ADVMISSING-TSC: ADVERSARIAL MISSINGNESS AND MASK-AWARE DEFENSE FOR TIME SERIES CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Time series classifiers can perform well under random missingness yet misclassify inputs when discriminative observations are lost. We propose AdvMissing-TSC to evaluate this risk by masking a fixed fraction of observations in each sample while leaving all retained values unchanged. Its Target-Guided Sparse Missingness Attack (TG-SMA) combines Gumbel relaxation with Top-k projection to optimize masks from classification feedback, either reducing accuracy or approaching a specified target accuracy. Temporal regularization encourages contiguous missing blocks. Using masks optimized under mean or zero imputation, we examine how missingness structure affects the missingness–imputation–classification process and compare imputers under fixed masks. We also propose Mask-aware single-Imputation consistency Defense (MID), which learns to gate value and missingness features and strengthens prediction consistency between complete and imputed views at higher missing rates. In the fixed-budget paired comparisons, adversarial missingness produces lower accuracy than corresponding random missingness. Among the four evaluated missingness–imputation combinations, MID restores accuracy for the most degraded combination to a level comparable to or slightly above that of the least degraded undefended combination.

1 INTRODUCTION

Which observations can a time series classifier afford to lose? The missing rate specifies how many observations are unavailable, but does not identify them. The original time series classification bake-off assumed complete inputs, and the 2018 UEA multivariate archive excluded incomplete sequences (Bagnall et al., 2017; 2018). The 2024 bake-off also removes incomplete cases from four datasets (Middlehurst et al., 2024). These evaluations do not establish how classifiers behave when discriminative observations are selectively removed. We study which missing positions most disrupt classification under a fixed budget and how to assess and mitigate that risk.

Previous time series attacks use distilled surrogate models or additive perturbations to change observed values (Karim et al., 2021; Rathore et al., 2020). Our attacker selects time–channel entries to hide while preserving every retained value (Figure 1). Classification feedback guides this selection. At a fixed missingness budget, this reveals sensitivity to particular information losses beyond the average behavior measured under random missingness. The same removal count can be spread across isolated entries or concentrated in contiguous segments. These patterns leave different temporal evidence available to a classifier. Comparing adversarial masks with random masks of the same structure separates the effect of selecting locations from the effect of imposing that structure.

The risk also depends on imputation. An imputed input combines observed values with estimates at the removed positions. Numerical reconstruction and classification have different objectives: recovering values with low reconstruction error does not by itself establish that the imputed sequence preserves the original decision. We therefore evaluate the missingness–imputation–classification process directly. Holding generated masks fixed gives every imputer the same retained information and prevents missing positions from confounding their comparison. This evaluation asks how each imputer supports classification under the selected masks; its scope is conditional on those masks.

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.

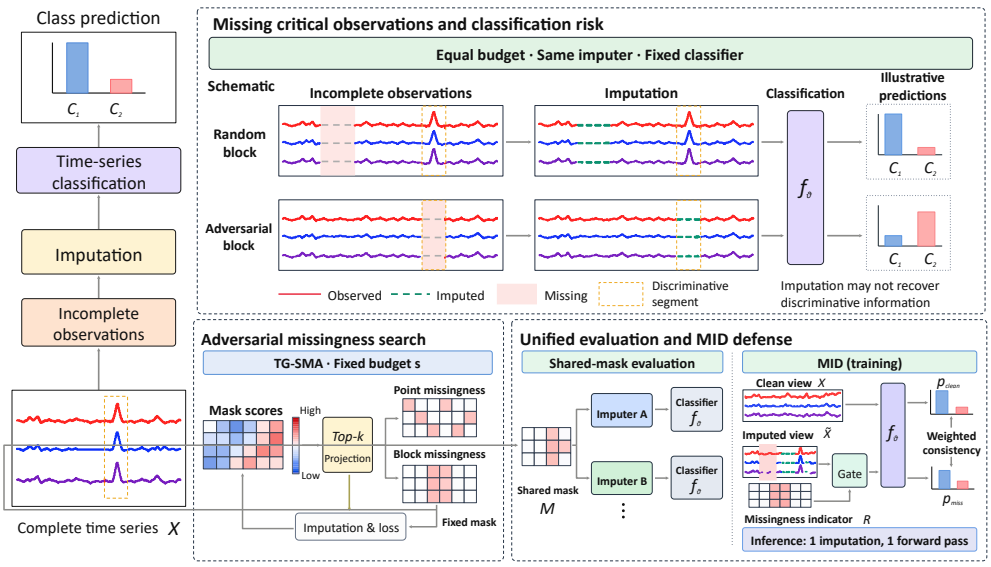


Figure 1: Selective missingness in the imputation–classification pipeline. Missing locations affect the information available for classification after imputation. The lower panels motivate task-level evaluation and summarize the attack, fixed-mask comparison, and MID.

For defense, the imputed values alone do not explicitly identify which entries were estimated. MID supplies this information through the missingness mask and learns how to combine mask features with value features. During training, complete and incomplete views of the same sample share a classifier, with a consistency term constraining changes in their predictions. Inference uses the imputed sequence and its mask, without requiring the complete view. AdvMissing-TSC connects this mask-aware defense with budget-constrained attacks and evaluation under fixed masks (Figure 2).

Our contributions are as follows:

We formulate an adversarial missingness threat model for time series classification, using optimizable binary masks to control per-sample removal counts while preserving retained values.

We propose TG-SMA and the zeroth-order search strategy NES-Mask. They use gradient and natural-evolution-strategy query feedback, respectively, to optimize missing positions for accuracy degradation or a specified aggregate accuracy, without targeting an incorrect class or guaranteeing global optimality.

We propose MID, which gates value and mask features and trains for consistency between complete and incomplete views after a single imputation. In the Chinatown experiment with Informer, Adv-Block + Mean is the most degraded of four undefended combinations. With MID, it matches or exceeds the least degraded undefended combination, Adv-Point + CDRec, at the four missing rates from 0.10 to 0.40.

We evaluate robustness on 10 time series classification datasets, comparing adversarial point, adversarial block, and random block missingness across seven imputers and four classifiers to assess the joint effects of missingness structure and imputation choice.

2 Related Work

Time series attacks use adversarial transformation networks or additive noise to modify observed values (Karim et al., 2021; Rathore et al., 2020). Recent attacks localize value perturbations through tree search (Ding et al., 2023) or gradient-selected intervals with a trend-adaptive loss (Kim et al., 2025). Our attack variable is observation availability. For nondifferentiable or black-box classifiers, NES-Mask adapts the query-based approach of Ilyas et al. (2018): paired positive and negative Gaussian perturbations and their query feedback estimate a search direction for mask generation.

Bagnall et al. (2017) compared representations based on full series, intervals, and shapelets using complete, equal-length univariate series. ROCKET extracts random convolutional features for a linear classifier (Dempster et al., 2020). TimesNet converts periodic variation into two-dimensional representations and supports classification (Wu et al., 2023). Informer uses efficient attention for long-sequence forecasting (Zhou et al., 2021); we adapt it for classification. We compare these representations' sensitivity to missing time segments and channels under equal budgets.

Imputation methods recover unavailable observations through approaches including low-rank completion and GAIN's adversarial estimation (Mazumder et al., 2010; Yoon et al., 2018). Reconstruction quality, however, does not always track downstream classification performance. Shadbahr et al. (2023) report this discrepancy on clinical and simulated data, showing why reconstruction error alone may not capture an imputer's effect on classification. Task-oriented imputation evaluation also estimates downstream effects on forecasting (Wang et al., 2024). Learning-based methods also incorporate missingness information and task requirements: BRITS jointly trains imputation and supervised prediction (Cao et al., 2018), and SAITS uses attention and missingness information to combine temporal representations (Du et al., 2023). Recent work jointly learns temporal imputation and noise-robust features for classifying incomplete time series (Liu et al., 2025). MID addresses discriminative stability under adversarial missingness. Following a single imputation, it gates value and mask features and constrains prediction consistency between complete and incomplete views of the same sample to reduce decision shifts caused by missing observations. We correspondingly compare imputers under fixed masks and evaluate whether their outputs retain useful discriminative information.

3 PRELIMINARIES

Let $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^n$, where $X_i \in \mathbb{R}^{T\times C}$ has C channels and $y_i \in \{1, \dots, K\}$. A frozen classifier f_ω maps complete inputs to class probabilities. We use an **observation mask** $M_i \in \{0,1\}^{T\times C}$, with one denoting an observed entry and zero a missing entry, and write $R_i = 1 - M_i$. An imputer receives both masked values and the observation mask:

$$\hat{X}_i = I(X_i \odot M_i, M_i), \qquad \hat{y}_i = \arg\max_c f_\omega(\hat{X}_i)_c. \tag{1}$$

The mask distinguishes missing entries from observed zeros, and imputation preserves observed values. The attacker selects unavailable time–channel entries without altering retained values. For a nominal missing rate s, each sample has **budget**

$$k = \lfloor sTC \rfloor, \qquad \mathcal{M}_k = \{M \in \{0,1\}^{T\times C} : \|1 - M\|_0 = k\}. \tag{2}$$

The realized rate is $k/(TC) \le s$. This budget counts entries, not entire time steps or channels; soft masks used during optimization need not belong to $\mathcal{M}_k$. For a labeled batch $\mathcal{B}$ of size B and masks $\mathcal{M} = \{M_i\}_{i\in\mathcal{B}}$, define

$$A_{\mathcal{B}}(\mathcal{M}; I, f_\omega) = \frac{1}{B}\sum_{i\in\mathcal{B}} \mathbf{1}\Big\{\arg\max_c f_\omega\big(I(X_i \odot M_i, M_i)\big)_c = y_i\Big\}. \tag{3}$$

We minimize either $A_{\mathcal{B}}$ or $|A_{\mathcal{B}} - a^\star|$, subject to $M_i \in \mathcal{M}_k$ for every sample. Here $a^\star = a_{\text{clean}}(1-\delta)$ specifies a relative accuracy reduction δ, rather than a target class. Budgets apply per sample, while accuracy feedback couples the samples within a batch. Neither search certifies a globally worst mask.

TG-SMA uses classifier gradients. NES-Mask queries accuracy without classifier or imputer gradients. Both use ground-truth evaluation labels. Search uses a Mean/Zero proxy imputer I_{search}; subsequent comparisons hold the masks fixed across evaluation imputers I_{eval}. Attacks are not reoptimized for each imputer. Whether masks are shared or regenerated across classifiers is unspecified.

4 ADVMISSING-TSC

Figure 2 connects mask search against a frozen classifier, evaluation with fixed masks, and MID training.

