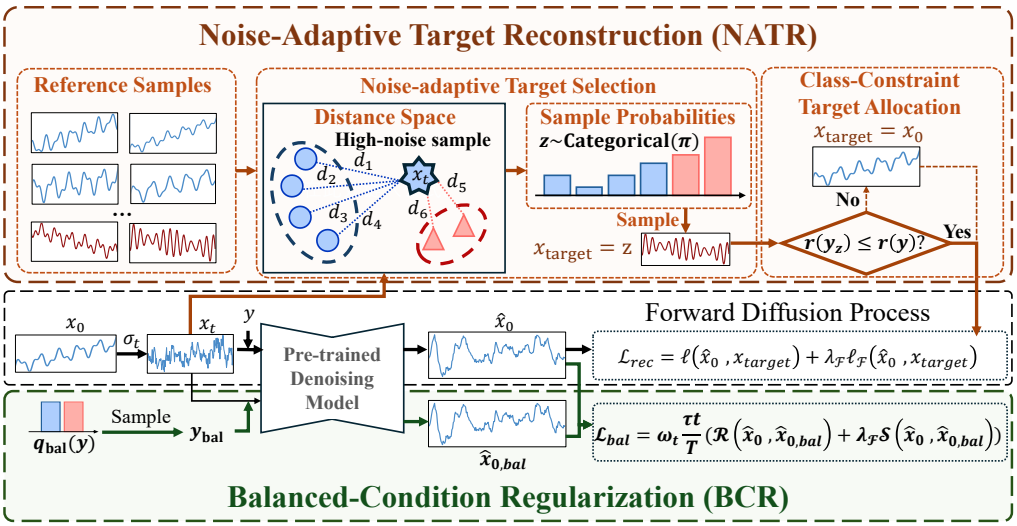


Figure 2: **Overview of TED-ITS.** A pretrained diffusion model is fine-tuned using two complementary components. Noise-Adaptive Target Reconstruction (NATR) increases reconstruction supervision for minority fault patterns through noise-aware target transfer, while Balanced-Condition Regularization (BCR) introduces balanced auxiliary class conditions and enforces temporal- and frequency-domain consistency across different conditions to reduce dependence on explicit class labels.

strongly, allowing some noisy observations originating from frequent operating states to also be compatible with rare fault trajectories (see Appendix A.2 for details). Motivated by this property, we introduce NATR, which probabilistically replaces a subset of reconstruction targets with rare fault trajectories. As shown in Fig. 2, NATR consists of two steps: noise-adaptive target selection and class-constrained target allocation. The resulting targets are then used as supervision to optimize the model.

Noise-adaptive target selection. Fig. 2 illustrates the target selection process. For a sample $(x_0^{(i)}, y_i)$, we first normalize the noisy observation $\widetilde{x}_{t_i}^{(i)}$, and map it into a distance space by computing its squared Frobenius distance to each clean candidate in the reference pool. Using the clean sequences in the reference pool, we compute the distance between the normalized observation and each candidate as

$$d_{ij} = \left\| \widetilde{x}_{t_i}^{(i)} - x_0^{(j)} \right\|_F^2 . \tag{1}$$

These distances are converted into sampling probabilities as

$$\pi_{ij} = \frac{\exp\left[-d_{ij}/(2\sigma_{t_i}^2)\right]}{\sum_{k=1}^{M} \exp\left[-d_{ik}/(2\sigma_{t_i}^2)\right]}, \tag{2}$$

where σ_{t_i} denotes the noise standard deviation of the normalized observation at timestep t_i, and M is the number of samples in the reference pool. Based on the sampling probabilities $\boldsymbol{\pi}_i$, we select a candidate target $x_0^{(z_i)}$, where $z_i \sim \mathrm{Categorical}(\boldsymbol{\pi}_i)$.

Proposition 3.1 (Noise-dependent reconstruction-target selection)**.** *Fix a reference pool $\mathcal{M}$ containing a source time series $x_0^{(i)}$ and a candidate $x_0^{(j)}$ with $j \neq i$. Let $x_{t_i}^{(i)}$ follow the forward diffusion distribution $q(x_{t_i}^{(i)} \mid x_0^{(i)})$ with normalized noise variance $\sigma_{t_i}^2 > 0$. Under the matching probabilities in Eq. (2), the expected probability of selecting the clean trajectory $x_0^{(j)}$ as the candidate reconstruction target satisfies*

$$\mathbb{E}_{x_{t_i}^{(i)} \sim q(x_{t_i}^{(i)} | x_0^{(i)})}[\pi_{ij}] \leq \frac{1}{2} \exp\left(-\frac{\|x_0^{(i)} - x_0^{(j)}\|_F^2}{8\sigma_{t_i}^2} \right). \tag{3}$$

Proposition 3.1 shows that selecting candidates far from the source relative to the noise scale is suppressed. Larger noise scales make this bound less restrictive, without implying that the actual selection probability increases monotonically. The candidate distribution is recomputed as noisy observations and mini-batch composition change, enabling dynamic target selection. This matching captures observation compatibility rather than fault-semantic equivalence. To direct target allocation toward rare states, we next introduce a class-constraint target allocation.

Class-constraint target allocation. To direct target allocation toward rare states, the sampled candidate must satisfy the class-frequency constraint shown in Fig. 2. Let $r(c) = N_c/N$ denote the empirical frequency of class c. The reconstruction target is defined as

$$x_{\text{target}}^{(i)} = \begin{cases} x_0^{(z_i)}, & r(y_{z_i}) \le r(y_i) \text{ and } t_i < t_{\text{cut}}, \\ x_0^{(i)}, & \text{otherwise}, \end{cases} \tag{4}$$

where t_{cut} specifies the diffusion stages in which target transfer is permitted. If the candidate satisfies the constraints, it becomes the reconstruction target; otherwise, the original target is retained through the fallback path. This rule allows some training instances from frequent states to use compatible rare-state trajectories as supervision, increasing opportunities for existing fault samples to contribute to reconstruction learning.

Reconstruction with transferred targets. For each noisy input $x_{t_i}^{(i)}$, the original branch retains the training label y_i and produces $\widehat{x}_0^{(i)} = f_\theta(x_{t_i}^{(i)}, t_i, y_i)$. We supervise this prediction using the tail-aware target $x_{\text{target}}^{(i)}$ constructed by NATR:

$$\mathcal{L}_{\text{rec}} = \frac{1}{N}\sum_{i=1}^{N} w(t_i)\Big[\ell\Big(\widehat{x}_0^{(i)}, x_{\text{target}}^{(i)}\Big) + \lambda_f \ell_{\mathcal{F}}\Big(\widehat{x}_0^{(i)}, x_{\text{target}}^{(i)}\Big)\Big], \tag{5}$$

where ℓ and $\ell_{\mathcal{F}}$ denote the ℓ_1 reconstruction losses in the temporal and frequency domains, respectively. When target transfer is applied, compatible minority fault trajectories can supervise predictions under other original class conditions. This broadens fault-pattern learning beyond fault-labeled training examples.

Theoretical justification. We further investigate the reliability of NATR in Appendix A.2. Our analysis establishes likelihood and distance bounds for noise-adaptive target matching, and shows that the class-frequency constraint guarantees a non-decreasing allocation of reconstruction targets to the tail-class set. Importantly, since NATR modifies reconstruction targets, we characterize the resulting reconstruction bias. Under squared-error loss, the deviation between the original and transferred population-optimal reconstruction functions is bounded by the expected squared target displacement. These results characterize the trade-off between increasing minority-target supervision and introducing reconstruction bias, providing theoretical motivation for selective target transfer.

3.2 Balanced-Condition Regularization (BCR)

To address the mismatch in class-label availability between training and inference, we further introduce Balanced-Condition Regularization (BCR). As shown in Fig. 2, BCR samples auxiliary labels from a balanced class distribution during fine-tuning. Then, it encourages time- and frequency-domain consistency between predictions of the same noisy observation under the original and auxiliary class conditions. BCR jointly optimizes the denoising predictions under the original class condition and the balanced auxiliary condition. Together with NATR, it aims to strengthen fault-pattern learning while reducing reliance on ground-truth class labels at inference.

