C2FDA: COARSE-TO-FINE DIFFUSION FOR LOCAL BLACK-BOX ATTACK ON TIME SERIES CLASSIFIERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Time series classifiers remain vulnerable to adversarial perturbations that can alter model predictions while preserving much of the original temporal structure. Local black-box attacks must determine both *where* to perturb and *what* perturbation to generate, while restricting perturbations to a small subset of timestamps. Existing methods often separate temporal support selection from perturbation generation, despite both decisions depending on broader temporal context. We propose **C2FDA**, the first diffusion-based framework for local black-box adversarial attacks on time series classifiers. C2FDA couples the *where* and *what* decisions through shared coarse-to-fine temporal representations. Coarse attention captures region-level relevance, while fine attention refines it into timestamp-level evidence. For *where*, region-level and timestamp-level evidences are combined to determine an exact-budget Top-k temporal support. For *what*, coarse relevance provides regional context for fine refinement, and the resulting timestamp-level evidence modulates diffusion features before noise prediction to shape perturbation content. A shared hybrid boundary objective encourages boundary crossing on the final masked perturbation, jointly optimizing temporal support and perturbation content. C2FDA performs black-box attacks through cross-model transfer. Experiments across diverse time series datasets and classifier architectures show state-of-the-art attack success rates while maintaining low perturbation distortion and high temporal similarity. Code is available at `https://anonymous.4open.science/r/C2FDA-2CC0/`.

1 INTRODUCTION

Time series classification, widely used in human activity recognition, physiological monitoring, industrial diagnosis, and intelligent sensing (Foumani et al., 2024), requires models to capture both observed values and their temporal dependencies. Despite continued advances in temporal representation learning (Jung & Kim, 2026; Xie et al., 2026; Huang et al., 2026), time series classifiers remain vulnerable to adversarial examples (Fawaz et al., 2019; Karim et al., 2021; Pialla et al., 2025), where small perturbations can alter predictions while largely preserving the original temporal structure. This vulnerability has shifted increasing attention toward generating effective adversarial perturbations under limited temporal support, where only a small subset of timestamps can be modified (Yang et al., 2022; 2026).

A central challenge in time series attacks is deciding both *where* to perturb and *what* perturbation to generate. As shown in Figure 1(a), FGSM (Goodfellow et al., 2015) determines perturbation content directly from model gradients but applies it across the full sequence; BIM (Kurakin et al., 2018) and PGD (Madry et al., 2018) follow a similar gradient-driven strategy. These methods therefore provide a direct mechanism for constructing *what* to perturb, but do not explicitly address *where* to perturb under a limited temporal budget or coordinate perturbation content with a selected local support. Local attacks introduce temporal support constraints, yet typically handle localization and perturbation generation through separate procedures. As shown in Figure 1(b), SGAN (Yang et al., 2026) uses shapelets learned beforehand from data as localization priors, then generates perturbations through a GAN with adaptive adjustment. TSadv (Yang et al., 2022) similarly relies on precomputed shapelets before differential-evolution search. Both therefore require a separate shapelet-discovery stage to determine *where*, rather than learning support jointly with perturbation content. Moreover, their interval-level support can be too coarse when only a few timestamps may

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.

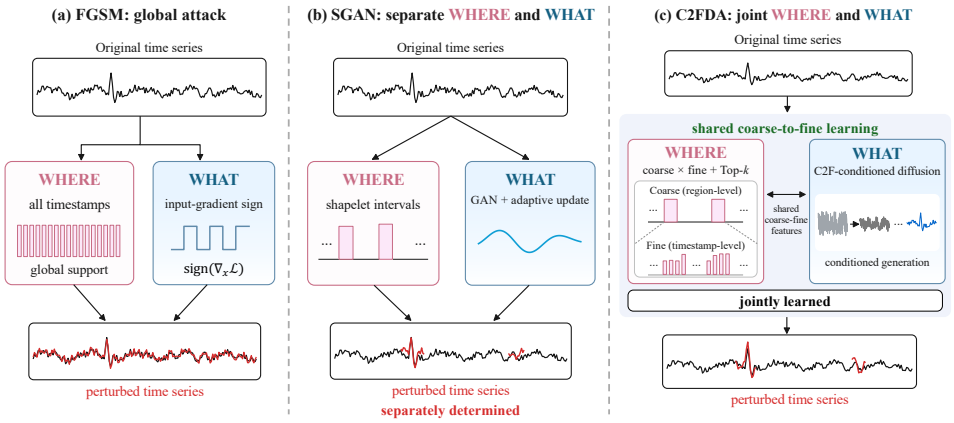


Figure 1: Strategies for determining perturbation support (*where*) and content (*what*). FGSM perturbs the full sequence, SGAN determines local support and perturbation content separately, whereas C2FDA learns both from shared coarse-to-fine temporal representations.

be modified. More generally, separating the two decisions is limiting for time series because both depend on temporal context: the relevance of an individual timestamp often depends on broader temporal patterns (Ye & Keogh, 2009). In other words, for *where*, coarse interval selection may allocate the limited budget imprecisely; for *what*, perturbation content generated independently of support selection may be poorly matched to the selected region or introduce unnecessary distortion, leaving the *where–what* interaction under multi-scale temporal dependencies insufficiently modeled.