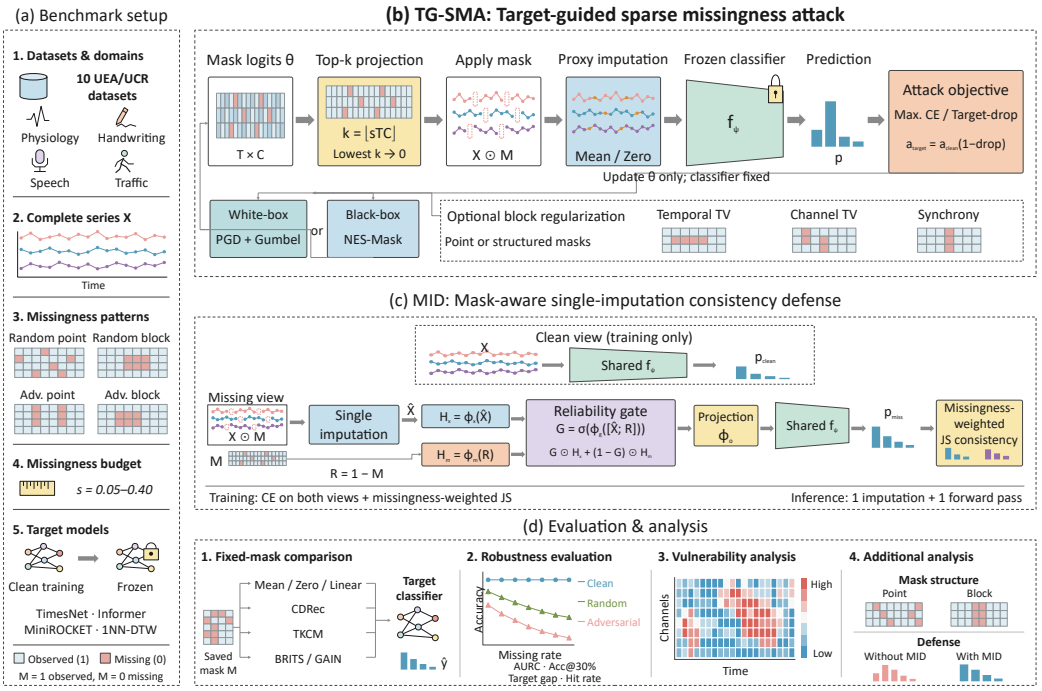


Figure 2: AdvMissing-TSC: (a) benchmark setting; (b) gradient or query mask search and Top-k projection; (c) MID; (d) fixed-mask evaluation. C denotes the number of channels. Sections 4.1–4.3 define the modules; Appendix B.4 gives the complete- and missing-view training settings.

4.1 Target-Guided Sparse Missingness Attack

We introduce continuous parameters $\Theta_i \in \mathbb{R}^{T\times C}$ and use the binary Gumbel-Softmax relaxation (Jang et al., 2017):

$$P_i = \mathrm{sigmoid}\left(\frac{\Theta_i + G_i^1 - G_i^0}{\tau}\right), \qquad G_i^0, G_i^1 \overset{\text{i.i.d.}}{\sim} \mathrm{Gumbel}(0,1). \tag{4}$$

The temperature $\tau > 0$ decreases during optimization, moving soft observation masks toward binary states while allowing classification feedback to reach the mask parameters. After each update, a **per-sample Top-k projection** enforces the budget:

$$M_i = \Pi_k(P_i), \qquad \Pi_k(P_i) \in \arg\min_{M\in\mathcal{M}_k} \|M - P_i\|_F^2. \tag{5}$$

The k smallest entries of P_i become missing. Appendix A.1 proves optimality. Soft masks provide gradients; hard masks enforce the budget, score candidates, and define the output. Sorting is excluded from backpropagation.

A differentiable proxy imputer I_0 maps soft-masked inputs to classifier inputs. Define the **classification surrogate** and **soft budget penalty** as

$$\begin{aligned} \mathcal{L}_{\mathrm{cls}} &= \tfrac{1}{B}\textstyle\sum_{i=1}^{B} \ell_{\mathrm{CE}}\big(f_\omega(I_0(X_i \odot P_i, P_i)), y_i\big), \\ \mathcal{L}_{\mathrm{sp}} &= \big[\tfrac{1}{BTC}\textstyle\sum_{i,t,c}(1 - P_{i,t,c}) - s\big]_+, \end{aligned} \tag{6}$$

where $[z]_+ = \max(z, 0)$. The penalty limits excessive masking during relaxation; Top-k projection supplies the exact per-sample constraint. Given hard-mask batch accuracy a_r at iteration r, set

$$d_r = \begin{cases} -1, & \text{worst-case mode,} \\ -1, & \text{target-guided mode and } a_r > a^\star, \\ +1, & \text{target-guided mode and } a_r \le a^\star. \end{cases} \tag{7}$$

Worst-case mode maximizes cross-entropy as a surrogate for reducing accuracy. **Target-guided mode** strengthens the attack above the target and reverses direction at or below it. Accuracy controls direction and candidate selection; gradients come from cross-entropy. A prescribed target need not be attainable under a given budget.

To encourage **contiguous missingness**, we optimize

$$\mathcal{L}_{\mathrm{TG}} = d_r \mathcal{L}_{\mathrm{cls}} + \lambda_s \mathcal{L}_{\mathrm{sp}} + \lambda_t \,\mathrm{TV}_t(P) + \lambda_c \,\mathrm{TV}_c(P) + \lambda_{\mathrm{sync}} V_c(P), \tag{8}$$

with temporal variation, channel variation, and channel synchronization terms

$$\begin{aligned} \mathrm{TV}_t(P) &= \frac{1}{B(T-1)C} \sum_{i=1}^{B} \sum_{t=1}^{T-1} \sum_{c=1}^{C} |P_{i,t+1,c} - P_{i,t,c}|, \\ \mathrm{TV}_c(P) &= \frac{1}{BT(C-1)} \sum_{i=1}^{B} \sum_{t=1}^{T} \sum_{c=1}^{C-1} |P_{i,t,c+1} - P_{i,t,c}|, \\ V_c(P) &= \frac{1}{BT} \sum_{i=1}^{B} \sum_{t=1}^{T} \mathrm{Var}_c(P_{i,t,c}). \end{aligned} \tag{9}$$

Temporal TV discourages changes between adjacent times; channel TV encourages similar states in adjacent channels; synchronization reduces differences across channels at the same time. Channel terms are zero for univariate inputs. Disabling these regularizers yields adversarial point missingness. Enabling them encourages temporal continuity or synchronization across channels, without enforcing fixed-length segments or exact rectangles. For a binary missingness sequence $R_t = 1 - M_t$ in one channel, the number of transitions $V = \sum_{t=1}^{T-1} |R_{t+1} - R_t|$ and missing segments satisfy $n_{\mathrm{seg}} = (V + R_1 + R_T)/2$. Fewer transitions therefore consolidate missing entries into fewer intervals at a fixed budget.

Search alternates continuous updates and hard projection, retaining past candidates. It returns the feasible candidate with lowest accuracy in worst-case mode, or smallest $|a_r - a^\star|$ in target-guided mode. Unless stated otherwise, block attacks use $\lambda_t = 0.1$, $\lambda_c = 0.05$, $\lambda_{\mathrm{sync}} = 0.1$, and $\lambda_s = 1.0$. Temporal continuity is the primary structural constraint. Channel TV receives a smaller weight and is mainly applicable when channel adjacency is meaningful. Appendix G.4 examines score initialization.

4.2 Query-based mask search

NES-Mask uses natural evolution strategies when classifier gradients are unavailable or prediction contains nondifferentiable operations. Following gradient estimation in query-based attacks (Ilyas et al., 2018), it updates continuous retention scores from candidate classification feedback. Each sample uses the same Top-k projection as TG-SMA. For batch masks $\mathcal{M}$ with proxy-imputed accuracy $a(\mathcal{M})$, the query objective is

$$J(\mathcal{M}) = \begin{cases} a(\mathcal{M}), & \text{worst-case mode,} \\ |a(\mathcal{M}) - a^\star|, & \text{target-guided mode.} \end{cases} \tag{10}$$

Given the evaluation labels, predicted classes suffice; full class probabilities are unnecessary. Each candidate is imputed with I_0 before querying the classifier. At iteration r, draw q independent Gaussian directions $\epsilon_j \sim \mathcal{N}(0, I)$ and form antithetic candidates

$$\Theta^{\pm}_{r,j} = \Theta_r \pm \nu \epsilon_j, \qquad \mathcal{M}^{\pm}_{r,j} = \Pi_k(\Theta^{\pm}_{r,j}), \tag{11}$$

where $\nu > 0$ sets the perturbation scale and Π_k ranks retention scores. Projection ensures that every query obeys the same per-sample budget. Paired differences estimate a local search direction. We rank-normalize all candidate objectives within the iteration to reduce sensitivity to their numerical scale. Writing these values as $\widetilde{J}^{\pm}_{r,j}$, with larger values indicating higher cost, we update

$$\hat{g}_r = \frac{1}{2q\nu} \sum_{j=1}^{q} (\widetilde{J}^{+}_{r,j} - \widetilde{J}^{-}_{r,j}) \epsilon_j, \qquad \Theta_{r+1} = \Theta_r - \eta_{\mathrm{NES}} \hat{g}_r. \tag{12}$$

The search retains the feasible mask with smallest objective. Target-guided search may stop within a prescribed tolerance; otherwise it runs to the iteration limit. Each iteration evaluates $2q$ batches of candidate masks, so the number of direction pairs and iterations determine query cost.

Fixed masks isolate imputer differences from changes in missing locations. This measures conditional robustness, without adaptive optimization for each imputer–classifier pair.

4.3 MASK-AWARE SINGLE-IMPUTATION CONSISTENCY DEFENSE

An imputed sequence contains both observed and estimated values. A classifier that receives values alone cannot explicitly distinguish them, and imputed patterns can become unreliable evidence when discriminative regions are missing. MID uses missing-location information through **gated fusion** and constrains prediction changes through **consistency between complete and missing views**. Its objective is classification stability under missingness, without directly optimizing reconstruction of individual missing values.

For $\hat{X} = I(X \odot M, M)$ from a fixed imputer and $R = 1 - M$, $R_{t,c} = 1$ identifies an imputed entry. We project values and missingness indicators separately and compute a gate from their concatenation:

$$\begin{aligned} H_x &= \phi_x(\hat{X}), \qquad H_m = \phi_m(R), \qquad G = \mathrm{sigmoid}\big(\phi_g([\hat{X}; R])\big), \\ H_{\mathrm{MID}} &= G \odot H_x + (1 - G) \odot H_m, \qquad Z = \phi_o(H_{\mathrm{MID}}). \end{aligned} \tag{13}$$

Here $[\cdot;\cdot]$ concatenates features. The lightweight maps ϕ_x, ϕ_m, and ϕ_g produce equal-dimensional value features, mask features, and gates. The gate controls their relative contributions according to the missingness structure. The linear output projection ϕ_o maps the C-dimensional fused features at each time to the $2C$-dimensional input of the classification backbone. The backbone receives Z; concatenation is used to compute the gate.

During training, each complete sample has a corresponding missing-and-imputed view. Let F_θ include the input projections, gated fusion, output projection, and classification backbone. Both views share parameters:

$$p_i^{\mathrm{clean}} = F_\theta(X_i, 0), \qquad p_i^{\mathrm{miss}} = F_\theta(\hat{X}_i, R_i). \tag{14}$$

The complete view has a zero missingness indicator and supplies the consistency reference. Define

$$\rho_i = \frac{\|1 - M_i\|_1}{TC}, \qquad w_i = 1 + \alpha\rho_i, \qquad \alpha \geq 0. \tag{15}$$

The training objective is

$$\begin{aligned} \mathcal{L}_{\mathrm{MID}} = \frac{1}{B}\sum_{i=1}^{B}\big[&2\ell_{\mathrm{CE}}(p_i^{\mathrm{clean}}, y_i) + 1.5\ell_{\mathrm{CE}}(p_i^{\mathrm{miss}}, y_i) \\ &+ \lambda_{e,\mathcal{B}} w_i \,\mathrm{JS}\big(p_i^{\mathrm{clean}} \,\|\, p_i^{\mathrm{miss}}\big)\big], \end{aligned} \tag{16}$$

where JS is Jensen–Shannon divergence. Both views receive gradients from their classification losses and the consistency term. We use $\alpha = 0.3$; Appendix B.4 defines the consistency schedule $\lambda_{e,\mathcal{B}}$. Larger missing rates receive greater consistency weight; if all samples have the same rate, this weight uniformly rescales the consistency coefficient.