Balanced auxiliary-condition sampling. For the same noisy input $x_{t_i}^{(i)}$, we sample an auxiliary class condition uniformly:

$$y_i^{\text{bal}} \sim \text{Uniform}(\{1, \dots, K\}), \tag{6}$$

where K is the number of operating states. As illustrated in Fig. 2, the balanced branch uses the shared denoising model to produce $\widehat{x}_{0,\text{bal}}^{(i)} = f_\theta\Big(x_{t_i}^{(i)}, t_i, y_i^{\text{bal}}\Big)$. Uniform sampling balances the

exposure of class conditions during fine-tuning. The auxiliary label is used only as an alternative condition; it does not change the noisy input or serve as a new ground-truth label.

Time- and frequency-domain consistency. We regularize the predictions of the original and balanced branches for the same noisy input to encourage the reconstructed patterns to remain accessible across class conditions. For notational simplicity, let $h_i = \widehat{x}_0^{(i)}$ and $h_i^{\mathrm{bal}} = \widehat{x}_{0,\mathrm{bal}}^{(i)}$. The temporal consistency term is

$$\mathcal{R}_i = \ell\big(h_i, \mathrm{sg}(h_i^{\mathrm{bal}})\big) + \lambda_c \ell\big(\mathrm{sg}(h_i), h_i^{\mathrm{bal}}\big)\,, \tag{7}$$

where $\mathrm{sg}(\cdot)$ stops gradient propagation and λ_c controls the relative contribution of the auxiliary-branch update. The frequency-domain term compares the real and imaginary components of the temporal Fourier transforms:

$$\mathcal{S}_i = \ell\big(\mathrm{Re}\,\mathcal{F}(h_i), \mathrm{Re}\,\mathcal{F}(h_i^{\mathrm{bal}})\big) + \ell\big(\mathrm{Im}\,\mathcal{F}(h_i), \mathrm{Im}\,\mathcal{F}(h_i^{\mathrm{bal}})\big)\,. \tag{8}$$

Both predictions receive gradients from the frequency-domain term. The resulting balanced-condition loss is

$$\mathcal{L}_{\mathrm{bal}} = \frac{1}{N}\sum_{i=1}^{N} w(t_i)\frac{\tau t_i}{T}\left(\mathcal{R}_i + \lambda_f \mathcal{S}_i\right), \tag{9}$$

where τ controls the regularization strength, T is the total number of diffusion steps, and $w(t_i)$ is the backbone's timestep weight.

Joint fine-tuning objective. The whole objective of the fine-tuning framework combines reconstruction with transferred targets and balanced-condition regularization:

$$\mathcal{L}_{\mathrm{FT}} = \mathcal{L}_{\mathrm{rec}} + \mathcal{L}_{\mathrm{bal}}. \tag{10}$$

The reconstruction term strengthens fault-pattern learning beyond fault-labeled examples, while the consistency term encourages these patterns to be retained across class conditions. Together, they aim to reduce the dependence of minority fault modeling on explicit target-state labels at inference.

4 Experiments

We evaluate TED-ITS on long-tailed industrial time-series benchmarks. After introducing the experimental setup, we assess label-free generation against real fault references and examine the model's dependence on target-state labels. We then conduct ablation studies of NATR and BCR, evaluate the generated data on downstream imputation and forecasting tasks, and analyze sensitivity to missing ratios and forecasting horizons.

4.1 Experimental Setup

Datasets. We evaluate TED-ITS on five datasets covering both public benchmarks and real-world industrial scenarios, including two Electricity Transformer Temperature (ETT) benchmarks (ETTh1 and ETTm1), two industrial anomaly detection datasets (SWaT and SKAB), and a real-world wind turbine monitoring dataset (WTSD).

1. **ETTh1** and **ETTm1** are two variants of the ETT dataset containing seven sensor variables collected from electricity transformers. As ground-truth anomaly labels are unavailable, we follow prior work (Hu et al., 2026) and regard observations with oil temperature values below 4 as anomalous, resulting in anomaly ratios of 9.87% and 9.94%, respectively.
2. **SWaT** is a real-world water treatment dataset containing 51 sensor and actuator variables with labeled attack events, yielding an anomaly ratio of 12.14%.
3. **SKAB** is an 8-dimensional industrial IoT dataset collected from a circulation pump testbed. Following prior work, we use the valve1 subset, whose anomaly ratios range from 1.86% to 2.23%.
4. **WTSD** is a wind turbine SCADA dataset containing measurements such as wind speed, wind direction, and power output. Since anomaly labels are unavailable, anomalies are identified according to deviations between actual generated power and the theoretical power curve.

Table 1: **Comparison of generation quality on four benchmark datasets.** TED-ITS ranks first or second in 22 of the 24 evaluation settings, demonstrating its effectiveness for tail industrial time-series generation.

Dataset		S	Diffusion-TS (Yuan & Qiao, 2024)	DiffWave (Kong et al., 2021)	DiffMTS (Ren et al., 2024)	TabDDPM (Kotelnikov et al., 2023)	DLTTS (Hu et al., 2026)	TED-ITS
ETTh1	Predictive score (Lower the Better)	32	0.2153	<u>0.1029</u>	0.1631	4.3581	0.1768	**0.0657**
		64	0.1165	0.1189	0.1327	5.3624	<u>0.1143</u>	**0.0600**
		128	0.1821	0.1586	0.1515	6.5121	<u>0.1446</u>	**0.0596**
	Discriminative score (Lower the Better)	32	<u>0.4155</u>	0.4889	0.4348	0.5000	0.4820	**0.3188**
		64	0.4698	0.4720	0.4516	0.4991	<u>0.4505</u>	**0.3420**
		128	0.4415	0.4639	0.4453	0.4995	<u>0.4341</u>	**0.3542**
ETTm1	Predictive score (Lower the Better)	32	0.1433	0.1974	0.6972	3.3010	<u>0.1349</u>	**0.1096**
		64	0.1678	0.2534	0.6926	3.7892	<u>0.1604</u>	**0.0790**
		128	0.1874	0.2700	0.6852	4.4655	<u>0.1798</u>	**0.0841**
	Discriminative score (Lower the Better)	32	0.4399	0.4731	**0.3115**	0.4826	0.4071	<u>0.3652</u>
		64	0.4269	0.4874	<u>0.3726</u>	0.4999	0.4048	**0.3491**
		128	0.4461	0.4801	<u>0.3862</u>	0.4991	0.3833	**0.3677**
SWaT	Predictive score (Lower the Better)	32	0.0010	0.0057	0.0153	**0.0001**	0.0020	<u>0.0004</u>
		64	0.0040	0.0029	0.0163	**0.0001**	0.0078	<u>0.0020</u>
		128	0.0010	0.0026	0.0132	<u>0.0003</u>	**0.00005**	0.0005
	Discriminative score (Lower the Better)	32	0.4999	0.4995	0.4999	0.4999	<u>0.4848</u>	**0.4315**
		64	0.4999	0.5000	0.5000	0.4999	<u>0.4476</u>	**0.4148**
		128	0.4999	0.4958	0.4951	0.4998	<u>0.4665</u>	**0.4595**
SKAB	Predictive score (Lower the Better)	32	<u>0.1971</u>	**0.1336**	0.6000	30.587	0.2486	0.2082
		64	0.2100	**0.1479**	0.6065	30.595	0.2442	<u>0.2068</u>
		128	**0.1795**	0.1972	0.6287	29.342	0.1934	<u>0.1899</u>
	Discriminative score (Lower the Better)	32	0.4993	0.4940	0.4893	0.4529	<u>0.4291</u>	**0.3632**
		64	0.4966	0.4945	0.4986	<u>0.4322</u>	0.4622	**0.3656**
		128	0.4853	0.4882	0.4995	0.3928	**0.3623**	<u>0.3751</u>