To address these limitations, we propose **C2FDA**, a coarse-to-fine diffusion framework for local black-box attacks on time series classifiers. As illustrated in Figure 1(c), C2FDA couples temporal support selection and perturbation synthesis through shared coarse-to-fine temporal representations: coarse attention captures region-level relevance, while fine attention refines it into timestamp-level evidence. The resulting temporal evidence jointly determines an exact-budget support for *where* to perturb and guides diffusion-based generation of *what* perturbation to apply, integrating the two decisions within a unified pipeline rather than separate procedures. A shared hybrid boundary objective further optimizes the final masked perturbation toward boundary crossing, and cross-model transfer enables target-query-free black-box attacks under temporal coverage and perturbation magnitude constraints. Our main contributions are summarized as follows:

(1) We propose C2FDA, the first diffusion-based framework for local black-box adversarial attacks on time series classifiers, formulating local attacks as a coupled *where–what* problem.

(2) A coarse-to-fine attention mechanism provides shared temporal evidence for both attack decisions: it guides diffusion-based perturbation synthesis (*what*) and determines exact-budget temporal support (*where*), while a shared hybrid boundary objective jointly optimizes both paths.

(3) Extensive experiments across heterogeneous time series datasets and classifier architectures show that C2FDA achieves state-of-the-art attack success rates while maintaining low perturbation distortion and high temporal similarity.

2 Related Work

Adversarial attacks on time series differ in both *what* perturbation is generated and *where* it is applied. Unlike white-box attacks that require model internals, black-box attacks better reflect practical settings where such access is unavailable. Existing methods are mainly gradient-, query-, or transfer-based, while local attacks further constrain temporal support, making the coordination between *where* and *what* central to our setting.

Global adversarial attacks on time series. Gradient-based methods are among the most widely studied white-box attacks. FGSM (Goodfellow et al., 2015) constructs perturbation content (*what*) from the input-gradient sign, while BIM (Kurakin et al., 2018) and PGD (Madry et al., 2018) itera-

tively refine it under a norm constraint; in these methods, the perturbation support (*where*) typically spans the full sequence. Their effectiveness on time series classifiers has been demonstrated empirically (Fawaz et al., 2019), while Diff-PGD (Xue et al., 2023) incorporates a diffusion prior into projected-gradient optimization. Under surrogate-to-target transfer, these methods perform black-box attacks without target-model queries.

Query-based black-box attacks optimize perturbation content (*what*) through repeated target-model evaluations, while temporal support (*where*) generally remains global. ZOO (Chen et al., 2017) uses zeroth-order optimization, NES (Ilyas et al., 2018) and SPSA (Uesato et al., 2018) estimate update directions from queries, and AutoZOOM (Tu et al., 2019) reduces the search dimension using an autoencoder. These methods avoid target gradients but still require target-model queries during attack generation.

Learned generators instead synthesize perturbation content through transferability without target queries at inference. AdvGAN (Xiao et al., 2018) learns perturbations from a surrogate or distilled model; GATN (Karim et al., 2021) extends learned adversarial transformation to time series, while GE-AdvGAN (Zhu et al., 2024) improves transferability through gradient editing. Their *what* generation is learned, but *where* is generally not explicitly localized, so the sequence is perturbed globally. Table 1 summarizes these methods together with local attacks in terms of black-box access, temporal locality, target-query requirements, and joint *where–what* modeling.

Local black-box attacks on time series. Local black-box attacks constrain *where* perturbations can be applied, reflecting that discriminative information in time series may be concentrated in short temporal patterns. TSadv (Yang et al., 2022) selects shapelet intervals and searches within them using differential evolution, requiring repeated target-model queries. BlackTreeS (Ding et al., 2023) uses tree search to identify influential temporal positions and estimates black-box gradients at the selected positions, likewise relying on target queries. SGAN (Yang et al., 2026) combines shapelet-guided localization with GAN-based generation and adaptive adjustment for query-free inference.

Despite their different mechanisms, these methods still handle perturbation support and content through separate procedures. TSadv and SGAN determine shapelet-based support independently of perturbation generation, while BlackTreeS alternates position search with black-box gradient estimation rather than jointly learning the two decisions. As summarized in Table 1, among existing local black-box attacks, only SGAN avoids target-model queries at inference, and none jointly learns *where* and *what* from shared representations. C2FDA instead couples temporal support selection and perturbation synthesis through shared coarse-to-fine representations in a unified diffusion pipeline.

Table 1: Comparison of attacks under transfer- or query-based black-box access.

Method	Black-box	Local Support	No Target Queries	Joint Learning
FGSM	✓	–	✓	–
BIM	✓	–	✓	–
PGD	✓	–	✓	–
Diff-PGD	✓	–	✓	–
ZOO	✓	–	–	–
NES	✓	–	–	–
SPSA	✓	–	–	–
AutoZOOM	✓	–	–	–
AdvGAN	✓	–	✓	–
GATN	✓	–	✓	–
GE-AdvGAN	✓	–	✓	–
TSadv	✓	✓	–	–
BlackTreeS	✓	✓	–	–
SGAN	✓	✓	✓	–
C2FDA	✓	✓	✓	✓

3 METHOD

3.1 PROBLEM FORMULATION

Let $\mathbf{x}$ denote a length-T, C-channel time series with label y. We consider a transfer-based black-box setting with a frozen surrogate classifier f_{s} and a held-out target classifier f_{t}. Their predictions are $\hat{y}_j(\mathbf{x}) = \arg\max_r f_j(\mathbf{x})_r$ for $j \in \{\mathrm{s}, \mathrm{t}\}$. C2FDA is trained using only f_{s}, while f_{t} is used only for final transfer evaluation. Given $\mathbf{x}$, the attack generates $\mathbf{x}_{\mathrm{adv}} = \mathbf{x} + \boldsymbol{\delta}$. A successful untargeted attack satisfies $\hat{y}_{\mathrm{t}}(\mathbf{x}) = y$ and $\hat{y}_{\mathrm{t}}(\mathbf{x}_{\mathrm{adv}}) \neq y$.