Gating acts on feature propagation, while consistency acts on predictions. With G and H_m locally fixed, value-feature perturbations satisfy

$$\|\Delta H_{\mathrm{MID}}\|_p \leq \|G\|_\infty \|\Delta H_x\|_p. \tag{17}$$

Smaller gate values attenuate direct propagation of the corresponding perturbations. This is a local fixed-gate result; input-dependent gates prevent it from constituting a global robustness guarantee. For $y^\star = \arg\max_j p_j^{\mathrm{clean}}$, let $\gamma = p_{y^\star}^{\mathrm{clean}} - \max_{j \neq y^\star} p_j^{\mathrm{clean}}$. If $\gamma > 0$ and, using natural logarithms,

$$\mathrm{JS}(p^{\mathrm{clean}} \| p^{\mathrm{miss}}) < \frac{\gamma^2}{8}, \tag{18}$$

the two views predict the same class. It guarantees class agreement, not correctness of the complete-view prediction. Inference uses one imputation and one forward pass through F_θ with the imputed sequence and missingness indicator. The complete view and consistency loss are used only during training. Testing requires neither the hidden true values nor repeated imputation or prediction ensembles.

Table 1: Classification accuracy at $s = 0.3$ with linear interpolation. Random-B denotes random time blocks. Bold values are the lowest accuracy within each dataset–classifier combination.

Dataset	Model	Complete	Random-P	Adv-P	Random-B	Adv-B
Heartbeat	TimesNet	0.872	0.829	0.764	0.793	**0.718**
	Informer	0.841	0.806	0.742	0.772	**0.704**
JapaneseVowels	TimesNet	0.978	0.958	0.916	0.932	**0.879**
	Informer	0.961	0.943	0.898	0.919	**0.861**
Handwriting	TimesNet	0.392	0.373	0.334	0.312	**0.242**
	Informer	0.367	0.346	0.301	0.287	**0.216**
Chinatown	TimesNet	0.983	0.974	0.953	0.933	**0.676**
	Informer	0.956	0.941	0.909	0.884	**0.758**
EthanolLevel	TimesNet	0.692	0.655	0.604	0.623	**0.581**
	Informer	0.642	0.616	0.571	0.590	**0.548**
UWaveGestureLibrary	TimesNet	0.904	0.867	0.824	0.808	**0.754**
	Informer	0.862	0.825	0.776	0.792	**0.736**

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

The experiments cover ten public time-series classification datasets, including Heartbeat, JapaneseVowels, Handwriting, and Chinatown, with univariate and multivariate sequences of varying lengths from different application domains. Variable-length sequences are padded to a common length, and existing missing values are interpolated to construct complete-input baselines.

We use TimesNet and Informer as differentiable classifiers and MiniROCKET and 1NN-DTW for black-box evaluation, with TG-SMA and NES-Mask, respectively. Classifiers are trained on complete training data and frozen during attacks. Mask search uses Mean or Zero as a surrogate imputer. We then fix the masks and compare seven imputers: Zero, Mean, Linear, CDRec, BRITS, GAIN, and TKCM.

The main attack experiment compares random point, random time-block, adversarial point, and adversarial block missingness on six datasets, using linear interpolation throughout evaluation. Missingness budgets are $s \in \{0.05, 0.10, 0.20, 0.30, 0.40\}$. We also test target relative accuracy drops of 20% and 30% on Heartbeat, Handwriting, and Chinatown.

We report complete-input accuracy, robust accuracy, relative accuracy drop, and area under the robustness curve (AURC). Target control is assessed by the absolute gap between attacked and target accuracy and the hit rate, with a hit defined as a gap of at most 0.03. MID is evaluated through accuracy gains and recovery relative to alternative missingness–imputation combinations.

5.2 ATTACK STRENGTH AT A FIXED BUDGET

Under a common budget and interpolation method (Table 1), adversarial points and blocks yield lower accuracy than their matching random baselines in all twelve dataset–classifier combinations. Equally weighted mean reductions are 4.51 and 8.10 percentage points, respectively. Adversarial blocks produce the lowest accuracy in every combination.

For TimesNet on Heartbeat, adversarial point missingness reduces accuracy from the random-point baseline of 0.829 to 0.764, a further 6.5 percentage points. Adversarial blocks reduce accuracy from 0.793 to 0.718, a further 7.5 points. On Chinatown, complete-input accuracy is 0.983, falling to 0.933 under random blocks and 0.676 under adversarial blocks. Position selection thus adds a 25.7-point loss. Missingness rate and random-removal performance alone do not fully characterize vulnerability to selective observation loss.

Comparisons across structures depend on the dataset and model. On Handwriting, random blocks yield lower accuracy than adversarial points for both TimesNet (0.312 versus 0.334) and Informer (0.287 versus 0.301). Optimizing block positions further reduces these accuracies to 0.242 and 0.216. Thus, the advantage over a matching random baseline does not imply that every adversarial

pattern is stronger than every random pattern. Both position and structure affect attack strength. Appendix G.1 summarizes relative drops across datasets.

5.3 Control of prescribed accuracy drops

We evaluate TG-SMA on Heartbeat, Handwriting, and Chinatown with TimesNet and Informer, relative drops $r \in \{0.2, 0.3\}$, and both point and block missingness at the five budgets above. Given complete-input accuracy a_{clean}, we define

$$a^\star = (1 - r)a_{\text{clean}}, \qquad \text{TargetGap} = |a_{\text{robust}} - a^\star| \,. \tag{19}$$

A gap of at most 0.03 counts as a hit. The target concerns aggregate accuracy, not a particular incorrect class.

Target gaps decrease with increasing budget in most combinations. Among the evaluated budgets, the minimum gap for adversarial points on Heartbeat with Informer is 0.0254 at $r = 0.2$; for adversarial blocks on Heartbeat with TimesNet, it is 0.0234 at $r = 0.3$. Both fall within tolerance. Handwriting with TimesNet has a mean gap of 0.0346 and a hit rate of 40%, among the better results. By comparison, the two Heartbeat models have mean gaps of approximately 0.1403 and 0.1463: isolated hits can coexist with substantial average error. Chinatown is harder to control, especially at $r = 0.3$, where most attacked accuracies remain above the target.

TG-SMA provides partial control, but arbitrary drops remain unguaranteed. Target reachability depends on dataset, classifier, and missingness structure; mean gaps and hit rates should accompany best-case results. Appendix F.1 and F.2 provide the full target-gap curves and representative accuracy curves.

5.4 Configuration sensitivity

An independent Informer experiment aggregates three imputation results (Table 6). On Heartbeat, Acc@30% is 0.717 under adversarial points and 0.746 under adversarial blocks, illustrating task-dependent sensitivity.

The fixed-mask comparison covers seven imputers (Table 7). For adversarial points on JapaneseVowels, Linear achieves 0.96 in both AURC and Acc@30%. Under adversarial blocks on Heartbeat, CDRec and TKCM both achieve Acc@30% of 0.80, exceeding Linear's 0.78. Recovery methods have different advantages under different conditions.

Table 9 compares four classifiers. On UWaveGestureLibrary under adversarial blocks, TimesNet has AURC 0.781, compared with MiniROCKET's 0.681. In this comparison, all four classifiers have lower adversarial-block accuracy than adversarial-point accuracy on Heartbeat. These results support considering missingness structure when choosing an imputer and classifier. Appendix G.2 and G.3 provide classifier radar plots and missing-position diagnostics.

5.5 Query-based attacks

We evaluate NES mask search on MiniROCKET with interpolation (Table 8). At $s = 0.30$, NES and random-block accuracies are $0.889/0.933$ on Chinatown, $0.154/0.194$ on Handwriting, and $0.884/0.938$ on JapaneseVowels. Query search finds more damaging masks on all three datasets. At $s = 0.40$, Handwriting accuracy falls to 0.108 under NES, compared with 0.168 under random blocks.

Heartbeat has the opposite ordering: random-block accuracy is 0.727 and 0.741 at $s = 0.30$ and 0.40, below the corresponding NES accuracies of 0.741 and 0.751. NES therefore outperforms the random-block baseline on three of four datasets under the query budget used.

A supplementary 1NN-DTW experiment on Chinatown gives accuracy 0.8805 under NES at $s = 0.30$, below random blocks (0.8921) and random points (0.9096). Mask search exposes distance-based classifier vulnerabilities without gradients. Appendix F.3 contains the query-attack curves.

Table 2: Chinatown–Informer accuracy (%) for four undefended configurations and MID applied to adversarial blocks with Mean imputation.

	Adversarial points		Adversarial blocks		MID
s	Mean	CDRec	Mean	CDRec	Blocks + Mean
0.05	92.0	93.5	86.0	88.0	92.0
0.10	86.0	86.5	82.0	84.0	87.0
0.20	80.0	80.5	75.0	77.0	83.0
0.30	76.0	76.5	70.0	72.0	81.0
0.40	72.0	75.5	65.0	67.0	76.0
AURC	79.9	80.9	74.4	76.4	82.9

5.6 MID UNDER AN UNFAVORABLE CONFIGURATION

On Chinatown with Informer (Table 2; Figure 3 in the appendix), adversarial blocks with Mean yield the lowest undefended accuracy at all five budgets; adversarial points with CDRec perform best. Applying MID to the former measures gains over its matching baseline and recovery toward the latter.

MID improves accuracy by 6, 5, 8, 11, and 11 percentage points at budgets 0.05, 0.10, 0.20, 0.30, and 0.40, respectively. At $s = 0.40$, accuracy rises from 65.0% to 76.0%, slightly above the 75.5% of adversarial points with CDRec. MID exceeds all four undefended configurations at budgets from 0.10 through 0.40, while retaining the block-missingness and Mean-imputation configuration.

Over $[0.05, 0.40]$, MID has AURC 82.9%, compared with 74.4% for its matching baseline and 80.9% for adversarial points with CDRec. Unrounded trapezoidal calculations give gains of 8.50 and 2.07 percentage points, respectively. Appendix E details the five-point gains and normalized integration; Appendix F.5 reports supplementary curves under other missingness settings.

6 LIMITATIONS

Missingness robustness depends on budget, position, and structure. Adversarial patterns outperform matching random baselines in the main experiment, but cross-structure rankings vary. Reconstruction error alone cannot determine imputation quality for classification.

Surrogate search followed by fixed-mask comparison controls missing positions but measures conditional robustness; it cannot replace adaptive attacks against the complete imputation–classification pipeline. TG-SMA's control depends on dataset, model, and budget. Exact missing counts do not guarantee reachable targets. Query search is computationally limited, so missing a target does not establish greater robustness.

MID gains are not universal. Training requires complete samples, and gating analysis under locally fixed conditions offers no global robustness guarantee. Artificial missingness on preprocessed complete sequences, used in most experiments, does not cover all real missingness mechanisms. Further evaluation should examine real missingness, generalization across missingness patterns, adaptive attacks against MID, and computational cost.