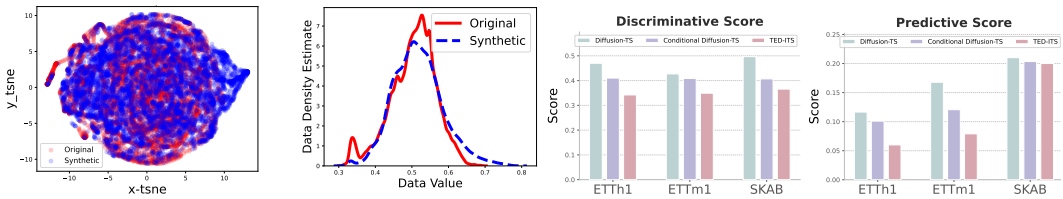
Evaluation metrics. We evaluate generated time series using qualitative visualization, Predictive Score (PS), Discriminative Score (DS), and downstream task performance. PCA and t-SNE visualize the distributions of real and generated samples. PS (↓) measures forecasting error when a predictor trained on generated data is evaluated on real data, while DS (↓) measures how easily a classifier distinguishes generated samples from real samples. Downstream imputation and forecasting performance is evaluated using Mean Squared Error (MSE, ↓). Implementation details are provided in Appendix B.

Baselines. We compare TED-ITS with five representative baselines: **Diffusion-TS** (Yuan & Qiao, 2024), **DiffWave** (Kong et al., 2021), and **DiffMTS** (Ren et al., 2024), which are diffusion-based generation methods; **TabDDPM** (Kotelnikov et al., 2023), a diffusion model for tabular data generation; and **DLTTS** (Hu et al., 2026), the current state-of-the-art method for long-tailed industrial time series generation. For all baselines, the reported tail-class generation results are taken from DLTTS (Hu et al., 2026). We evaluate TED-ITS under the same experimental settings and evaluation protocol to ensure a fair comparison.

Implementation details. TED-ITS is implemented in PyTorch and trained on a single NVIDIA RTX 4090 GPU. We adopt a class-conditional diffusion model and employ a Transformer-based denoising backbone. The model is optimized using AdamW with a learning rate of 1×10^{-4} and a batch size of 128. We adopt classifier-free training by replacing class labels with a null condition with probability 0.1, enabling joint learning of class-conditional and unconditional denoising. At inference, we use the null condition to generate samples without ground-truth class labels.

4.2 Tail Time Series Generation

Unconditional tail time-series generation. Table 1 reports the quantitative comparison of **tail time series generation** across four benchmark datasets. TED-ITS achieves the best or second-best performance in 22 out of 24 settings, consistently outperforming existing generation methods under long-tailed industrial distributions. The improvements are particularly evident on the ETT datasets,



(a) t-SNE visualization. (b) Kernel visualization. (c) Discriminative Score. (d) Predictive Score.

Figure 3: **Qualitative and quantitative evaluation of tail time-series generation.** (a) t-SNE visualization of real and generated tail time series. (b) Kernel-based visualization of real and generated tail time series. (c–d) Impact of conditional pretraining and tail-enhanced fine-tuning on generation quality in terms of the Discriminative Score and Predictive Score, respectively.

Table 2: **Ablation study of NATR and BCR on three datasets.** Lower values indicate better performance. The best results are highlighted in bold.

Method	ETTh1			ETTm1			SWaT		
	$S = 32$	$S = 64$	$S = 128$	$S = 32$	$S = 64$	$S = 128$	$S = 32$	$S = 64$	$S = 128$
Predictive score (Lower is better)									
w/o NATR	0.0825	0.0799	0.0642	0.1183	0.1192	0.0960	0.0012	0.0031	0.0030
w/o BCR	0.0920	0.0907	0.1022	0.2123	0.1227	0.1123	0.0019	0.0024	0.0043
TED-ITS	**0.0657**	**0.0600**	**0.0596**	**0.1096**	**0.0790**	**0.0841**	**0.0004**	**0.0020**	**0.0005**
Discriminative score (Lower is better)									
w/o NATR	0.3606	0.3968	0.3925	0.4158	0.3948	0.4020	0.4675	0.4561	0.4668
w/o BCR	0.3728	0.3954	0.3899	0.4123	0.4205	0.4080	0.4386	0.4722	0.4621
TED-ITS	**0.3188**	**0.3420**	**0.3542**	**0.3652**	**0.3491**	**0.3677**	**0.4315**	**0.4148**	**0.4595**

where TED-ITS achieves the lowest predictive scores across all sequence lengths, indicating that its generated samples are more useful for training predictors evaluated on real tail sequences. On the industrial datasets SWaT and SKAB, TED-ITS consistently achieves superior Discriminative scores while maintaining competitive Predictive performance. These results indicate that the generated samples better preserve the underlying characteristics of rare events and exhibit improved distributional fidelity under long-tailed settings. Fig. 3 (a) and (b) provide further qualitative evidence for the effectiveness of TED-ITS, suggesting that TED-ITS generates samples with high fidelity and distributional alignment to real data.

Effectiveness of conditional pretraining and tail-enhanced fine-tuning. Fig. 3 (c) and (d) compare Diffusion-TS, Conditional Diffusion-TS, and TED-ITS under the same label-free generation protocol, using real tail sequences as the evaluation reference. Conditional Diffusion-TS improves generation performance over Diffusion-TS, showing the value of class supervision during training even when target-state labels are unavailable at inference. TED-ITS achieves further improvements, indicating that tail-enhanced fine-tuning builds on this supervised foundation to better match the real tail reference.

4.3 Ablation Study

We evaluate the contributions of NATR and BCR under the label-free generation protocol, using real tail sequences as the evaluation reference. As shown in Table 2, removing either component worsens both Predictive and Discriminative scores across all evaluated settings. The performance drop without NATR supports the benefit of increasing the participation of rare fault trajectories in reconstruction supervision. Removing BCR also degrades performance, indicating that its balanced auxiliary conditions and cross-condition consistency regularization contribute to generation quality when ground-truth class labels are unavailable.

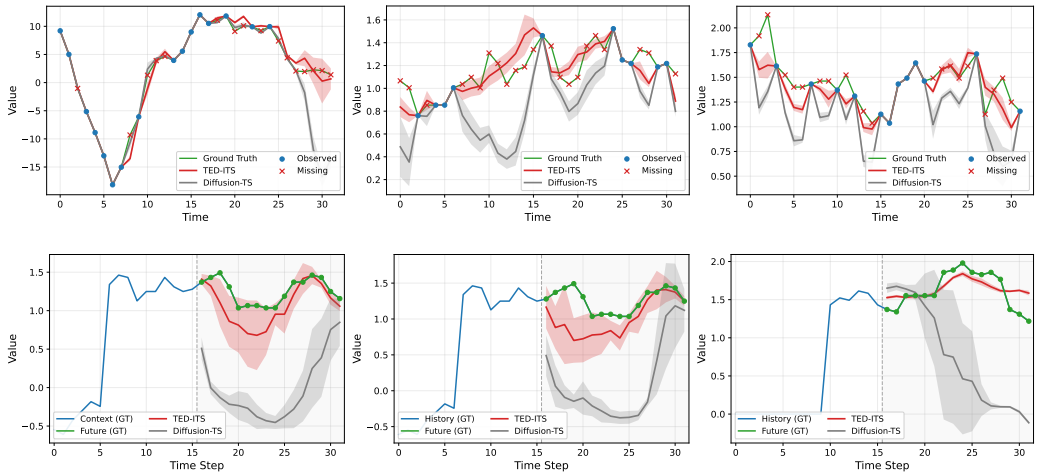


Figure 4: **Representative imputation (top) and forecasting (bottom) results on ETTh1.** TED-ITS consistently produces trajectories closer to the ground truth than Diffusion-TS.