We decompose the local perturbation into perturbation content $\boldsymbol{\Delta}$ and temporal support $\mathbf{m}$: $\boldsymbol{\delta} = \mathbf{m} \odot \boldsymbol{\Delta}$, where the temporal mask is shared across channels. Here, $\mathbf{m}$ determines *where* to perturb, while $\boldsymbol{\Delta}$ determines *what* perturbation to generate. The attack satisfies

$$\|\boldsymbol{\Delta}\|_\infty \leq \epsilon, \qquad \|\mathbf{m}\|_0 = k, \tag{1}$$

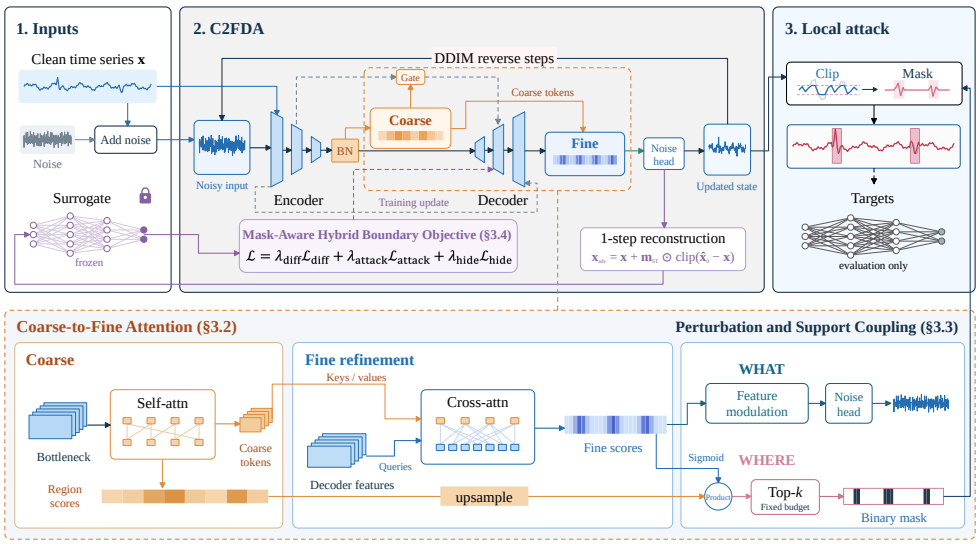


Figure 2: Overview of C2FDA. Coarse attention captures region-level relevance, which fine attention refines into timestamp-level evidence for perturbation synthesis (*what*) and temporal support selection (*where*).

where ϵ is the perturbation magnitude budget and $k = \max\{1, \min(T-1, \lfloor \rho T + 0.5 \rfloor)\}$ is the temporal support budget determined by coverage ratio ρ. Since f_t is unavailable during attack optimization, we use the untargeted surrogate margin $\mathcal{M}_s(\mathbf{x}, y) = f_s(\mathbf{x})_y - \max_{r \neq y} f_s(\mathbf{x})_r$, where $\mathcal{M}_s(\mathbf{x}, y) < 0$ indicates boundary crossing.

C2FDA uses a conditional one-dimensional diffusion U-Net. Given $\mathbf{x}$ and its noisy state $\mathbf{x}^{(\tau)}$, the encoder produces a bottleneck representation $\mathbf{b}$, and the predicted noise reconstructs the clean-state estimate $\hat{\mathbf{x}}_0$. Diffusion and DDIM details are provided in Appendix A.

3.2 Coarse-to-Fine Attention

The attention module in Figure 2 comprises coarse attention and fine attention.

Coarse attention. The coarse branch in Figure 2 pools bottleneck features into K segment descriptors $\mathbf{S} = \text{Pool}_K(\mathbf{b}) \in \mathbb{R}^{K \times d_b}$. With projections $\mathbf{Q}_c = \mathbf{U}_c \mathbf{W}_Q^c + \mathbf{P}_c$, $\mathbf{K}_c = \mathbf{S}\mathbf{W}_K^c$, and $\mathbf{V}_c = \mathbf{S}\mathbf{W}_V^c$, multi-head attention models interactions among segments:

$$\mathbf{Z}_c = \text{LN}\left(\text{MHA}(\mathbf{Q}_c, \mathbf{K}_c, \mathbf{V}_c)\mathbf{W}_O^c\right), \tag{2}$$

where $\mathbf{P}_c$ contains learned segment embeddings. We set $\mathbf{U}_c = \mathbf{S}$ by default and optionally augment it with a surrogate-derived temporal prior.

A linear head produces coarse relevance logits $\boldsymbol{\ell}_c$, normalized as $\mathbf{w}_c = \text{softmax}(\boldsymbol{\ell}_c / \tau_c)$, where τ_c is a temperature parameter. This relevance drives the Gate in Figure 2 to condition an intermediate skip feature.

Coarse relevance remains a soft prior rather than a fixed interval. Fine attention can therefore use context from multiple candidate regions before exact-budget selection. This addresses a limitation of interval-based localization: broadly relevant regions need not contain equally useful timestamps under a small temporal budget.