7 CONCLUSION

AdvMissing-TSC evaluates time-series classification through observation availability. TG-SMA and NES-Mask search within nominal budgets; fixed-mask comparisons separate missingness patterns from subsequent imputation. Adversarially selected masks can cause greater damage than matching random patterns, with exceptions that depend on structure and configuration. Through mask-aware classification and consistency training, MID improves the reported Chinatown–Informer setting with block missingness and Mean imputation, although supplementary results show no uniform benefit. Broader robustness conclusions require evaluation of attacks optimized against MID.

AI USE STATEMENT

Generative AI assisted with translating the Chinese manuscript into English, polishing the prose, and preparing and checking the LaTeX layout.

REPRODUCIBILITY STATEMENT

Appendix A states the assumptions and proofs of the conditional mathematical results. Appendices B and C describe the experimental protocol. The tables distinguish separate experiment groups. The additional empirical summaries in Appendix E use only the five points reported in Table 2.

REFERENCES

Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31(3):606–660, 2017. doi: 10.1007/s10618-016-0483-9.

Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The UEA multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*, 2018. URL https://arxiv.org/abs/1811.00075.

Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. BRITS: Bidirectional recurrent imputation for time series. In *Advances in Neural Information Processing Systems*, volume 31, pp. 6776–6786, 2018.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8:6085, 2018. doi: 10.1038/s41598-018-24271-9.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2206–2216. PMLR, 2020.

Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The UCR time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019. doi: 10.1109/JAS.2019.1911747.

Angus Dempster, François Petitjean, and Geoffrey I. Webb. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5):1454–1495, 2020. doi: 10.1007/s10618-020-00701-z.

Angus Dempster, Daniel F. Schmidt, and Geoffrey I. Webb. MiniRocket: A very fast (almost) deterministic transform for time series classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 248–257. ACM, 2021. doi: 10.1145/3447548.3467231.

Daizong Ding, Mi Zhang, Fuli Feng, Yuanmin Huang, Erling Jiang, and Min Yang. Black-box adversarial attack on time series classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7358–7368, 2023. doi: 10.1609/aaai.v37i6.25896.

Wenjie Du, David Cote, and Yan Liu. SAITS: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023. doi: 10.1016/j.eswa.2023.119619.

Vincent Fortuin, Dmitry Baranchuk, Gunnar Rätsch, and Stephan Mandt. GP-VAE: Deep probabilistic time series imputation. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1651–1661. PMLR, 2020.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2137–2146. PMLR, 2018.

Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. InceptionTime: Finding AlexNet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020. doi: 10.1007/s10618-020-00710-y.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*, 2017.

Fazle Karim, Somshubra Majumdar, and Houshang Darabi. Adversarial attacks on time series. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3309–3320, 2021. doi: 10.1109/TPAMI.2020.2986319.

Mourad Khayati, Philippe Cudré-Mauroux, and Michael H. Böhlen. Scalable recovery of missing blocks in time series with high and low cross-correlations. *Knowledge and Information Systems*, 62(6):2257–2280, 2020a. doi: 10.1007/s10115-019-01421-7.

Mourad Khayati, Alberto Lerner, Zakhar Tymchenko, and Philippe Cudré-Mauroux. Mind the gap: An experimental evaluation of imputation of missing values techniques in time series. *Proceedings of the VLDB Endowment*, 13(5):768–782, 2020b. doi: 10.14778/3377369.3377383.

Hoki Kim, Yunyoung Lee, Woojin Lee, and Jaewook Lee. Towards undetectable adversarial attack on time series classification. *Information Sciences*, 715:122216, 2025. doi: 10.1016/j.ins.2025.122216.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. R-Drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 10890–10905, 2021.

Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Mengna Liu, Pengshuai Yao, Xu Cheng, and Shengyong Chen. An end-to-end model for time series classification in the presence of missing values. *Expert Systems with Applications*, 284:127954, 2025. doi: 10.1016/j.eswa.2025.127954.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.

Matthew Middlehurst, James Large, Michael Flynn, Jason Lines, Aaron Bostrom, and Anthony Bagnall. HIVE-COTE 2.0: a new meta ensemble for time series classification. *Machine Learning*, 110(11–12):3211–3243, 2021. doi: 10.1007/s10994-021-06057-9.

Matthew Middlehurst, Patrick Schäfer, and Anthony Bagnall. Bake off redux: A review and experimental evaluation of recent time series classification algorithms. *Data Mining and Knowledge Discovery*, 38(4):1958–2031, 2024. doi: 10.1007/s10618-024-01022-1.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2019. doi: 10.1109/TPAMI.2018.2858821.

Pradeep Rathore, Arghya Basak, Sri Harsha Nistala, and Venkataramana Runkana. Untargeted, targeted and universal adversarial attacks and defenses on time series. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2020. doi: 10.1109/IJCNN48605.2020.9207272.

Tolou Shadbahr, Michael Roberts, Jan Stanczuk, Julian Gilbey, Philip Teare, Sören Dittmer, Matthew Thorpe, Ramon Viñas Torné, Evis Sala, Pietro Lió, Mishal Patel, Jacobus Preller, AIX-COVNET Collaboration, James H. F. Rudd, Tuomas Mirtti, Antti Sakari Rannikko, John A. D. Aston, Jing Tang, and Carola-Bibiane Schönlieb. The impact of imputation quality on machine learning classifiers for datasets with missing values. *Communications Medicine*, 3:139, 2023. doi: 10.1038/s43856-023-00356-z.

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.

Zhixian Wang, Linxiao Yang, Liang Sun, Qingsong Wen, and Yi Wang. Task-oriented time series imputation evaluation via generalized representers. In *Advances in Neural Information Processing Systems*, volume 37, pp. 137403–137431, 2024. doi: 10.52202/079017-4365.

Kevin Wellenzohn, Michael H. Böhlen, Anton Dignös, Johann Gamper, and Hannes Mitterer. Continuous imputation of missing values in streams of pattern-determining time series. In *Proceedings of the 20th International Conference on Extending Database Technology*, pp. 330–341, 2017. doi: 10.5441/002/edbt.2017.30.

Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. Natural evolution strategies. *Journal of Machine Learning Research*, 15:949–980, 2014.

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

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing data imputation using generative adversarial nets. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5689–5698. PMLR, 2018.

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

A THEORETICAL ANALYSIS

This appendix gives the assumptions and proofs for exact budget projection, temporal continuity regularization, and MID consistency training.

A.1 PROJECTION UNDER AN EXACT ELEMENT BUDGET

Proposition 1. Let $P \in [0,1]^m$, where $m = TC$, and let $0 \le k \le m$ be an integer. Setting the k smallest entries of P to zero and all remaining entries to one solves

$$\Pi_k(P) \in \arg\min_{M \in \{0,1\}^m : \|1-M\|_0 = k} \|M - P\|_2^2. \tag{20}$$

Proof. Let S be the set of removed entries, with $|S| = k$. Expanding the objective gives

$$\|M - P\|_2^2 = \sum_{j=1}^{m} (1 - P_j)^2 + \sum_{j \in S} (2P_j - 1). \tag{21}$$

The first term is independent of S. Selecting the k smallest values therefore minimizes the objective. Ties can be resolved by choosing any sufficient subset of tied indices; removing every tied entry need not satisfy the budget. The result also holds for $k = 0$ and $k = m$.

Continuous parameters determine the locations searched, while index projection enforces the removal count. Candidates can thus be compared at the same number of removed observations.

A.2 MISSING INTERVALS

Proposition 2. For channel c of a binary missingness indicator $R = 1 - M$, define

$$V_c = \sum_{t=1}^{T-1} |R_{t+1,c} - R_{t,c}|. \tag{22}$$

If B_c counts its disjoint nonempty missing intervals, then

$$B_c = \frac{V_c + R_{1,c} + R_{T,c}}{2} \le \frac{V_c}{2} + 1. \tag{23}$$

Proof. An interval that touches neither sequence boundary contributes an entry transition and an exit transition. An interval touching the left or right boundary contributes one fewer transition, supplied by $R_{1,c}$ or $R_{T,c}$. Each interval is therefore counted twice. If the entire sequence is missing, $V_c = 0$ and both boundary indicators equal one, so the identity still holds. The upper bound follows from $R_{1,c} + R_{T,c} \le 2$.

At a fixed budget, fewer mask transitions encourage missing entries to form fewer contiguous intervals. This identity concerns the final binary mask. A continuous TV penalty supplies a structural preference but does not by itself fix the number, length, or shape of hard-mask blocks.

A.3 LOCAL ATTENUATION OF NUMERICAL FEATURE PERTURBATIONS

Lemma 1. For $H_{\mathrm{MID}} = G \odot H_x + (1-G) \odot H_m$, locally fix $G \in [0,1]^m$ and H_m. For $1 \le p \le \infty$,

$$\|\Delta H_{\mathrm{MID}}\|_p \le \|G\|_\infty \|\Delta H_x\|_p. \tag{24}$$

Proof. Subtraction gives $\Delta H_{\mathrm{MID}} = G \odot \Delta H_x$. For finite p,

$$\sum_j |G_j \Delta H_{x,j}|^p \le \|G\|_\infty^p \sum_j |\Delta H_{x,j}|^p. \tag{25}$$

Taking the pth root proves the result. For $p = \infty$, take the coordinatewise maximum. Since $\|G\|_\infty \le 1$, the numerical pathway does not amplify this feature perturbation when the gate is fixed. Smaller gate values attenuate the corresponding components further. The actual gate depends on the input; this result concerns the fused representation under a fixed gate and does not describe the global response of the downstream classifier.

A.4 A SUFFICIENT CONDITION FOR PREDICTION AGREEMENT

Theorem 1. Let p and q be the class probabilities for the complete and missing views. Suppose p has a unique largest-probability class $y^\star$, with margin

$$\gamma = p_{y^\star} - \max_{j \neq y^\star} p_j > 0. \tag{26}$$

If Jensen–Shannon divergence (Lin, 1991) uses natural logarithms and

$$\mathrm{JS}(p\|q) < \gamma^2/8, \tag{27}$$

then $\arg\max_j q_j = y^\star$.

Proof. Let $\mathrm{TV}(p,q) = \frac{1}{2}\|p-q\|_1$. Using $\mathrm{TV}(p,q) \leq \sqrt{2\,\mathrm{JS}(p\|q)}$ and $|p_j - q_j| \leq \mathrm{TV}(p,q)$, for every $j \neq y^\star$ we have

$$q_{y^\star} - q_j \geq \gamma - 2\,\mathrm{TV}(p,q) \geq \gamma - 2\sqrt{2\,\mathrm{JS}(p\|q)} > 0. \tag{28}$$

The missing view therefore predicts $y^\star$. A positive complete-view margin and sufficiently small distribution shift suffice to preserve the predicted class. Agreement does not imply correctness, and a consistency training loss does not ensure that every test sample satisfies this condition.