Table 3: **Imputation and forecasting performance on ETTh1 and WTSD.** S denotes the missing ratio for imputation and the prediction horizon for forecasting. Lower MSE indicates better performance. The best results are highlighted in bold.

Method	ETTh						WTSD					
	Imputation			Prediction			Imputation			Prediction		
	$S=0.6$	$S=0.7$	$S=0.8$	$S=8$	$S=16$	$S=24$	$S=0.6$	$S=0.7$	$S=0.8$	$S=8$	$S=16$	$S=24$
Diffusion-TS	0.0117	0.0124	0.0128	0.0404	0.0451	0.0529	0.0263	0.0275	0.0282	0.0892	0.1253	0.1504
TED-ITS	**0.0109**	**0.0115**	**0.0118**	**0.0381**	**0.0409**	**0.0486**	**0.0262**	**0.0274**	**0.0281**	**0.0888**	**0.1237**	**0.1480**

4.4 Imputation and Forecasting Performance

We further evaluate whether the benefits of tail-enhanced fine-tuning extend to downstream tasks where target-state labels are unavailable. Fig. 4 shows representative imputation and forecasting results on ETTh1. Compared with Diffusion-TS, TED-ITS more closely follows the ground-truth trajectories, including regions with pronounced fluctuations. As reported in Table 3, TED-ITS also achieves lower MSE across different missing ratios and forecasting horizons on ETTh1 and WTSD. These results demonstrate the downstream utility of jointly strengthening minority-pattern learning through NATR and reducing dependence on specific class conditions through BCR.

5 Conclusion

We proposed TED-ITS, a tail-enhanced conditional diffusion fine-tuning framework for label-free generation of industrial fault time series. Built upon pretrained Conditional Diffusion-TS, TED-ITS leverages class supervision during training and retains its benefits when target-state labels are unavailable at inference. Noise-Adaptive Target Reconstruction (NATR) increases the opportunities for minority fault trajectories to provide reconstruction supervision, strengthening rare fault-pattern learning. Building on these targets, Balanced-Condition Regularization (BCR) jointly optimizes reconstruction and cross-condition prediction consistency to reduce the dependence of learned patterns on specific class conditions. Experiments show that TED-ITS improves label-free generation evaluated against real tail sequences, as well as downstream imputation and forecasting performance. TED-ITS currently relies on fault patterns observed during training and has not been evaluated on unseen fault types. Future work could explore zero-shot generalization to unseen fault types, extending TED-ITS beyond the fault patterns observed during training.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean or reformat datasets, support qualitative or thematic data analysis, or interpret experimental results. Additionally, we used generative AI tools to improve the grammar, clarity, and readability of the manuscript.

All AI-assisted text was carefully reviewed and revised by the authors to ensure technical correctness and consistency with the intended scientific content. Generative AI tools were not used to determine the scientific claims or conclusions of this work.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of the results presented in this work. Section 3 in the main paper provides a detailed description of the proposed method and its training procedure. Implementation details, including model architectures, training configurations, hyperparameters, dataset construction and preprocessing, and evaluation protocols, are provided in Section 4 and appendix. Additional experimental results and ablation studies are included in the appendix to further support the reported findings.

ETHICS STATEMENT

This work focuses on the development and evaluation of machine learning methods using publicly available benchmark datasets. The study does not involve human subjects, personally identifiable information, or sensitive private data. To the best of our knowledge, the datasets used in our experiments are employed in accordance with their intended research purposes and applicable terms of use. Like other generative modeling techniques, the proposed method may be used in applications beyond those evaluated in this work. Our experiments are intended to evaluate the technical properties of the proposed method rather than to support harmful uses.

REFERENCES

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in neural information processing systems*, volume 34, pp. 8780–8794, 2021.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Weijun Hu, Haifeng Zhang, Wenbin Chen, Kaihong Chen, Jinhao Long, Wenli Shang, and Zhong Cao. Dltts: Diffusion model for long-tailed time series generation in industrial scenarios. *IEEE Transactions on Knowledge and Data Engineering*, 2026.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. 2021.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International conference on machine learning*, pp. 17564–17579. PMLR, 2023.

Bo Liu, Haoxiang Li, Hao Kang, Gang Hua, and Nuno Vasconcelos. Gistnet: A geometric structure transfer network for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8209–8218, October 2021.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2537–2546, 2019.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *arXiv preprint arXiv:2007.07314*, 2020.

Frank Nielsen. Generalized bhattacharyya and chernoff upper bounds on bayes error using quasi-arithmetic means. *Pattern Recognition Letters*, 42:25–34, 2014.

Yiming Qin, Huangjie Zheng, Jiangchao Yao, Mingyuan Zhou, and Ya Zhang. Class-balancing diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18434–18443, 2023.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International conference on machine learning*, pp. 8857–8868. PMLR, 2021.

Lei Ren, Haiteng Wang, and Yuanjun Laili. Diff-mts: Temporal-augmented conditional diffusion-based aigc for industrial time series toward the large model era. *IEEE Transactions on Cybernetics*, 54(12):7187–7197, 2024.

Jie Shao, Ke Zhu, Hanxiao Zhang, and Jianxin Wu. Diffult: Diffusion for long-tail recognition without external knowledge. *Advances in Neural Information Processing Systems*, 37:123007–123031, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

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, pp. 24804–24816, 2021.

Haiteng Wang, Lei Ren, Yikang Li, and Yuqing Wang. Metaindux-ts: Frequency-aware aigc foundation model for industrial time series. *IEEE transactions on neural networks and learning systems*, 2025.

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021.

Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10857–10866, 2021.

Yi Xu, Zhigang Chen, Rui Wang, Yangfan Li, Fengxiao Tang, Ming Zhao, and Jiaqi Liu. Faultdiffusion: Few-shot fault time series generation with diffusion model. *arXiv preprint arXiv:2511.15174*, 2025.

Yilun Xu, Shangyuan Tong, and Tommi Jaakkola. Stable target field for reduced variance score estimation in diffusion models. *arXiv preprint arXiv:2302.00670*, 2023.

Yiyuan Yang, Ming Jin, Haomin Wen, Chaoli Zhang, Yuxuan Liang, Lintao Ma, Yi Wang, Chenghao Liu, Bin Yang, Zenglin Xu, et al. A survey on diffusion models for time series and spatio-temporal data. *ACM Computing Surveys*, 58(8):1–39, 2026.

Xinyu Yuan and Yan Qiao. Diffusion-ts: Interpretable diffusion for general time series generation. In *International Conference on Learning Representations*, volume 2024, pp. 41582–41610, 2024.

Tianjiao Zhang, Huangjie Zheng, Jiangchao Yao, Xiangfeng Wang, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Long-tailed diffusion models with oriented calibration. In *The twelfth international conference on learning representations*, 2024.

A Theoretical Analysis

A.1 Bayesian View of Label Supervision.

The optimal denoising score of a diffusion model can be expressed as

$$\nabla_{x_t} \log q_t(x_t) = \mathbb{E}_{q_t(x_0|x_t)} \left[\nabla_{x_t} \log q_t(x_t|x_0) \right], \tag{11}$$

which indicates that the score function is obtained by averaging all possible denoising directions associated with clean samples x_0 that could generate the noisy observation x_t.

Under long-tailed industrial data distributions, the posterior distribution $q(x_0|x_t)$ is dominated by abundant head-class samples. Consequently, the learned score field tends to favor normal operating patterns, while the denoising directions associated with rare fault states receive insufficient emphasis.