Fine attention. In the fine branch of Figure 2, let $\mathbf{h} = D(\mathbf{b}, \tilde{\mathbf{e}}_1, \mathbf{e}_0, \tau) \in \mathbb{R}^{T \times d_h}$ denote the full-resolution decoder representation, where $\tilde{\mathbf{e}}_1$ denotes the coarse-conditioned intermediate skip feature. Each timestamp feature attends to the coarse tokens using projections $\mathbf{q}_{f,t} = \mathbf{h}_t \mathbf{W}_Q^f$,

$\mathbf{k}_{\mathrm{f},k} = \mathbf{z}_{\mathrm{c},k}\mathbf{W}_{\mathrm{K}}^{\mathrm{f}}$, and $\mathbf{v}_{\mathrm{f},k} = \mathbf{z}_{\mathrm{c},k}\mathbf{W}_{\mathrm{V}}^{\mathrm{f}}$:

$$\begin{aligned} a_{t,k} &= \mathrm{softmax}_k\left(\frac{\langle \mathbf{q}_{\mathrm{f},t}, \mathbf{k}_{\mathrm{f},k}\rangle}{\sqrt{d_{\mathrm{head}}}} + \log(\max(w_{\mathrm{c},k}, \varepsilon_w))\right), \\ \mathbf{c}_{\mathrm{f},t} &= \sum_{k=1}^{K} a_{t,k}\mathbf{v}_{\mathrm{f},k}. \end{aligned} \tag{3}$$

The log-prior term carries region-level relevance into timestamp-level refinement. A pointwise head produces $\ell_{\mathrm{f}}(t) = \mathrm{MLP}([\mathbf{q}_{\mathrm{f},t}; \mathbf{c}_{\mathrm{f},t}])$. Surrogate-derived temporal guidance and intermediate coarse conditioning are detailed in Appendix A.

3.3 Perturbation and Support Coupling

The two paths in Figure 2 share coarse-to-fine evidence to couple *where* and *what* before masking. Fine logits shape the candidate waveform through feature modulation and rank timestamps for support selection. This lets the generator adapt its perturbation to the locations that remain available, instead of relying on changes later discarded by masking.

Perturbation synthesis: the *what* path. The feature-modulation block in Figure 2 centers the fine logits as $\tilde{\ell}_{\mathrm{f}}(t) = \ell_{\mathrm{f}}(t) - \frac{1}{T}\sum_{j=1}^{T}\ell_{\mathrm{f}}(j)$ and converts them into a feature gain $g_{\mathrm{f}}(t) = 1 + \alpha_{\mathrm{f}}\tanh(\tilde{\ell}_{\mathrm{f}}(t))$, where $0 \le \alpha_{\mathrm{f}} \le 1$. The decoder features are modulated before noise prediction:

$$\mathbf{h}'_t = g_{\mathrm{f}}(t)\mathbf{h}_t, \qquad \hat{\boldsymbol{\eta}}_\theta = \mathrm{EpsHead}(\mathbf{h}'). \tag{4}$$

The bounded perturbation content is $\boldsymbol{\Delta} = \mathrm{clip}(\hat{\mathbf{x}}_0 - \mathbf{x}, -\epsilon, \epsilon)$.

Temporal support selection: the *where* path. The upsampling path in Figure 2 expands coarse relevance to timestamp resolution as $s_{\mathrm{c}}(t) = w_{\mathrm{c},\pi(t)}/\max_j w_{\mathrm{c},j}$, where $\pi(t)$ maps timestamp t to its coarse segment. Fine relevance is $s_{\mathrm{f}}(t) = \mathrm{sigmoid}(\ell_{\mathrm{f}}^{\mathrm{mask}}(t))$, where $\boldsymbol{\ell}_{\mathrm{f}}^{\mathrm{mask}}$ denotes the optionally smoothed fine logits used for support selection.

The Product and Top-k blocks in Figure 2 select exactly k timestamps under the budget in equation 1:

$$s_{\mathrm{joint}}(t) = s_{\mathrm{c}}(t)s_{\mathrm{f}}(t), \qquad m(t) = \mathbb{I}[t \in \mathrm{TopK}(\mathbf{s}_{\mathrm{joint}}, k)]\,. \tag{5}$$

The final local perturbation is $\boldsymbol{\delta} = \mathbf{m}\odot\boldsymbol{\Delta}$, with the temporal mask shared across channels. Fine-logit smoothing and the straight-through relaxation used during optimization are detailed in Appendix A.

3.4 Mask-Aware Hybrid Boundary Objective

The objective block in Figure 2 combines the exact adversarial margin with a first-order estimate around a Gaussian-smoothed neighborhood of the clean input. For $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathrm{sg}}^2\mathbf{I})$, the smoothed surrogate margin and its input gradient are

$$\mu_0 = \mathbb{E}_{\boldsymbol{\xi}}\left[\mathcal{M}_{\mathrm{s}}(\mathbf{x}+\boldsymbol{\xi}, y)\right], \qquad \mathbf{g}_{\mathrm{sg}} = \nabla_{\mathbf{x}}\mu_0. \tag{6}$$

The expectation is estimated from Gaussian perturbations following Smilkov et al. (2017); μ_0 and $\mathbf{g}_{\mathrm{sg}}$ are detached during generator optimization.