A.5 CONSISTENCY WEIGHTED BY MISSINGNESS

Corollary 1. Let $w(M) = 1 + \alpha\rho(M)$, where $\alpha \geq 0$ and $\rho(M) = \|1 - M\|_1/(TC)$. If a sample's weighted consistency term satisfies $\mathcal{L}^w_{\mathrm{con}} = w(M)\,\mathrm{JS}(p\|q) \leq \epsilon$, then

$$\mathrm{JS}(p\|q) \leq \frac{\epsilon}{w(M)}, \qquad \mathrm{TV}(p,q) \leq \sqrt{\frac{2\epsilon}{w(M)}}. \tag{29}$$

Proof. Divide the loss bound by the positive weight and apply the divergence inequality above. With training masks fixed, the weight is independent of classifier parameters and scales each sample's consistency gradient by $w(M)$. This gives a conditional interpretation of assigning greater consistency weight to samples with higher missing rates; increasing the weight does not necessarily reduce empirical divergence.

B ALGORITHM DETAILS

Gradient-based attacks and robust optimization provide standard approaches to adversarial evaluation (Goodfellow et al., 2015; Madry et al., 2018). Insufficient attacks can overestimate robustness (Croce & Hein, 2020).

Table 3: TG-SMA candidate generation and retention. Soft masks provide gradients; hard masks enforce the budget and determine candidate scores and outputs.

Step	Operation
1	Construct a soft observation mask P using binary Gumbel relaxation.
2	Obtain the classification loss and accuracy feedback from the soft search pathway.
3	Project each sample to a hard mask and evaluate it with the proxy imputer and frozen classifier.
4	Retain the hard mask if its target gap is the smallest so far, or its accuracy is the lowest in minimization mode.
5	Compute the signed cross-entropy objective and regularizers. Target mode uses negative CE above the target and positive CE otherwise; worst-case mode always uses negative CE.
6	Backpropagate through the soft pathway, update the mask parameters, and repeat.
7	Return the retained Top-k binary mask.

B.1 MASK SAMPLING

Each observation has two states: observed and missing. Combining the observed-state score with a zero score gives binary logits. After adding independent Gumbel noise, the observed-state softmax component is

$$P = \frac{\exp((\Theta + G_1)/\tau)}{\exp((\Theta + G_1)/\tau) + \exp(G_0/\tau)} = \text{sigmoid}\left(\frac{\Theta + G_1 - G_0}{\tau}\right). \tag{30}$$

Binary Gumbel-Softmax and Gumbel-Sigmoid are two expressions for this relaxation (Jang et al., 2017; Maddison et al., 2017). Lowering the temperature moves the soft mask toward discrete states. Top-k index selection determines the final element count independently of whether soft values exceed 0.5.

The retention scores are flattened per sample. The $k = \lfloor sTC \rfloor$ smallest indices are set to zero before restoring the time–channel shape. Tied scores are resolved by selecting exactly k indices. Each retained candidate associates its hard mask with its removal count, proxy imputer, and score.

B.2 SEARCH OBJECTIVE

Target-guided search uses batch accuracy to choose the cross-entropy direction: negative cross-entropy strengthens the attack when accuracy exceeds the target; positive cross-entropy weakens it when accuracy reaches or falls below the target. Accuracy minimization always uses negative cross-entropy. Candidate retention depends on hard-mask accuracy or target gap. This permits location changes within the same budget while retaining the closest candidate found.

The structural terms control temporal transitions, neighboring-channel differences, and synchronization across channels at the same time. Block-missingness coefficients are $\lambda_t = 0.1$, $\lambda_c = 0.05$, $\lambda_{\text{sync}} = 0.1$, and $\lambda_s = 1.0$. Point missingness disables the structural terms. The budget penalty assists the soft search; projection fixes the final count. Channel terms are omitted for univariate inputs, and channel TV is used when channel order has a meaningful interpretation. Initialization distributions appear in Appendix G.4.

B.3 QUERY SEARCH

Natural evolution strategies provide a basis for query-based optimization (Wierstra et al., 2014). Each NES-Mask iteration forms paired positive and negative Gaussian perturbations around the current scores. Every candidate undergoes Top-k projection, proxy imputation, and a model query before feedback updates the scores. Model parameters remain frozen, and only predicted classes are required. Evaluation labels are used to calculate accuracy. Neither classifier nor imputer gradients are needed.

B.4 MID TRAINING

Prediction consistency also underlies self-ensembling (Laine & Aila, 2017) and virtual adversarial regularization (Miyato et al., 2019). R-Drop regularizes agreement between dropout-generated predictions (Liang et al., 2021).

MID supplies a missingness indicator alongside a single imputation. The missing view fuses two C-dimensional feature pathways and maps the result to $2C$ dimensions. Both views share the classifier. The gating network is a $2C \to h \to C$ MLP with GELU and a final sigmoid. The main defense setting uses $h = 2C$, zero gate dropout, gating, and per-sample missing-rate weighting. Inference uses only the missing view and requires neither a complete sample nor repeated imputation or prediction ensembles.

Let epochs start at $e = 0$, let E_{warm} be the warmup length, and let $\rho_{\mathcal{B}}$ be the batch missing rate. The effective consistency coefficient is

$$\lambda_{e,\mathcal{B}} = \lambda_{\max} u_e[\gamma_0 + (1 - \gamma_0)\rho_{\mathcal{B}}], \tag{31}$$

where $u_e = \min\{1, (e+1)/E_{\text{warm}}\}$ if $E_{\text{warm}} > 0$, and $u_e = 1$ otherwise. The main defense loss is

$$\mathcal{L}_{\mathcal{B}} = 2\,\text{CE}_{\text{clean}} + 1.5\,\text{CE}_{\text{miss}} + \frac{\lambda_{e,\mathcal{B}}}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}(1+\alpha\rho(M_i))\,\text{JS}(p_i^{\text{clean}}\|p_i^{\text{miss}}). \tag{32}$$

Table 4: MID training configuration.

Setting	Main defense
Complete-view stop-gradient	Off
Clean/missing CE weights	2 / 1.5
Maximum consistency coefficient $\lambda_{\max}$	0.2
Consistency floor γ_0	0.2
Missing-rate weight α	0.3
Warmup epochs	5

C Evaluation protocol

C.1 Experimental coverage

The ten datasets are Heartbeat, JapaneseVowels, Handwriting, Chinatown, EthanolLevel, UWaveGestureLibrary, SelfRegulationSCP1, SelfRegulationSCP2, EthanolConcentration, and Fungi. These benchmarks belong to the UCR and UEA classification archives (Dau et al., 2019; Bagnall et al., 2018). Different subsets assess attack strength, target control, imputation, classifier differences, and defense recovery.

Table 5: Coverage of the reported comparisons.

Comparison	Datasets	Models / imputers	Results
Fixed-budget attacks	Six datasets	TimesNet, Informer; Linear	Table 1; Fig. 9
Target control	Heartbeat, Handwriting, Chinatown	TimesNet, Informer; two target drops	Sec. 5.3; Figs. 4, 5
Missingness structures	Same three datasets	Informer; Mean, Linear, CDRec	Table 6; Fig. 7
Imputers	Eight datasets	Zero, Mean, Linear, CDRec, BRITS, GAIN, TKCM	Table 7
Query attacks	Four datasets	MiniROCKET; Linear; additional 1NN-DTW results	Table 8; Fig. 6
Classifiers	Eight datasets	TimesNet, Informer, 1NN-DTW, MiniROCKET	Table 9; Fig. 10
MID recovery	Chinatown	Informer; Adv-Block + Mean	Table 2; Fig. 3; App. E
Mask diagnostics	Representative visualizations	Window frequencies, spectra, soft-mask histograms	Figs. 11, 12

The eight-dataset tables contain Heartbeat, JapaneseVowels, EthanolLevel, Fungi, Handwriting, Chinatown, SelfRegulationSCP2, and UWaveGestureLibrary. Linear and interpolation refer to the same imputation method. MiniROCKET uses a nearly deterministic variant of random convolutional features (Dempster et al., 2021). Other classification approaches include InceptionTime convolutional ensembles (Ismail Fawaz et al., 2020) and HIVE-COTE 2.0 ensembles across representations (Middlehurst et al., 2021).

C.2 MISSINGNESS PROTOCOL

Random-P denotes random point missingness; Time-B, random temporal blocks; Rect-B, random rectangular blocks; and Adv-P and Adv-B, adversarial point and block missingness. Generated masks remain fixed across the seven imputers. Point attacks are paired with random point missingness, and block attacks with the corresponding random block baseline. Comparisons across structures reflect both structure and location selection.

C.3 EVALUATION METRICS

Accuracy a_s is the fraction of correctly classified samples at nominal budget s. The relative drop is $100(a_{\text{clean}} - a_s)/a_{\text{clean}}$, whereas $100(a_1 - a_2)$ is a difference in percentage points. Negative relative drops indicate accuracy above the corresponding complete-input baseline and are not clipped.

$$\text{AURC}_{[s_0,s_1]} = \frac{1}{s_1 - s_0} \int_{s_0}^{s_1} a_s \, \mathrm{d}s. \tag{33}$$

AURC is normalized by interval length. The reported interval is $[0.05, 0.40]$, evaluated at $\{0.05, 0.10, 0.20, 0.30, 0.40\}$. Acc@30% describes a fixed budget. Because these five points are not equally spaced, their arithmetic mean differs from the normalized trapezoidal area. Appendix E calculates this area from Table 2; other AURC values are reported aggregates.

C.4 TARGET ATTAINMENT

For relative drop r, the target is $a^\star = (1 - r)a_{\text{clean}}$, and the gap is $|a_s - a^\star|$. A gap of at most 0.03 counts as a hit. For a specified evaluation set $\mathcal{C}$,

$$\text{HitRate}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathbf{1}\{\text{TargetGap}_c \leq 0.03\}. \tag{34}$$

The minimum reported gap is 0.0254 for Heartbeat–Informer with Adv-Point at $r = 0.2$, and 0.0234 for Heartbeat–TimesNet with Adv-Block at $r = 0.3$. Handwriting–TimesNet has a mean gap of 0.0346 and a reported hit rate of 0.40.

D EXTENDED RESULTS

The result groups below examine missingness structure, imputation, and classifiers. Each result group has its own complete-input baselines and experimental settings; results from different groups are not combined into one robustness curve.

D.1 TASK DEPENDENCE OF MISSINGNESS STRUCTURE

Table 6 compares five missingness types for Informer. On Heartbeat, adversarial point missingness has Acc@30% of 0.717, below random point missingness at 0.780. On Handwriting, adversarial blocks yield 0.181, below random temporal blocks at 0.200 and random points at 0.299. On Chinatown, random rectangular blocks cause more damage than adversarial blocks, so structural constraints matter when interpreting attack strength.

Table 6: Missingness structures for Informer, aggregating Mean, Linear, and CDRec imputation. Drop values are the reported relative drops; no shared baseline is inferred from rounded accuracies.

Dataset	Pattern	AURC	Acc@30%	Drop@30%
Heartbeat	Random-P	0.777	0.780	0.00%
Heartbeat	Time-B	0.766	0.763	2.29%
Heartbeat	Rect-B	0.769	0.767	1.67%
Heartbeat	Adv-P	0.739	0.717	8.13%
Heartbeat	Adv-B	0.759	0.746	4.38%
Handwriting	Random-P	0.305	0.299	7.64%
Handwriting	Time-B	0.232	0.200	38.18%
Handwriting	Rect-B	0.258	0.245	24.24%
Handwriting	Adv-P	0.276	0.269	16.97%
Handwriting	Adv-B	0.216	0.181	44.00%
Chinatown	Random-P	0.855	0.832	11.66%
Chinatown	Time-B	0.815	0.781	17.03%
Chinatown	Rect-B	0.726	0.693	26.42%
Chinatown	Adv-P	0.805	0.793	15.79%
Chinatown	Adv-B	0.754	0.710	24.56%

D.2 Imputers under fixed masks

CDRec uses centroid decomposition (Khayati et al., 2020a), while TKCM uses top-k case matching (Wellenzohn et al., 2017). Dedicated benchmarks compare methods for missing-block recovery (Khayati et al., 2020b). Probabilistic imputation can use a latent Gaussian-process model (Fortuin et al., 2020) or conditional diffusion (Tashiro et al., 2021); GRU-D incorporates missingness masks and observation intervals into prediction (Che et al., 2018).