When class labels are available, the posterior distribution can be refined as $q(x_0|x_t) \to q(x_0|x_t, y)$, where the label y provides additional semantic information. According to Bayes' theorem,

$$q(x_0|x_t, y) = \frac{q_t(x_t|x_0)q(x_0|y)}{q_t(x_t|y)}. \tag{12}$$

Compared with the unconditional posterior, the conditional posterior constrains the reference space to samples belonging to the target class, thereby reducing interference from dominant head classes and enabling the model to learn more discriminative denoising directions for tail faults. Therefore, class labels can be regarded as informative semantic priors that constrain the posterior reference space and guide the model toward class-specific denoising directions, thereby mitigating long-tail bias and improving tail-class generation. Moreover, labels reveal the head-tail structure of industrial data, enabling knowledge transfer from data-rich head classes to data-scarce tail classes.

A.2 Theoretical Analysis of Noise-Adaptive Target Reconstruction

This subsection analyzes Noise-Adaptive Target Reconstruction (NATR) from three perspectives: the compatibility of candidate targets with noisy observations, the allocation of reconstruction targets across classes, and the reconstruction bias induced by target replacement.

Labels are available during training only for the frequency-based target-selection rule and for analyzing supervision allocation. Our analysis distinguishes reconstruction-target allocation from unconditional generation quality. Increasing the frequency with which minority samples serve as reconstruction targets does not, by itself, guarantee improved fidelity or coverage of minority states in unconditional samples.

Consider a labeled dataset $\mathcal{D} = \{(x_0^{(i)}, y_i)\}_{i=1}^N$, where $x_0^{(i)}$ is a multivariate time series. We use L for the sequence length and T_{diff} for the number of diffusion steps to avoid ambiguity. Let $r(c) = N_c/N$ denote the empirical frequency of class c.

At diffusion timestep t, the forward process is

$$x_t^{(i)} = \sqrt{\bar{\alpha}_t}\, x_0^{(i)} + \sqrt{1 - \bar{\alpha}_t}\, \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, I). \tag{13}$$

For timesteps satisfying $0 < \bar{\alpha}_t < 1$, define the normalized noisy sample as

$$\widetilde{x}_{t_i}^{(i)} = \frac{x_t^{(i)}}{\sqrt{\bar{\alpha}_t}} = x_0^{(i)} + \sigma_t \epsilon_i, \qquad \sigma_t^2 = \bar{\alpha}_t^{-1} - 1. \tag{14}$$

For a fixed timestep, this normalization is invertible. Let $\mathcal{M}$ be a reference pool containing the original sample indexed by i. NATR assigns each candidate $j \in \mathcal{M}$ the matching weight

$$a_{ij} = \exp\left(-\frac{\|\widetilde{x}_{t_i}^{(i)} - x_0^{(j)}\|_F^2}{2\sigma_t^2}\right) \tag{15}$$

and selection probability

$$\pi_{ij} = \frac{a_{ij}}{\sum_{k\in\mathcal{M}} a_{ik}}. \tag{16}$$

A candidate index is drawn as $J_i \sim \mathrm{Categorical}(\{\pi_{ij}\}_{j\in\mathcal{M}})$.

Define the non-self accepted replacement indicator $A_i = \mathbf{1}\{J_i \neq i,\ r(y_{J_i}) \leq r(y_i),\ t < t_{\mathrm{cut}}\}$. The reconstruction target is

$$x_{\mathrm{target}}^{(i)} = x_0^{(i)} + A_i\big(x_0^{(J_i)} - x_0^{(i)}\big). \tag{17}$$

This expression is equivalent to the NATR rule: rejected candidates and self-selected candidates leave the target unchanged. A non-self replacement can also have zero displacement if two indexed samples coincide.

A.2.1 Likelihood and Geometric Compatibility

For a source sample indexed by i, fix its diffusion timestep t_i with $\sigma_{t_i} > 0$, and let $\widetilde{x}_{t_i}^{(i)} = x_{t_i}^{(i)}/\sqrt{\bar{\alpha}_{t_i}}$ denote the normalized noisy observation. Let M be the dimension of a flattened sequence. The Frobenius norm of a matrix-valued sequence equals the Euclidean norm of its flattened representation. Let z denote a normalized noisy observation and x a clean reference sequence, the normalized forward density is

$$q_{t_i}(z \mid x) = (2\pi\sigma_{t_i}^2)^{-M/2} \exp\left(-\frac{\|z - x\|_F^2}{2\sigma_{t_i}^2}\right). \tag{18}$$

Since the Gaussian normalization factor is common to all candidates,

$$\pi_{ij} = \frac{q_{t_i}(\widetilde{x}_{t_i}^{(i)} \mid x_0^{(j)})}{\sum_{k\in\mathcal{M}} q_{t_i}(\widetilde{x}_{t_i}^{(i)} \mid x_0^{(k)})}. \tag{19}$$

Thus, the matching probabilities are normalized likelihoods over the reference samples. They can also be interpreted as posterior probabilities over candidate indices under a uniform prior on the fixed reference pool. This interpretation does not imply semantic equivalence between the original and candidate trajectories.

Proof of Proposition 3.1. The proof follows the Gaussian-overlap argument in Appendix C of Zhang et al. (2024), using normalized forward densities without class-prior weighting in the matching probabilities.

Proof. Fix the reference pool $\mathcal{M}$ and timestep t_i. For notational simplicity, define $p_k(z) = q_{t_i}(z \mid x_0^{(k)})$, where q_{t_i} is the normalized forward density in Eq. (18). The normalized noisy observation $\widetilde{x}_{t_i}^{(i)}$ has density p_i. Because normalization is an invertible deterministic transformation at fixed t_i, the expectation over the original noisy observation can be written as

$$\mathbb{E}_{x_{t_i}^{(i)} \sim q(x_{t_i}^{(i)} \mid x_0^{(i)})}[\pi_{ij}] = \int \frac{p_i(z)p_j(z)}{\sum_{k\in\mathcal{M}} p_k(z)}\,\mathrm{d}z. \tag{20}$$

Since $\mathcal{M}$ contains both i and $j \neq i$, and all Gaussian densities are strictly positive,

$$\int \frac{p_i(z)p_j(z)}{\sum_{k\in\mathcal{M}} p_k(z)}\,\mathrm{d}z \leq \int \frac{p_i(z)p_j(z)}{p_i(z) + p_j(z)}\,\mathrm{d}z \leq \frac{1}{2}\int \sqrt{p_i(z)p_j(z)}\,\mathrm{d}z. \tag{21}$$

The second inequality follows from $a + b \geq 2\sqrt{ab}$ for $a, b > 0$. The remaining integral is the Bhattacharyya coefficient, which measures distributional overlap and admits a closed-form expression for Gaussian densities (Nielsen, 2014).

Let $m_{ij} = (x_0^{(i)} + x_0^{(j)})/2$. Completing the square gives

$$\|z - x_0^{(i)}\|_F^2 + \|z - x_0^{(j)}\|_F^2 = 2\|z - m_{ij}\|_F^2 + \frac{1}{2}\|x_0^{(i)} - x_0^{(j)}\|_F^2. \tag{22}$$

Consequently,

$$\sqrt{p_i(z)p_j(z)} = \exp\left(-\frac{\|x_0^{(i)} - x_0^{(j)}\|_F^2}{8\sigma_{t_i}^2}\right) q_{t_i}(z \mid m_{ij}). \tag{23}$$

Since $q_{t_i}(z \mid m_{ij})$ integrates to one, we obtain

$$\mathbb{E}_{x_{t_i}^{(i)} \sim q(x_{t_i}^{(i)} | x_0^{(i)})}[\pi_{ij}] \leq \frac{1}{2} \exp\left(-\frac{\|x_0^{(i)} - x_0^{(j)}\|_F^2}{8\sigma_{t_i}^2}\right). \tag{24}$$

□

Likelihood-based weighting over reference samples has been used to construct diffusion training targets (Xu et al., 2023; Zhang et al., 2024). For NATR, we examine how a candidate's normalized selection probability constrains its forward likelihood relative to the source sample.