Using the final masked perturbation $\boldsymbol{\delta}$,

$$\begin{aligned} \mu_{\mathrm{adv}} = \mathcal{M}_{\mathrm{s}}(\mathbf{x}+\boldsymbol{\delta}, y), \qquad \mu_{\mathrm{tay}} = \mu_0 + \langle \mathbf{g}_{\mathrm{sg}}, \boldsymbol{\delta}\rangle, \\ \mu_\gamma = (1-\gamma)\mu_{\mathrm{adv}} + \gamma\mu_{\mathrm{tay}}, \qquad \gamma \in [0,1]. \end{aligned} \tag{7}$$

The attack loss is

$$\mathcal{L}_{\mathrm{attack}} = \mathbb{E}\left[\mathrm{softplus}\left(\kappa + \mu_\gamma\right)\right], \tag{8}$$

where $\kappa \ge 0$ controls the desired attack confidence. Because the loss evaluates the masked input, changes outside the selected support cannot directly improve the attack objective. Its gradients update perturbation content through noise prediction and support selection through the differentiable mask relaxation, coordinating *where* and *what* around the same boundary-crossing objective.

Noise prediction uses the diffusion loss $\mathcal{L}_{\text{diff}} = \mathbb{E}[\|\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_\theta\|_{\text{F}}^2/(TC)]$, while the perturbation regularization term $\mathcal{L}_{\text{hide}} = \mathbb{E}[\|\boldsymbol{\delta}\|_2]$ penalizes the applied perturbation magnitude. The regularized training objective shown in Figure 2 is

$$\mathcal{L} = \lambda_{\text{diff}}\mathcal{L}_{\text{diff}} + \lambda_{\text{attack}}\mathcal{L}_{\text{attack}} + \lambda_{\text{hide}}\mathcal{L}_{\text{hide}}. \tag{9}$$

Training and adversarial example generation are detailed in Appendix A.

4 Experimental Evaluation

We evaluate C2FDA through comparisons with existing attacks, ablation studies, defense evaluations, and visual and quantitative analyses of perturbation locality.

4.1 Datasets and Experimental Setup

We use 14 time series datasets (3 multivariate and 11 univariate), with a 7:1:2 train–validation–test split and per-variable standardization to zero mean and unit variance. C2FDA uses exact-budget temporal support and a default pointwise magnitude budget of $\epsilon = 0.5$ across all 14 datasets.

A frozen 1D ResNet serves as the surrogate, with LSTM (Sherstinsky, 2020), BiRNN (Lynn et al., 2019), CNN (Kiranyaz et al., 2021), TCN (Bai et al., 2018), and DynamicConv (Wu et al., 2019) as held-out targets. C2FDA, FGSM (Goodfellow et al., 2015), PGD (Madry et al., 2018), MI-FGSM (Dong et al., 2018), GATN (Karim et al., 2021), AdvGAN (Xiao et al., 2018), Diff-PGD (Xue et al., 2023), GE-AdvGAN (Zhu et al., 2024), and SGAN (Yang et al., 2026) use the same surrogate for transfer attacks without target-model queries. NES (Ilyas et al., 2018), SPSA (Uesato et al., 2018), AutoZOOM (Tu et al., 2019), and TSadv (Yang et al., 2022) query targets to generate attacks.

Evaluation metrics. Attack success rate (ASR) is the fraction of initially correctly classified target inputs misclassified after attack (Rathore et al., 2020). Mean squared error (MSE) and dynamic time warping (DTW) measure pointwise and temporal-shape distortion, respectively. Defense success rate (DSR) is the relative ASR reduction after purification (%). Definitions are in Appendix B; dataset statistics, implementation settings, and efficiency results are in Appendix C.

4.2 Attack Performance Comparison

Table 2 compares attacks on three multivariate datasets against five held-out targets. "–" indicates that MSE and DTW are not reported when ASR is zero. C2FDA achieves the highest average ASR on all three. Additional results and distortion rankings are provided in Appendix D.

Global attacks. These methods primarily optimize perturbation content (*what*) over the full sequence. C2FDA achieves higher ASR than query-based AutoZOOM on 14 of 15 dataset–target pairs and ties on the remaining pair, while reducing average DTW on all three datasets. For example, on BasicMotions, average ASR rises from 18.5% to 69.6% and DTW falls from 72.68 to 39.11. AutoZOOM retains lower MSE, which may reflect its ability to refine perturbations through repeated target-model queries (Tu et al., 2019). C2FDA instead learns on a surrogate and must transfer both support (*where*) and content (*what*) to unseen targets. Global transfer baselines also achieve lower average ASR across all three datasets. FGSM and PGD obtain zero ASR on HandMovementDirection. On UWaveGestureLibrary, C2FDA also lowers MSE and DTW compared with AdvGAN, Diff-PGD, and GE-AdvGAN. Its advantage lies in combining higher ASR and lower temporal-shape distortion with localized, target-query-free generation.

Local black-box attacks. TSadv and SGAN constrain *where* through shapelet intervals before searching for or generating *what*. C2FDA achieves higher average ASR than both query-based TSadv and query-free SGAN on all three datasets. On UWaveGestureLibrary, for example, average ASR reaches 69.1%, compared with 18.5% for TSadv and 52.1% for SGAN. Against SGAN, C2FDA also reduces average DTW from 137.27 to 56.53 at similar MSE. TSadv can achieve lower distortion or higher ASR on individual targets, but its gains are less consistent across architectures. The overall improvement supports coordinating timestamp-level support with perturbation content, beyond constraining attacks to shapelet intervals alone.

Table 2: Adversarial attack results on the representative datasets.