The seven imputers face the same removed observations in Table 7. For adversarial point missingness on JapaneseVowels, Linear reaches 0.96 for both AURC and Acc@30%. For adversarial blocks on Heartbeat, CDRec and TKCM achieve Acc@30% of 0.80, above Linear at 0.78.

Table 7: Fixed-mask comparison of seven imputers. Rand-B(avg) is the reported random-block aggregate; a dash denotes an unreported value. J. Vowels, Ethanol, UWave, and SCP2 abbreviate JapaneseVowels, EthanolLevel, UWaveGestureLibrary, and SelfRegulationSCP2.

(a) AURC

Dataset	Pattern	Zero	Mean	Linear	CDRec	BRITS	GAIN	TKCM
Heartbeat	Adv-P	0.64	0.69	0.73	0.72	0.68	0.70	0.69
	Adv-B	0.75	0.75	0.79	0.78	0.78	0.77	0.77
	Rand-B(avg)	0.78	0.78	0.75	0.76	0.74	0.75	0.78
J. Vowels	Adv-P	0.88	0.89	0.96	0.94	0.96	0.96	0.94
	Adv-B	0.83	0.86	0.91	0.86	0.89	0.88	0.93
	Rand-B(avg)	0.91	0.94	0.94	0.91	0.96	0.95	0.94
Ethanol	Adv-P	0.52	0.50	0.62	0.50	0.50	0.55	0.54
	Adv-B	0.48	0.48	0.60	0.48	0.48	0.52	0.58
	Rand-B(avg)	0.55	0.55	0.63	0.55	0.56	0.57	0.6
Fungi	Adv-P	0.90	0.89	0.97	0.94	0.96	0.89	0.94
	Adv-B	0.90	0.89	0.97	0.89	0.94	0.61	0.88
	Rand-B(avg)	0.92	0.92	0.98	0.92	0.96	0.9	0.97
Handwriting	Adv-P	0.308	0.305	0.355	0.298	0.295	0.290	0.342
	Adv-B	0.355	0.352	0.390	0.355	0.360	0.352	0.378
	Rand-B(avg)	0.345	0.358	0.385	0.352	0.352	0.362	0.378
Chinatown	Adv-P	0.79	0.80	0.96	0.82	0.83	0.86	0.92
	Adv-B	0.82	0.85	0.98	0.84	0.86	0.86	0.93
	Rand-B(avg)	0.84	0.88	0.99	0.86	0.89	0.90	0.95
SCP2	Adv-P	0.51	0.57	0.54	0.51	0.46	0.53	0.54
	Adv-B	0.67	0.68	0.83	0.67	0.83	0.56	0.86
	Rand-B(avg)	0.75	0.76	0.87	0.74	0.87	0.60	0.88
UWave	Adv-P	0.67	0.70	0.83	0.61	0.75	0.74	0.76
	Adv-B	0.80	0.82	0.86	0.83	0.83	0.82	0.90
	Rand-B(avg)	0.82	0.84	0.88	0.83	0.86	0.86	0.92

(b) Acc@30%

Dataset	Pattern	Zero	Mean	Linear	CDRec	BRITS	GAIN	TKCM
Heartbeat	Adv-P	0.73	0.76	0.78	0.77	0.77	0.77	0.76
	Adv-B	0.76	0.76	0.78	0.80	0.76	0.76	0.80
	Rand-B(avg)	0.74	0.74	0.73	0.74	0.72	0.74	0.76
J. Vowels	Adv-P	0.89	0.94	0.96	0.94	0.96	0.89	0.94
	Adv-B	0.83	0.86	0.91	0.86	0.88	0.88	0.86
	Rand-B(avg)	0.91	0.94	0.94	0.91	0.96	0.77	0.91
Ethanol	Adv-P	0.52	0.50	0.62	0.50	0.50	0.36	0.54
	Adv-B	0.52	0.50	0.62	0.52	0.54	0.35	—
	Rand-B(avg)	0.52	0.52	0.63	0.52	0.54	0.55	0.58
Fungi	Adv-P	0.97	0.97	0.97	0.94	0.96	0.96	0.97
	Adv-B	0.90	0.89	0.97	0.89	0.94	0.67	0.88
	Rand-B(avg)	0.95	0.95	0.98	0.95	0.97	0.95	0.98
Handwriting	Adv-P	0.355	0.350	0.390	0.350	0.360	0.250	0.342
	Adv-B	0.350	0.350	0.390	0.355	0.360	0.352	0.378
	Rand-B(avg)	0.342	0.345	0.385	0.352	0.352	0.362	0.378
Chinatown	Adv-P	0.79	0.80	0.96	0.82	0.83	0.69	0.82
	Adv-B	0.82	0.85	0.98	0.84	0.86	0.86	0.83
	Rand-B(avg)	0.88	0.83	0.90	0.83	0.88	0.87	0.83
SCP2	Adv-P	0.47	0.51	0.54	0.51	0.46	0.53	0.49
	Adv-B	0.51	0.54	0.60	0.54	0.51	0.54	0.54
	Rand-B(avg)	0.57	0.56	0.56	0.57	0.53	0.56	0.56
UWave	Adv-P	0.76	0.80	0.86	0.83	0.84	0.82	0.83
	Adv-B	0.80	0.83	0.86	0.83	0.83	0.86	0.86
	Rand-B(avg)	0.85	0.86	0.90	0.86	0.88	0.88	0.92

AURC and Acc@30% can favor different methods. For adversarial blocks on SelfRegulationSCP2, TKCM has AURC 0.86, above Linear at 0.83, while their Acc@30% values are 0.54 and 0.60. Reporting both avoids relying solely on one budget or an interval summary. These metrics concern classification, not reconstruction error.

D.3 Query attack results

Table 8 reports absolute accuracy and relative drops. At budgets of 30% and 40%, NES yields lower accuracy than random blocks on Chinatown, Handwriting, and JapaneseVowels. On Handwriting at 40%, the relative drop reaches 58.56%. Random blocks are stronger on Heartbeat.

Table 8: Query attacks on MiniROCKET with interpolation. Accuracy values are fractions; drops are percentages relative to complete-input accuracy. Negative drops indicate accuracy above the complete-input baseline. J. Vowels denotes JapaneseVowels.

Dataset	Attack	Complete	AURC	Acc@30%	Acc@40%	Drop@30%	Drop@40%
Chinatown	Random-P	0.983	0.973	0.974	0.950	0.89%	3.26%
	Random-B	0.983	0.959	0.933	0.924	5.04%	5.93%
	NES-Mask	0.983	0.926	0.889	0.851	9.50%	13.35%
Handwriting	Random-P	0.261	0.259	0.256	0.255	1.80%	2.25%
	Random-B	0.261	0.214	0.194	0.168	25.68%	35.59%
	NES-Mask	0.261	0.203	0.154	0.108	40.99%	58.56%
Heartbeat	Random-P	0.751	0.759	0.766	0.756	-1.95%	-0.65%
	Random-B	0.751	0.743	0.727	0.741	3.25%	1.30%
	NES-Mask	0.751	0.749	0.741	0.751	1.30%	0.00%
J. Vowels	Random-P	0.989	0.981	0.978	0.973	1.09%	1.64%
	Random-B	0.989	0.953	0.938	0.878	5.19%	11.26%
	NES-Mask	0.989	0.917	0.884	0.836	10.66%	15.52%

D.4 Comparison of four classifiers

Table 9 compares two neural classifiers and two classifiers that do not require gradient access. Across all 32 dataset–classifier combinations, adversarial blocks have lower Acc@30% than adversarial points and the aggregate random-block baseline.

Table 9: Comparison of four classifiers in the separate eight-dataset result group. Rand-B(avg) is the reported random-block aggregate. Dataset abbreviations follow Table 7.

(a) AURC

Dataset	Pattern	TimesNet	Informer	1NN-DTW	MiniROCKET
Heartbeat	Adv-P	0.742	0.724	0.628	0.738
	Adv-B	0.691	0.674	0.572	0.702
	Rand-B(avg)	0.771	0.758	0.716	0.793
J. Vowels	Adv-P	0.902	0.937	0.928	0.913
	Adv-B	0.867	0.862	0.884	0.875
	Rand-B(avg)	0.908	0.914	0.946	0.929
Ethanol	Adv-P	0.523	0.502	0.462	0.558
	Adv-B	0.474	0.446	0.418	0.506
	Rand-B(avg)	0.574	0.552	0.497	0.596
Fungi	Adv-P	0.947	0.941	0.914	0.958
	Adv-B	0.895	0.872	0.852	0.902
	Rand-B(avg)	0.961	0.925	0.936	0.969
Handwriting	Adv-P	0.362	0.351	0.339	0.379
	Adv-B	0.294	0.281	0.267	0.309
	Rand-B(avg)	0.391	0.374	0.362	0.405
Chinatown	Adv-P	0.823	0.817	0.792	0.846
	Adv-B	0.768	0.755	0.728	0.786
	Rand-B(avg)	0.895	0.862	0.837	0.947
SCP2	Adv-P	0.563	0.512	0.505	0.617
	Adv-B	0.497	0.456	0.446	0.553
	Rand-B(avg)	0.846	0.736	0.802	0.864
UWave	Adv-P	0.826	0.614	0.804	0.724
	Adv-B	0.781	0.568	0.746	0.681
	Rand-B(avg)	0.848	0.826	0.876	0.916

(b) Acc@30%

Dataset	Pattern	TimesNet	Informer	1NN-DTW	MiniROCKET
Heartbeat	Adv-P	0.761	0.772	0.701	0.681
	Adv-B	0.708	0.713	0.646	0.632
	Rand-B(avg)	0.827	0.742	0.754	0.667
J. Vowels	Adv-P	0.904	0.941	0.931	0.902
	Adv-B	0.852	0.858	0.843	0.849
	Rand-B(avg)	0.887	0.907	0.889	0.946
Ethanol	Adv-P	0.506	0.498	0.481	0.538
	Adv-B	0.468	0.457	0.424	0.486
	Rand-B(avg)	0.528	0.523	0.516	0.572
Fungi	Adv-P	0.928	0.936	0.927	0.914
	Adv-B	0.872	0.864	0.846	0.836
	Rand-B(avg)	0.948	0.943	0.958	0.936
Handwriting	Adv-P	0.371	0.366	0.356	0.386
	Adv-B	0.289	0.291	0.276	0.304
	Rand-B(avg)	0.384	0.369	0.357	0.397
Chinatown	Adv-P	0.818	0.821	0.781	0.793
	Adv-B	0.739	0.746	0.708	0.734
	Rand-B(avg)	0.778	0.826	0.773	0.827
SCP2	Adv-P	0.664	0.514	0.538	0.646
	Adv-B	0.558	0.463	0.482	0.574
	Rand-B(avg)	0.892	0.573	0.846	0.898
UWave	Adv-P	0.832	0.827	0.826	0.718
	Adv-B	0.761	0.758	0.774	0.656
	Rand-B(avg)	0.837	0.858	0.895	0.846

Model rankings depend on the task. Under adversarial blocks on UWaveGestureLibrary, TimesNet has AURC 0.781, compared with MiniROCKET's 0.681. Under the same missingness type on Handwriting, MiniROCKET has Acc@30% 0.304, above the other three models.