Proposition A.1 (Likelihood compatibility). Suppose that $\mathcal{M}$ contains the original sample indexed by i and a candidate indexed by $j \neq i$. Then

$$q_{t_i}(\widetilde{x}_{t_i}^{(i)} \mid x_0^{(j)}) \geq \frac{\pi_{ij}}{1 - \pi_{ij}} q_{t_i}(\widetilde{x}_{t_i}^{(i)} \mid x_0^{(i)}). \tag{25}$$

Proof. Let $q_k = q_{t_i}(\widetilde{x}_{t_i}^{(i)} \mid x_0^{(k)})$. All Gaussian densities are strictly positive, and the reference pool contains both i and j. Therefore,

$$\pi_{ij} = \frac{q_j}{\sum_{k \in \mathcal{M}} q_k} \leq \frac{q_j}{q_i + q_j}. \tag{26}$$

Rearranging gives $\pi_{ij} q_i \leq (1 - \pi_{ij}) q_j$. Because $0 < \pi_{ij} < 1$, division by $1 - \pi_{ij}$ proves the claim. □

Proposition A.2 (Geometric compatibility). Under the assumptions of Proposition A.1, let $d_i = \|\widetilde{x}_{t_i}^{(i)} - x_0^{(i)}\|_F^2$. Then

$$\|x_0^{(j)} - x_0^{(i)}\|_F \leq \sqrt{d_i} + \sqrt{d_i - 2\sigma_{t_i}^2 \log \frac{\pi_{ij}}{1 - \pi_{ij}}}. \tag{27}$$

The quantity inside the second square root is nonnegative.

Proof. Let $d_j = \|\widetilde{x}_{t_i}^{(i)} - x_0^{(j)}\|_F^2$. Substituting the Gaussian densities into Eq. (25) and taking logarithms yields

$$d_j \leq d_i - 2\sigma_{t_i}^2 \log \frac{\pi_{ij}}{1 - \pi_{ij}}. \tag{28}$$

The right-hand side is nonnegative because it upper-bounds $d_j \geq 0$. By the triangle inequality and the monotonicity of the square root,

$$\begin{aligned} \|x_0^{(j)} - x_0^{(i)}\|_F &\leq \|x_0^{(i)} - \widetilde{x}_{t_i}^{(i)}\|_F + \|\widetilde{x}_{t_i}^{(i)} - x_0^{(j)}\|_F \\ &= \sqrt{d_i} + \sqrt{d_j} \\ &\leq \sqrt{d_i} + \sqrt{d_i - 2\sigma_{t_i}^2 \log \frac{\pi_{ij}}{1 - \pi_{ij}}}. \end{aligned} \tag{29}$$

This proves the stated bound. □

Interpretation. For fixed d_i and σ_t, the upper bound in Eq. (27) decreases with π_{ij}. Higher matching probabilities therefore yield tighter candidate-wise compatibility bounds.

These inequalities hold for all candidates and, in particular, for accepted replacements. However, NATR samples from the full categorical distribution; it does not impose a minimum selection probability. Consequently, the results do not provide a uniformly small displacement bound for every accepted target. They also do not establish fault-semantic equivalence: two geometrically similar trajectories may differ in diagnostically important components.

A.2.2 Class-Frequency Constraint Target Allocation.

We next examine the effect of the class-frequency constraint. Let (X, Y) denote an original training sample, (X_J, Y_J) its sampled reference candidate, and A the replacement indicator. Define the effective target label as

$$Y_{\text{target}} = \begin{cases} Y_J, & A = 1, \\ Y, & A = 0. \end{cases} \tag{30}$$

The transfer rule implies the pointwise relation

$$r(Y_{\text{target}}) \le r(Y). \tag{31}$$

For a threshold u, define the frequency-based tail set

$$\mathcal{C}_{\text{tail}}(u) = \{c \in \mathcal{C} : r(c) \le u\}. \tag{32}$$

Proposition A.3 (Non-decreasing tail-target allocation). For every threshold u,

$$\mathbb{P}\big(Y_{\text{target}} \in \mathcal{C}_{\text{tail}}(u)\big) - \mathbb{P}\big(Y \in \mathcal{C}_{\text{tail}}(u)\big) = \mathbb{P}\big(A = 1,\ Y \notin \mathcal{C}_{\text{tail}}(u),\ Y_J \in \mathcal{C}_{\text{tail}}(u)\big) \ge 0. \tag{33}$$

Proof. If $Y \in \mathcal{C}_{\text{tail}}(u)$, then $r(Y) \le u$. When $A = 0$, the target label is unchanged. When $A = 1$, the acceptance rule implies $r(Y_J) \le r(Y) \le u$. Hence an original tail sample cannot be assigned a target outside the tail set.

If $Y \notin \mathcal{C}_{\text{tail}}(u)$, the target enters the tail set precisely when $A = 1$ and $Y_J \in \mathcal{C}_{\text{tail}}(u)$. Therefore,

$$\mathbf{1}\{Y_{\text{target}} \in \mathcal{C}_{\text{tail}}(u)\} - \mathbf{1}\{Y \in \mathcal{C}_{\text{tail}}(u)\} = \mathbf{1}\{A = 1,\ Y \notin \mathcal{C}_{\text{tail}}(u),\ Y_J \in \mathcal{C}_{\text{tail}}(u)\}. \tag{34}$$

□

Interpretation. NATR cannot decrease the aggregate allocation of reconstruction targets to a tail set defined by a frequency threshold. The increase is strict if accepted transfers from outside that set into it occur with positive probability.

This is an aggregate statement. An individual tail class may lose target assignments to a rarer class, and the result does not imply a balanced target distribution. This statement concerns the labels of reconstruction targets; the denoiser itself receives no class label.

Because the argument is pointwise, the same non-decreasing allocation property holds when conditioning on a diffusion timestep or when counting target assignments with any common nonnegative per-instance weight. Such weighted counts should not be interpreted as gradient magnitudes or independent information gain.

A.3 Reconstruction Bias

Target replacement changes the reconstruction objective of the unconditional denoiser while retaining the original noisy input. We characterize this change under squared-error temporal reconstruction. Here, "unconditional" means that no class label is provided; the predictor still depends on the noisy observation and timestep.

Let S denote the random diffusion timestep and $Z = X + \sigma_S \epsilon$ the normalized noisy observation. Define

$$H = (Z, S), \qquad \Delta = A(X_J - X), \qquad X_{\text{target}} = X + \Delta. \tag{35}$$

We write $f(H) = f(Z, S)$ for a predictor in normalized coordinates. For the neural denoiser, this corresponds to $f(Z, S) = f_\theta(\sqrt{\bar{\alpha}_S}Z, S, \varnothing)$. Although Δ depends on labels through target selection, the predictor does not observe those labels. The conditional expectations below therefore average over labels and reference-pool randomness given (Z, S).

Assume that X and X_{target} have finite second moments. All expectations include training sample selection, diffusion noise, reference-pool sampling, and candidate selection. No independence between these quantities is required beyond the specified forward process.

Consider the population risks

$$\mathcal{R}_0(f) = \mathbb{E}\big[\|f(H) - X\|_F^2\big], \tag{36}$$

$$\mathcal{R}_{\text{NATR}}(f) = \mathbb{E}\big[\|f(H) - X_{\text{target}}\|_F^2\big]. \tag{37}$$

Over unrestricted square-integrable predictors, their minimizers are

$$f_0^*(H) = \mathbb{E}[X \mid H], \tag{38}$$

$$f_{\text{NATR}}^*(H) = \mathbb{E}[X_{\text{target}} \mid H]. \tag{39}$$

These identities hold almost surely. A strictly positive timestep weight $w(S)$ does not change the pointwise minimizers of these temporal risks. At timesteps with zero weight, the weighted risk does not determine a minimizer.