Dataset	Attack	BiRNN			LSTM			CNN			TCN			DynamicConv			Avg		
	Method	ASR↑	MSE↓	DTW↓	ASR↑	MSE↓	DTW↓	ASR↑	MSE↓	DTW↓	ASR↑	MSE↓	DTW↓	ASR↑	MSE↓	DTW↓	ASR↑	MSE↓	DTW↓
BasicMotions	FGSM	0.0%	–	–	7.1%	0.0900	167.94	12.5%	0.0900	169.37	0.0%	–	–	0.0%	–	–	3.9%	0.0900	168.66
	PGD	0.0%	–	–	0.0%	–	–	6.2%	0.0675	137.94	0.0%	–	–	0.0%	–	–	1.2%	0.0675	137.94
	MI-FGSM	0.0%	–	–	0.0%	–	–	6.2%	0.0750	147.15	0.0%	–	–	0.0%	–	–	1.2%	0.0750	147.15
	NES	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–
	SPSA	0.0%	–	–	7.1%	0.0527	120.96	43.8%	0.0551	122.06	6.2%	0.0542	121.83	22.2%	0.0551	121.77	15.9%	0.0543	121.66
	AutoZOOM	7.1%	**0.0198**	63.06	7.1%	**0.0165**	57.93	43.8%	**0.0351**	89.95	12.5%	**0.0275**	77.42	22.2%	**0.0281**	75.03	18.5%	**0.0254**	72.68
	GATN	0.0%	–	–	7.1%	0.2612	274.88	12.5%	0.2163	279.39	0.0%	–	–	11.1%	0.2131	288.45	6.1%	0.2302	280.91
	TSadv	35.7%	0.1670	102.60	21.4%	0.1616	99.39	**56.2%**	0.2301	143.78	**37.5%**	0.2229	139.46	22.2%	0.1664	103.99	34.6%	0.1896	117.84
	AdvGAN	0.0%	–	–	7.1%	0.2562	282.29	6.2%	0.2421	291.75	0.0%	–	–	11.1%	0.2597	284.73	4.9%	0.2527	286.26
	Diff-PGD	42.9%	0.3170	314.60	0.0%	–	–	**56.2%**	0.2316	297.58	6.2%	0.3236	349.76	33.3%	0.2867	303.35	27.7%	0.2897	316.32
	GE-AdvGAN	7.1%	0.2778	290.68	7.1%	0.2902	312.70	25.0%	0.2854	291.30	0.0%	–	–	22.2%	0.2693	286.35	12.3%	0.2807	295.26
	SGAN	50.0%	0.1023	187.45	21.4%	0.1244	175.51	**56.2%**	0.1045	151.23	25.0%	0.1341	176.29	66.7%	0.1234	195.28	43.9%	0.1177	177.15
	C2FDA	**92.9%**	0.0450	**33.74**	**78.6%**	0.0364	**35.66**	50.0%	0.0606	**71.74**	**37.5%**	0.0498	**27.20**	**88.9%**	0.0426	**27.22**	**69.6%**	0.0469	**39.11**
HandMovementDirection	FGSM	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–
	PGD	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–
	MI-FGSM	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–
	NES	0.0%	–	–	0.0%	–	–	0.0%	–	–	4.2%	0.0900	403.24	0.0%	–	–	0.8%	0.0900	403.24
	SPSA	4.5%	0.0515	843.88	0.0%	–	–	5.6%	0.0520	841.98	12.5%	0.0518	843.58	12.5%	0.0517	845.07	7.0%	0.0518	843.63
	AutoZOOM	0.0%	–	–	0.0%	–	–	11.1%	**0.0192**	435.56	12.5%	**0.0157**	393.55	12.5%	**0.0131**	360.08	7.2%	**0.0160**	396.40
	GATN	0.0%	–	–	0.0%	–	–	5.6%	0.1417	512.24	0.0%	–	–	6.2%	0.1394	498.53	2.4%	0.1406	505.39
	TSadv	0.0%	–	–	0.0%	–	–	**100.0%**	0.1308	113.82	**100.0%**	0.1112	115.47	68.8%	0.0960	**101.94**	53.8%	0.1127	110.41
	AdvGAN	0.0%	–	–	0.0%	–	–	0.0%	–	–	4.2%	0.1199	518.51	12.5%	0.1126	643.73	3.3%	0.1163	581.12
	Diff-PGD	4.5%	0.1895	760.72	0.0%	–	–	0.0%	–	–	4.2%	0.1966	813.59	6.2%	0.1969	862.05	3.0%	0.1943	812.12
	GE-AdvGAN	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–	12.5%	0.1806	752.09	2.5%	0.1806	752.09
	SGAN	4.5%	0.689	366.41	0.0%	–	–	77.8%	0.1123	337.57	8.3%	0.1093	307.12	37.5%	0.0832	299.83	25.6%	0.2485	327.73
	C2FDA	**81.8%**	**0.0436**	**118.76**	**64.7%**	**0.0526**	**99.64**	94.4%	0.0342	**91.48**	12.5%	0.0632	**104.54**	**87.5%**	0.0281	123.62	**68.2%**	0.0443	**107.61**
UWaveGestureLibrary	FGSM	2.4%	0.0900	203.94	15.7%	0.0900	204.90	28.6%	0.0900	208.97	14.1%	0.0900	212.17	16.9%	0.0900	207.24	15.5%	0.0900	207.44
	PGD	1.2%	0.0404	98.46	14.3%	0.0404	97.36	24.7%	0.0370	97.80	14.1%	0.0356	93.75	10.2%	0.0379	95.56	12.9%	0.0383	96.59
	MI-FGSM	1.2%	0.0544	120.39	17.1%	0.0541	118.98	26.0%	0.0515	121.26	15.4%	0.0508	118.81	13.6%	0.0520	118.10	14.7%	0.0526	119.51
	NES	2.4%	0.0873	255.70	8.6%	0.0900	255.68	18.2%	0.0900	255.29	10.3%	0.0900	255.01	13.6%	0.0900	255.67	10.6%	0.0895	255.47
	SPSA	28.9%	0.0522	158.44	47.1%	0.0524	161.16	89.6%	0.0509	153.60	60.3%	0.0519	154.26	47.5%	0.0522	159.73	54.7%	0.0519	157.44
	AutoZOOM	34.9%	0.0182	81.23	41.4%	**0.0127**	71.49	94.8%	0.0220	83.82	64.1%	0.0228	84.80	55.9%	0.0158	76.94	58.2%	0.0183	79.66
	GATN	3.6%	0.1199	321.99	5.7%	0.1256	329.43	9.1%	0.1258	304.46	6.4%	0.1008	323.51	6.8%	0.1144	320.57	6.3%	0.1173	319.99
	TSadv	1.2%	**0.0040**	**22.86**	0.0%	–	–	57.1%	**0.0038**	**21.65**	30.8%	**0.0036**	**20.95**	3.4%	**0.0025**	**14.17**	18.5%	**0.0035**	**19.91**
	AdvGAN	14.5%	0.1491	330.13	51.4%	0.1661	336.39	49.4%	0.1719	338.72	33.3%	0.1427	368.63	54.2%	0.1683	366.28	40.6%	0.1596	348.03
	Diff-PGD	20.5%	0.1283	282.90	41.4%	0.1407	263.42	59.7%	0.1342	261.71	35.9%	0.1127	241.39	37.3%	0.1394	308.29	39.0%	0.1311	271.54
	GE-AdvGAN	18.1%	0.1339	284.76	37.1%	0.1288	257.69	48.1%	0.1132	233.69	26.9%	0.1311	242.48	25.4%	0.1321	276.80	31.1%	0.1278	259.08
	SGAN	21.7%	0.0903	142.03	57.1%	0.0984	148.81	76.6%	0.0896	140.78	47.4%	0.0745	121.80	57.6%	0.0738	132.92	52.1%	0.0853	137.27
	C2FDA	**38.6%**	0.0973	60.19	**64.3%**	0.0819	**60.15**	**97.4%**	0.0845	48.21	**65.4%**	0.0935	55.32	**79.7%**	0.0807	58.79	**69.1%**	0.0876	56.53