E MID RECOVERY

E.1 INTEGRATION OVER THE FIVE REPORTED BUDGETS

Using the five accuracies in Table 2, with $(s_1, \ldots, s_5) = (0.05, 0.10, 0.20, 0.30, 0.40)$, the normalized trapezoidal statistic is

$$\mathrm{AURC} = \frac{1}{s_5 - s_1} \sum_{j=1}^{4} (s_{j+1} - s_j) \frac{a_j + a_{j+1}}{2}. \tag{35}$$

The first interval has width 0.05 and each subsequent interval has width 0.10. MID achieves 82.9286%, compared with 74.4286% for its Adv-Block + Mean baseline and 80.8571% for undefended Adv-Point + CDRec.

E.2 POINTWISE RECOVERY

The matched-baseline gain compares MID with undefended Adv-Block + Mean. The best-undefended difference compares it with Adv-Point + CDRec, the strongest of the four undefended configurations.

MID improves its matched baseline at all five budgets and exceeds the best undefended configuration at the four points from 0.10 to 0.40. Its interval gains are 8.50 and 2.07 percentage points, respectively, while retaining block missingness and mean imputation.

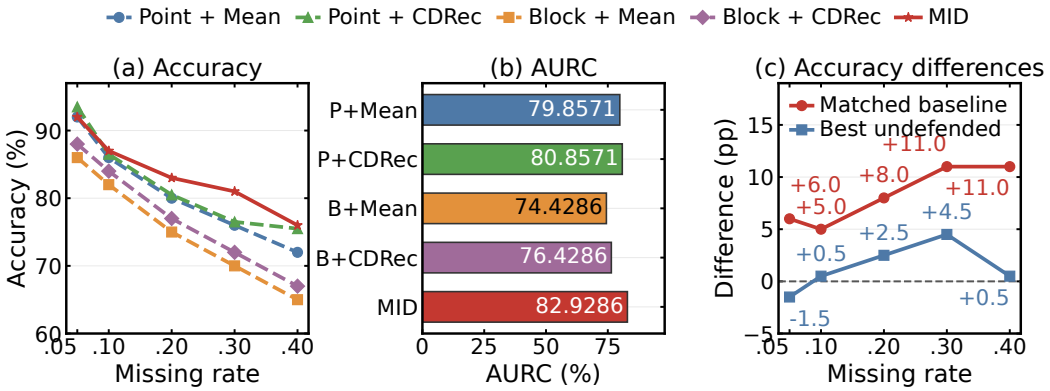


Figure 3: MID recovery on Chinatown–Informer. (a) Accuracy at the five reported budgets. (b) Normalized trapezoidal AURC (%). (c) Gain over the matched Adv-Block + Mean baseline and difference from the best undefended configuration, Adv-Point + CDRec, in percentage points. P and B denote point and block missingness. MID retains Adv-Block + Mean throughout.

F SUPPLEMENTARY EXPERIMENTS

F.1 TARGET GAP VERSUS BUDGET

Figure 4 shows all six dataset–model pairs. Handwriting–TimesNet approaches the tolerance range as the budget increases. Some Heartbeat configurations also approach their targets at larger budgets, consistent with Section 5.3.

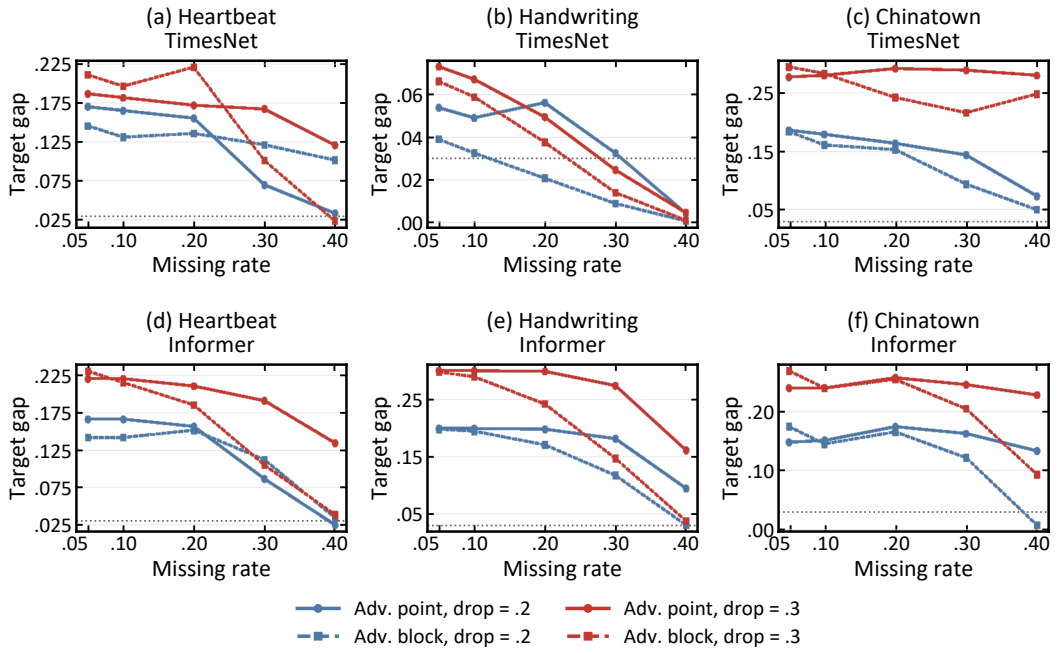


Figure 4: Target gaps for TimesNet (top row) and Informer (bottom row), with Heartbeat, Handwriting, and Chinatown from left to right. Relative drops are 0.2 and 0.3, and nominal budgets range from 0.05 to 0.40. The horizontal tolerance line is 0.03.

F.2 Representative target-accuracy curves

On Handwriting, point and block attacks approach the target at larger budgets. Chinatown remains above the target, showing task-dependent attainability. The complete-input levels in Figure 5 belong to this result group and do not replace the baselines in Table 1.

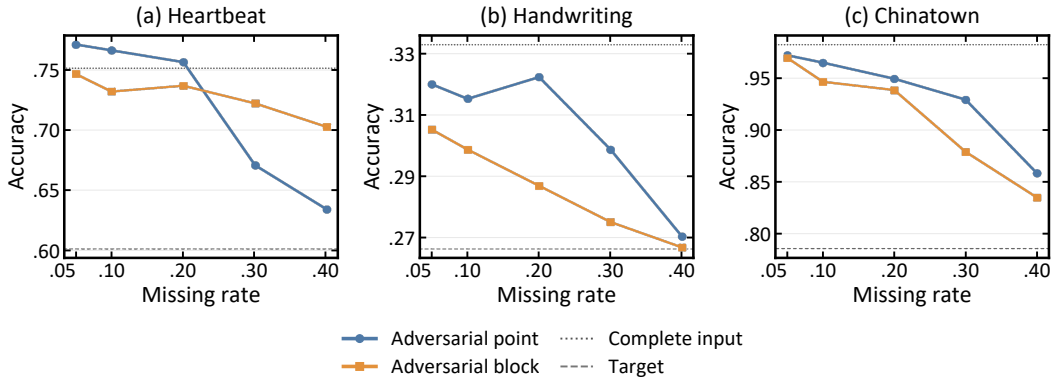


Figure 5: Representative TimesNet accuracy curves for a target relative drop of 0.2: Heartbeat, Handwriting, and Chinatown (left to right).

F.3 Query attacks without model gradients

On Chinatown, Handwriting, and JapaneseVowels, NES falls below the random-block baseline at both 0.30 and 0.40. Heartbeat has a different ranking. Figure 6(a–d) and Table 8 show that query feedback can find vulnerable locations, while its effectiveness depends on the task and search resources.

F.4 MISSINGNESS STRUCTURE

Figure 6(e–f) shows TimesNet under five missingness structures. The adversarial-block curve on Chinatown is below its corresponding random curve. On Heartbeat, adversarial points are stronger over part of the range. These patterns show how location optimization exposes different forms of observation dependence.

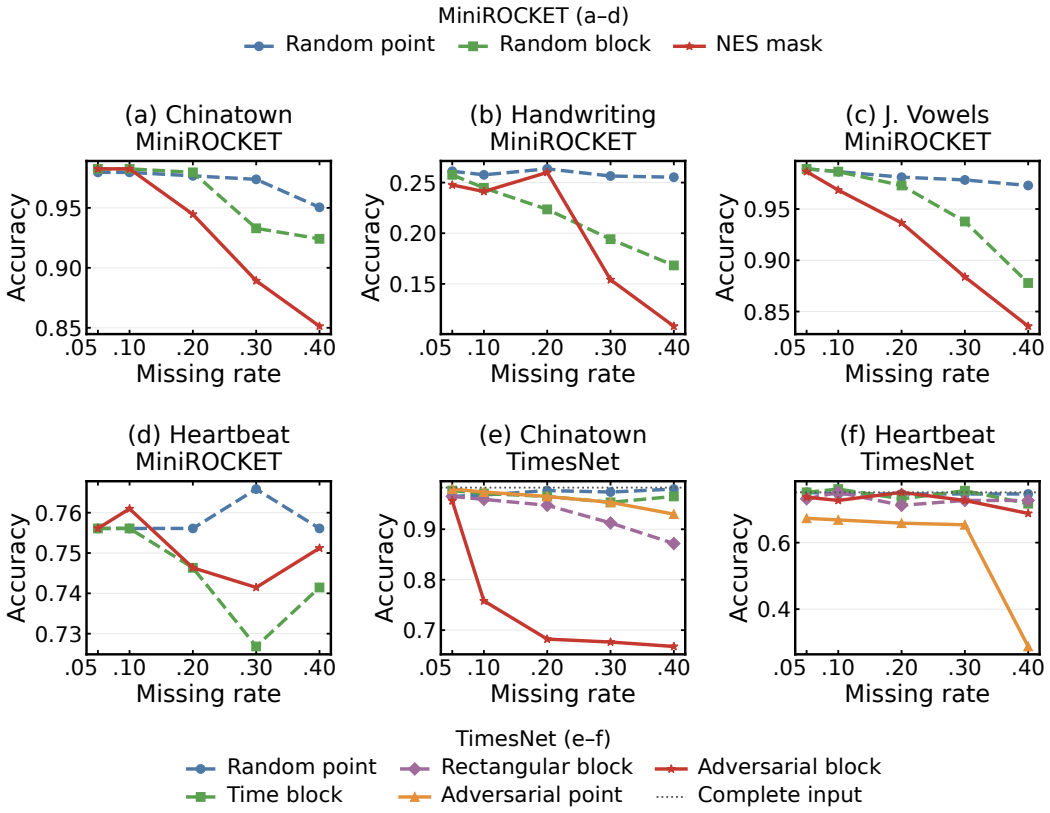


Figure 6: Robustness curves with interpolation. (a–d) MiniROCKET on Chinatown, Handwriting, JapaneseVowels, and Heartbeat, comparing random points, random blocks, and NES masks. (e–f) TimesNet on Chinatown and Heartbeat, comparing five missingness structures. The two legends identify the methods in their respective panel groups. Complete-input accuracies for TimesNet are 0.983 and 0.751 in this separate result group. Adversarial curves result from heuristic search, not certified worst-case bounds.