Proposition A.4 (Reconstruction bias). Define

$$b(H) = f_{\text{NATR}}^*(H) - f_0^*(H), \qquad p_A(H) = \mathbb{P}(A = 1 \mid H). \tag{40}$$

On inputs for which $p_A(H) > 0$, let

$$D_A(H) = \mathbb{E}\big[\|X_J - X\|_F^2 \mid H, A = 1\big]. \tag{41}$$

Set $D_A(H) = 0$ when $p_A(H) = 0$. Then

$$b(H) = \mathbb{E}[\Delta \mid H] \tag{42}$$

and

$$\|b(H)\|_F^2 \le p_A(H)^2 D_A(H) \le p_A(H) D_A(H) = \mathbb{E}\big[A\|X_J - X\|_F^2 \mid H\big]. \tag{43}$$

In particular, $b(H) = 0$ whenever $p_A(H) = 0$.

Proof. By linearity of conditional expectation,

$$f_{\text{NATR}}^*(H) = \mathbb{E}[X + \Delta \mid H] = f_0^*(H) + \mathbb{E}[\Delta \mid H], \tag{44}$$

which proves Eq. (42). When $p_A(H) > 0$,

$$\mathbb{E}[\Delta \mid H] = p_A(H)\,\mathbb{E}[X_J - X \mid H, A = 1]. \tag{45}$$

Conditional Jensen's inequality gives

$$\|b(H)\|_F^2 = p_A(H)^2\,\|\mathbb{E}[X_J - X \mid H, A = 1]\|_F^2 \le p_A(H)^2 \mathbb{E}\big[\|X_J - X\|_F^2 \mid H, A = 1\big]. \tag{46}$$

The remaining inequality follows from $0 \le p_A(H) \le 1$, and the final equality follows by conditioning on A.

If $p_A(H) = 0$, then $\Delta = 0$ conditionally almost surely, so $b(H) = 0$. The stated convention for $D_A(H)$ makes the bound valid in this case as well. □

Interpretation. The reconstruction bias of the unconditional predictor depends on both the accepted transfer probability and the displacement of the accepted targets. The refined bound additionally shows that displacement directions may cancel: positive transfer probability and nonzero transfer distances do not necessarily imply nonzero bias.

Conversely, likelihood compatibility and increased tail-target allocation do not imply that this bias is small. NATR preserves the original no-class-condition regression function $\mathbb{E}[X \mid Z, S]$ exactly when $\mathbb{E}[\Delta \mid H] = 0$ almost surely, a condition not guaranteed by its acceptance rule.

Corollary A.1 (Timestep-wise expected bias). For a timestep t with $\mathbb{P}(S = t) > 0$,

$$\mathbb{E}[\|b(H)\|_F^2 \mid S = t] \le \mathbb{E}[p_A(H)^2 D_A(H) \mid S = t] \le \mathbb{E}[A\|X_J - X\|_F^2 \mid S = t]. \tag{47}$$

Proof. Take conditional expectations in Eq. (43) given $S = t$. Since S is a component of H, the tower property gives the final inequality. □

Corollary A.2 (Excess original-target reconstruction risk). Define the original-target risk at timestep t by

$$\mathcal{R}_{0,t}(f) = \mathbb{E}[\|f(Z, t) - X\|_F^2 \mid S = t]. \tag{48}$$

Then

$$0 \le \mathcal{R}_{0,t}(f^*_{\mathrm{NATR}}) - \mathcal{R}_{0,t}(f^*_0) = \mathbb{E}[\|b(H)\|_F^2 \mid S = t] \le \mathbb{E}[A\|X_J - X\|_F^2 \mid S = t]. \tag{49}$$

The corresponding aggregate identity and bound are

$$0 \le \mathcal{R}_0(f^*_{\mathrm{NATR}}) - \mathcal{R}_0(f^*_0) = \mathbb{E}[\|b(H)\|_F^2] \le \mathbb{E}[A\|X_J - X\|_F^2]. \tag{50}$$

Proof. For any square-integrable predictor f, expand

$$\begin{aligned}\|f(H) - X\|_F^2 = \|f(H) - f^*_0(H)\|_F^2 + \|f^*_0(H) - X\|_F^2 \\ + 2\langle f(H) - f^*_0(H), f^*_0(H) - X\rangle_F.\end{aligned} \tag{51}$$

The cross term has zero conditional expectation given H, since $\mathbb{E}[f^*_0(H) - X \mid H] = 0$. Because S is measurable with respect to H, taking expectations given $S = t$ yields

$$\mathcal{R}_{0,t}(f) - \mathcal{R}_{0,t}(f^*_0) = \mathbb{E}[\|f(H) - f^*_0(H)\|_F^2 \mid S = t]. \tag{52}$$

Set $f = f^*_{\mathrm{NATR}}$ and apply Corollary A.1. Taking unconditional expectations proves the aggregate statement. □

Limits of class-wise interpretation. Labels may be used to stratify empirical statistics, but Y is not an input to the predictor. In general, $\mathbb{E}[f^*_0(H) - X \mid H, Y] \neq 0$, so the risk decomposition above cannot be asserted separately within each class without additional assumptions. Likewise, $\mathbb{E}[A\|X_J - X\|_F^2 \mid H]$ cannot in general be replaced by a class-specific displacement moment after averaging within a class. The guarantees stated here apply to the aggregate distribution and to timestep groups, not to each class separately.

Connection to candidate compatibility. For an accepted candidate, define

$$G_i(j, t) = \sqrt{d_i} + \sqrt{d_i - 2\sigma_t^2 \log \frac{\pi_{ij}}{1 - \pi_{ij}}}. \tag{53}$$

Proposition A.2 gives $\|x_0^{(j)} - x_0^{(i)}\|_F^2 \le G_i(j, t)^2$. Hence the displacement term in Eq. (47) can also be bounded by the corresponding expectation of $A\, G_i(J_i, t)^2$, with the product defined as zero when $A = 0$.

This connects the geometric analysis to the reconstruction-bias bound for the unconditional predictor. It does not establish that either bound is numerically small: that depends on the realized noise levels, candidate probabilities, and transfer distances.

B Additional Experimental Details and Results

This section supplements the experimental setup and results presented in the main text. We provide dataset resources and evaluation details, followed by an analysis of sensitivity to fine-tuning epochs and distributional visualizations of generated time series.

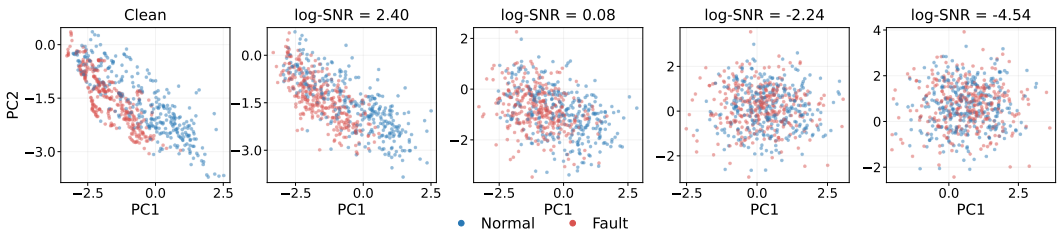


Figure 5: **Fixed PCA projections of normal and anomalous time-series windows at different noise levels.** Blue and red points denote normal and anomalous windows, respectively. PCA is fitted using clean training windows only. All panels use the same selected test windows and paired noise realizations. Lower log-SNR indicates stronger noise. Axis limits are adjusted independently across panels.

B.1 Dataset Resources

The dataset characteristics and operating-state label definitions are described in the main text. For reference, the public dataset resources are listed below:

- **ETTh1 and ETTm1:** `https://gitee.com/quark/ETTDataset/tree/main/ETT-small`.
- **SWaT:** `https://itrust.sutd.edu.sg/itrust-labs_datasets/dataset_info/#swat`.
- **SKAB:** `https://www.kaggle.com/dsv/1693952`.
- **WTSD:** `https://www.kaggle.com/datasets/berkerisen/wind-turbine-scada-dataset`.