Statistical comparison of ASR. We assess overall attack effectiveness across 14 datasets using the Friedman test ($p < 10^{-15}$), followed by the Nemenyi post-hoc test ($\alpha = 0.05$). As shown in Figure 3, C2FDA achieves the best (lowest) average rank of 1.64, ahead of AutoZOOM (2.29), SGAN (5.25), and TSadv (7.86). With a critical difference of 4.88, C2FDA significantly outperforms TSadv, Diff-PGD (6.68), AdvGAN (7.14), and GE-AdvGAN (7.57). C2FDA achieves the leading average ASR rank without target-model queries, whereas AutoZOOM relies on repeated queries to optimize each attack.

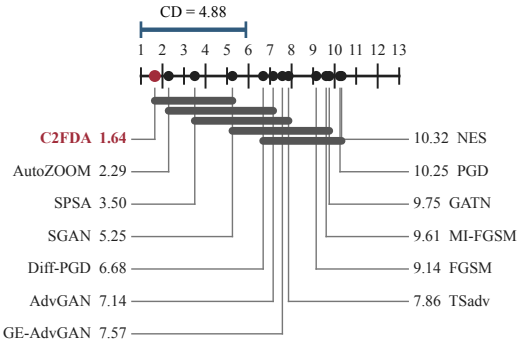


Figure 3: Critical-difference diagram for ASR across 14 datasets.

4.3 Ablation Experiments

Effectiveness of coarse-to-fine attention. Table 3 compares the full model (**Full Model**) with five ablations. Removing coarse attention (**w/o Coarse**) generally lowers ASR and increases distortion, while removing fine attention (**w/o Fine**) reduces ASR on all three datasets. On HandMovementDi-

Table 3: Ablation results on the representative datasets against CNN. Each entry reports ASR (%) / MSE / DTW.

Dataset	Full Model	w/o Coarse	w/o Fine	w/o C2F Attn.	Random Intervals	Exact Margin Only
BasicMotions	50.0 / 0.0606 / **71.74**	50.0 / 0.0897 / 97.83	31.2 / **0.0504** / 76.46	**56.2** / 0.1915 / 239.27	31.2 / 0.0979 / 102.61	43.8 / 0.0872 / 94.38
HandMovementDirection	**94.4** / 0.0342 / 91.48	77.8 / 0.0608 / 103.72	66.7 / **0.0302** / **90.34**	88.9 / 0.2046 / 504.85	50.0 / 0.0573 / 143.72	83.3 / 0.0415 / 97.63
UWaveGestureLibrary	**97.4** / 0.0845 / **48.21**	89.6 / 0.1024 / 57.67	76.6 / **0.0842** / 53.20	84.4 / 0.1490 / 248.41	49.4 / 0.0962 / 56.38	87.0 / 0.1183 / 61.85

rection, ASR falls from 94.4% to 77.8% and 66.7%, respectively. Coarse context and fine refinement thus provide complementary evidence for *where* and *what*.