For Informer on Handwriting, contiguous missing blocks cause greater degradation than random points. Heartbeat and Chinatown exhibit different rankings, supporting evaluation across several structures to identify task-specific risks.

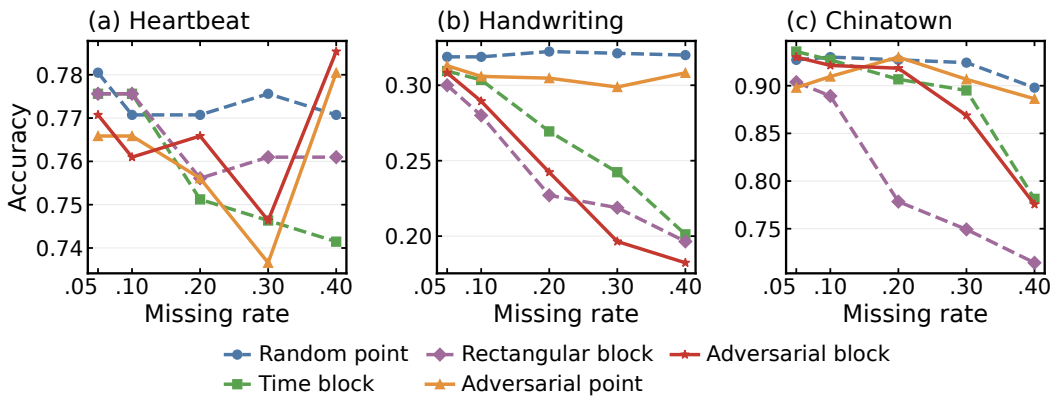


Figure 7: Informer robustness with interpolation on Heartbeat, Handwriting, and Chinatown (left to right). The shared legend identifies the five missingness structures.

F.5 MID UNDER OTHER MISSINGNESS SETTINGS

On Chinatown, the MID random-point curve is above the matching undefended curve. MID also improves adversarial-block accuracy at higher budgets (Figure 8). On Handwriting, improvements appear for random and adversarial points; on Heartbeat, some higher-budget block settings improve. These results retain the differences across tasks and missingness structures.

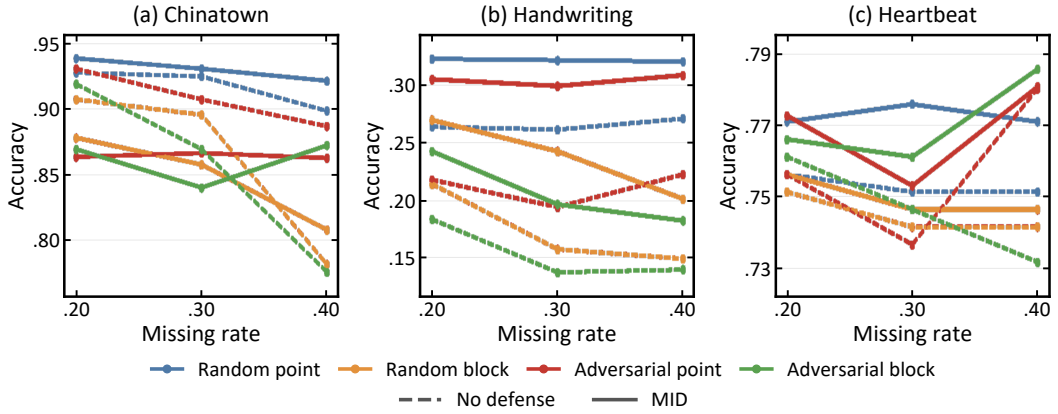


Figure 8: Additional MID results on Chinatown, Handwriting, and Heartbeat (left to right) at budgets of 0.20, 0.30, and 0.40. Color denotes missingness type; line style distinguishes MID from no defense. Each panel retains its own accuracy scale.

G MASK DIAGNOSTICS

G.1 RELATIVE DROPS ACROSS DATASETS

On all six datasets, adversarial points cause larger relative drops than random points, and adversarial blocks cause larger drops than random blocks. The adversarial-block and random-block drops are 39.7% and 21.1% on Handwriting, and 26.0% and 6.3% on Chinatown. Figure 9 summarizes the selective-missingness risk across tasks.

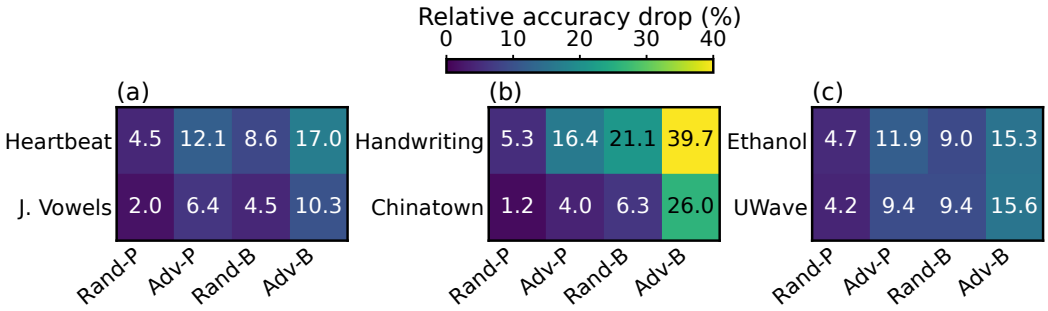


Figure 9: Relative accuracy drops across six datasets and four missingness types, averaged over TimesNet and Informer. The three panels retain the original dataset order, with two datasets per panel. Cells and the shared color bar show percentages. J. Vowels, Ethanol, and UWave abbreviate JapaneseVowels, EthanolLevel, and UWaveGestureLibrary.

G.2 Multiple robustness metrics across classifiers

Figure 10 combines an interval metric with fixed-budget accuracy. MiniROCKET performs prominently on the random-block AURC axis, while TimesNet is competitive on the adversarial-missingness accuracy axes. Table 9 provides the corresponding task-level comparisons.

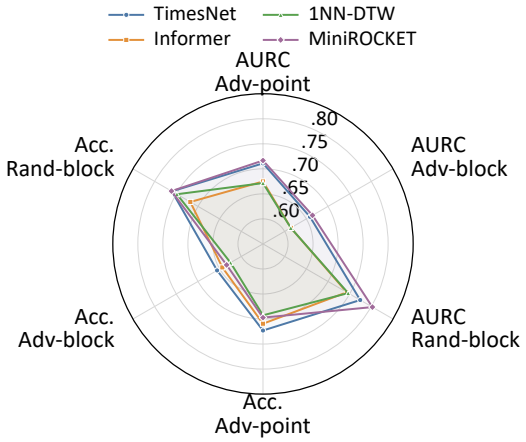


Figure 10: Robustness of four classifiers, averaged across the eight datasets in Table 9. The six axes give AURC and accuracy under adversarial point, adversarial block, and random block missingness.

G.3 Class-dependent missing locations

Figure 11 arranges the two Chinatown classes in rows, with sequences and masked positions, amplitude spectra, and missingness frequencies in four-hour windows in the three columns. High-frequency windows are nonuniform for both classes, and their locations differ. These preferences provide candidate time regions for observation protection and targeted evaluation.

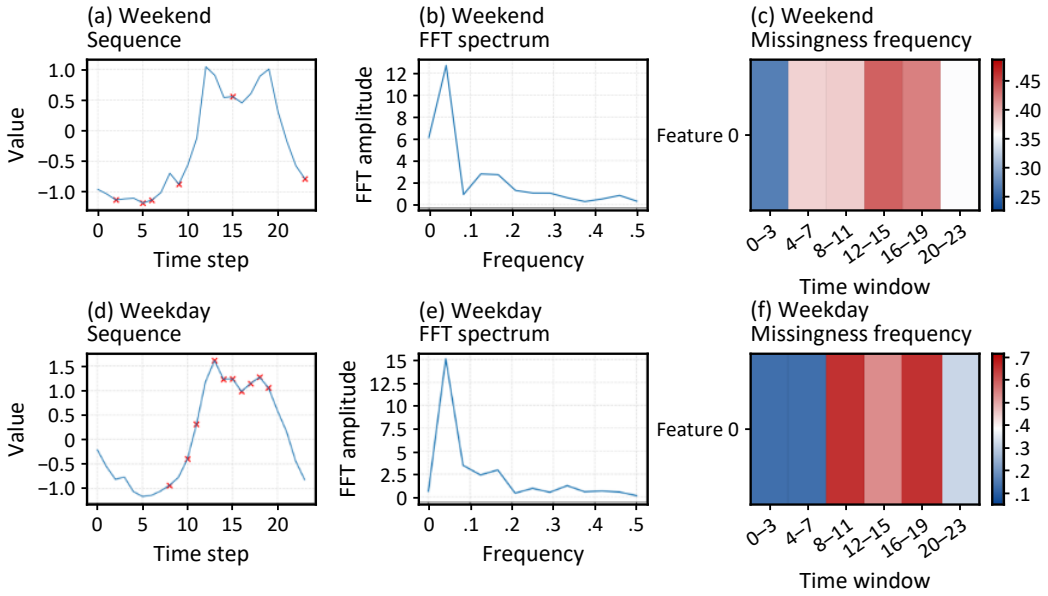


Figure 11: Class-dependent temporal diagnostics on Chinatown. Rows correspond to weekend and weekday samples. Columns show the sequence with missing locations marked by red crosses, its FFT amplitude spectrum, and missingness frequency in four-hour windows. The two frequency panels retain their original, separate color scales.

G.4 Mask-score initialization

Figure 12 compares soft observation masks initialized with scores of -3.0, -1.0, and 0.5. At -3.0, soft values concentrate near zero. At -1.0, mass at both ends becomes more apparent. At 0.5, a larger share lies near the observed state at one. These distributions help identify an early preference for either state. The final removal count is determined by exact Top-k projection; a soft-value distribution or a 0.5 threshold does not guarantee the hard-mask count.

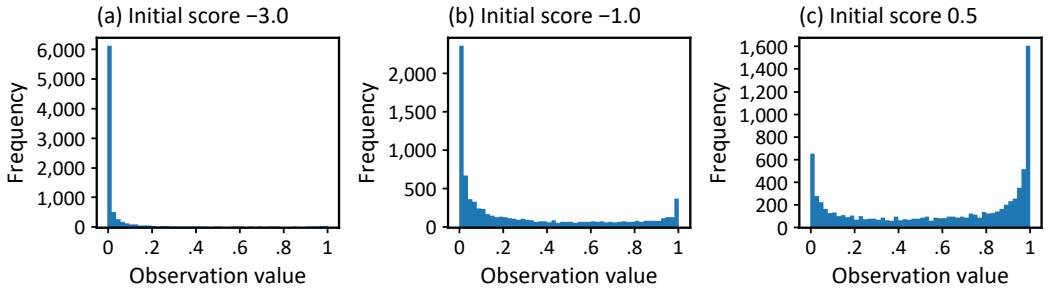


Figure 12: Soft observation-mask histograms for three score initializations. The horizontal axis is the soft observation value before binarization or projection; the vertical axis is frequency. Vertical scales differ.