B.2 Effect of Noise on Class Discriminability

We examine whether increasing diffusion noise reduces the discriminability of normal and anomalous time-series windows. This experiment assesses a premise underlying NATR, without assuming that shared operating information necessarily survives the noising process.

Experimental setup. We use the original dataset preprocessing and label definitions, with a training fraction of 0.8. Noisy observations are constructed using the model's forward diffusion schedule: $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I)$. We express the noise level as

$$\mathrm{logSNR}(t) = \ln\left(\frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}\right), \tag{54}$$

where lower values indicate stronger noise.

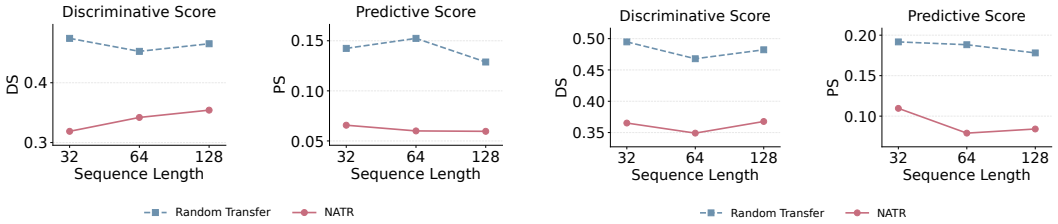
For visualization, we flatten each time-series window and fit PCA using clean training windows only. The resulting projection is fixed across all noise levels. We select equal numbers of normal and anomalous test windows and reuse the same windows and their associated Gaussian noise realizations across panels. Figure 5 shows the clean inputs and four representative noise levels. Axis limits are adjusted independently in each panel to improve visibility; apparent cloud sizes should therefore not be compared directly across panels.

To complement the two-dimensional visualization, we train a fixed-capacity lightweight 1D-CNN classification probe from scratch at each noise level, using the same training budget. Each probe receives the full noisy sequence, without PCA projection or a class condition. We use three training seeds and three test-noise realizations per seed. Metrics are first averaged over noise realizations within each seed, and then summarized across seeds.

Results. The clean and weakly perturbed samples retain visible class-dependent differences in the projected distributions. As noise increases, the elongated structure becomes less distinct and the two classes increasingly overlap. The classification results in Table 4 are consistent with this observation.

Table 4: **Classification-probe performance at representative noise levels.** Values are means ± standard deviations across three training seeds, after averaging test-noise realizations within each seed.

Input	AUROC	Balanced accuracy
Clean	0.9274 ± 0.0020	0.8576 ± 0.0064
$\mathrm{logSNR} = 2.40$	0.9024 ± 0.0035	0.8220 ± 0.0023
$\mathrm{logSNR} = 0.08$	0.7539 ± 0.0042	0.6802 ± 0.0057
$\mathrm{logSNR} = -2.24$	0.5894 ± 0.0025	0.5585 ± 0.0030
$\mathrm{logSNR} = -4.54$	0.5269 ± 0.0014	0.5165 ± 0.0019



(a) ETTh1 (b) ETTm1

Figure 6: **Comparison of NATR and random target transfer on ETTh1 and ETTm at sequence lengths of 32, 64, and 128.** Lower Discriminative Score (DS) and Predictive Score (PS) indicate better performance.

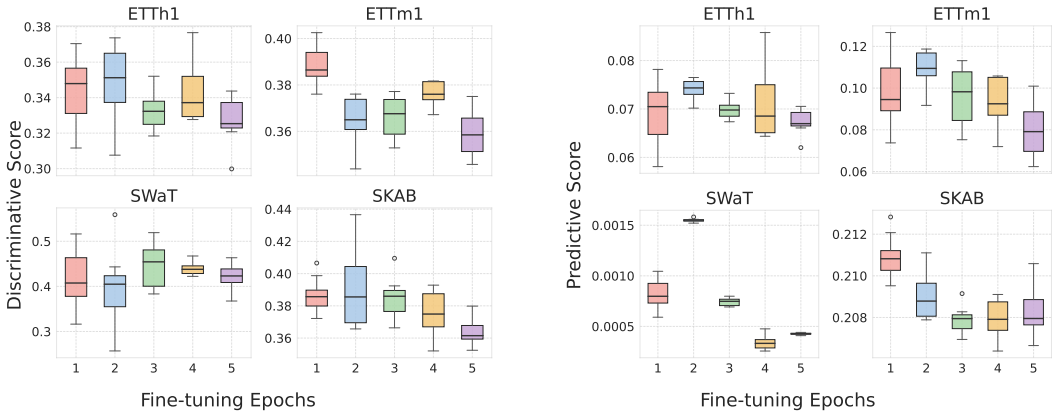
At $\mathrm{logSNR} = 0.08$, AUROC decreases from 0.9274 on clean inputs to 0.7539, indicating reduced discriminability while some class information remains recoverable. At $\mathrm{logSNR} = -4.54$, both metrics approach the binary chance level of 0.5. The probe therefore has limited ability to distinguish the two classes at this noise level.

B.3 Comparison with Random Target Transfer

Experimental setup. To evaluate the contribution of noise-compatibility-based target selection, we compare NATR with a random target transfer baseline on the ETTm1 dataset using sequence lengths of 32, 64, and 128. The baseline samples candidate targets uniformly from the current mini-batch, replacing NATR's compatibility-based sampling while retaining the same class-frequency and timestep constraints. Both methods use the same pretrained checkpoint, model configuration, training budget, and evaluation protocol. We assess generation quality using the Discriminative Score (DS) and Predictive Score (PS), with lower values indicating better performance for both metrics. This comparison evaluates the candidate-selection mechanism as a whole; the number of accepted transfers and the resulting target-class distributions are not explicitly matched.

Results and analysis. As shown in Figure 6, NATR achieves lower DS and PS than random target transfer at all three sequence lengths, demonstrating a consistent advantage across the evaluated window sizes. The lower DS indicates that generated samples are less distinguishable from real samples under the evaluation protocol, while the lower PS suggests better preservation of information useful for temporal prediction. Together, these results support the effectiveness of compatibility-guided target transfer over uniform random selection.

A plausible explanation is that uniform selection ignores the relationship between a candidate target and the current noisy observation. At lower noise levels, the observation retains substantial information about its source trajectory. Randomly replacing its reconstruction target can therefore introduce incompatible supervision and interfere with learning temporal dynamics. NATR instead assigns candidate probabilities according to their likelihood under the current noise level. When the observation remains informative, this mechanism tends to favor the original target or compatible alternatives. As noise increases and the observation becomes less informative about its source, a broader set of trajectories can receive appreciable matching probability. Combined with the class-frequency constraint,



(a) Discriminative Score.

(b) Predictive Score.

Figure 7: **Discriminative and Predictive scores under different numbers of fine-tuning epochs.** Lower values indicate better performance for both metrics.

this provides a noise-adaptive way to reallocate supervision toward rarer classes while potentially limiting interference from incompatible targets.

The observed improvements are consistent with this explanation and support the effectiveness of NATR's overall transfer strategy. However, because the two methods may differ in their accepted transfer counts, timestep distributions, and target-class distributions, this comparison does not isolate the benefit of selecting a particular target sample under a fixed transfer budget. The results therefore establish an advantage over the evaluated random-transfer baseline, rather than directly proving that shared operating structure is preserved during transfer.

B.4 Sensitivity to Fine-Tuning Epochs

We examine how the number of fine-tuning epochs affects generation quality. Fig. 7 reports the Discriminative and Predictive scores across the evaluated datasets for different fine-tuning durations.

The Discriminative scores generally improve or remain comparable as fine-tuning proceeds on ETTh1, ETTm1, and SKAB, whereas SWaT exhibits a less consistent trend. The Predictive scores also vary across epoch settings, and additional fine-tuning does not uniformly improve performance across datasets. These results indicate that the effect of fine-tuning duration depends on both the dataset and the evaluation metric.