Removing both attention stages (**w/o C2F Attn.**) substantially increases MSE and DTW across all three datasets, even when ASR improves. On BasicMotions, ASR rises from 50.0% to 56.2%, but DTW increases from 71.74 to 239.27. The shared attention mechanism therefore improves the balance between attack effectiveness and distortion.

Replacing learned support with random intervals (**Random Intervals**) lowers ASR and increases both distortion metrics on all three datasets. For example, ASR on HandMovementDirection drops from 94.4% to 50.0%. The result highlights the importance of sample-specific support selection: restricting coverage alone does not ensure that perturbations affect decision-relevant timestamps.

Effect of the hybrid boundary objective. The variant using only the exact margin (**Exact Margin Only**) removes the first-order component ($\gamma = 0$). Compared with this variant, the full model achieves higher ASR and lower MSE and DTW on all three datasets. On HandMovementDirection, for example, ASR improves from 83.3% to 94.4%, while MSE and DTW decrease by 17.6% and 6.3%, respectively. Thus, the hybrid objective improves the balance between attack success and distortion on these datasets. Evaluating both margin components on the final masked perturbation ties this boundary guidance to the locations and perturbation content actually applied, jointly refining *where* and *what*. Results on the remaining datasets are provided in Appendix Table 9.

4.4 Defense

We evaluate robustness against a time series adaptation of collaborative defense-GAN (Laykaviriyakul & Phaisangittisagul, 2023) on CNN, reporting pre-purification ASR / DSR (%) in Table 4.

C2FDA matches or exceeds SGAN's residual ASR on all three datasets while undergoing smaller relative ASR reductions. On UWaveGestureLibrary, for example, post-purification ASR is 87.0% for C2FDA versus 55.8% for SGAN. These results show that the jointly learned support and perturbation content retain substantial attack effectiveness after purification. Additional results appear in Appendix Table 10.

Table 4: Defense results on representative datasets (ASR / DSR, %).

Method	BasicMotions ASR / DSR	HMD ASR / DSR	UWave ASR / DSR
FGSM	12.5 / 50.4	0.0 / –	28.6 / 72.7
PGD	6.2 / 0.0	0.0 / –	24.7 / 78.9
MI-FGSM	6.2 / 0.0	0.0 / –	26.0 / 70.0
NES	0.0 / –	0.0 / –	18.2 / 64.3
SPSA	43.8 / 42.9	5.6 / 0.0	89.6 / 66.6
AutoZOOM	43.8 / 42.9	11.1 / 49.5	94.8 / 53.4
GATN	12.5 / 50.4	5.6 / 100.0	9.1 / 100.0
TSadv	56.2 / 22.1	100.0 / 27.8	57.1 / 18.0
AdvGAN	6.2 / 100.0	0.0 / –	49.4 / 94.7
Diff-PGD	56.2 / 66.5	0.0 / –	59.7 / 76.0
GE-AdvGAN	25.0 / 50.0	0.0 / –	48.1 / 100.0
SGAN	56.2 / 33.3	77.8 / 50.0	76.6 / 27.2
C2FDA	50.0 / 25.0	94.4 / 17.6	97.4 / 10.7

4.5 Case Study on Localized Perturbations

Figure 4 shows coarse-to-fine attention and the final DDIM support on SonyAIBORobotSurface1. Coarse attention identifies candidate regions; fine attention refines them into timestamp-level evidence, placing perturbations (*where*) around a salient waveform transition.

Figure 5 compares attacks on the same Car sequence. AutoZOOM introduces small deviations across broad regions, while TSadv and SGAN concentrate changes within intervals. C2FDA couples selected locations (*where*) with small deviations (*what*) near waveform transitions, preserving long regions elsewhere. Additional qualitative examples are provided in Appendix Figures 10 and 12.

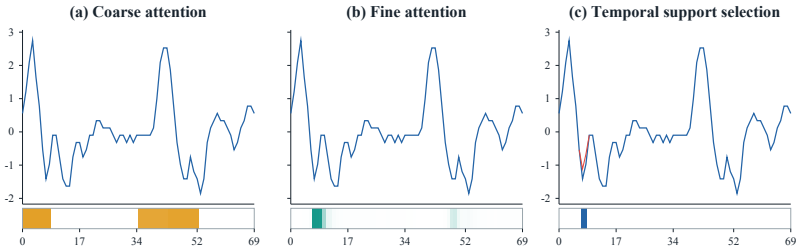


Figure 4: Coarse-to-fine attention and temporal support selection on SonyAIBORobotSurface1.

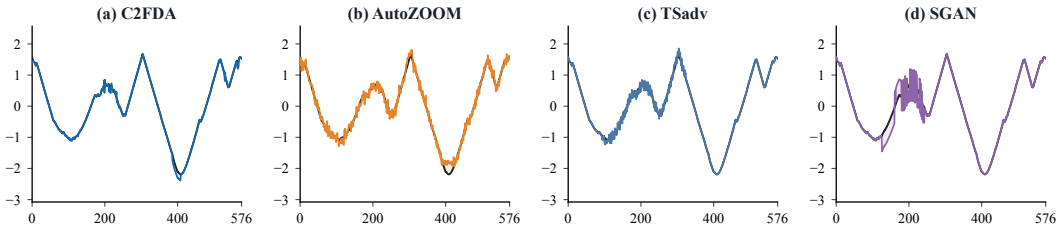


Figure 5: Adversarial examples on Car. Black: clean input; colored: adversarial input.

4.6 Temporal Coverage and Attack Success

Figure 6 compares mean temporal coverage and ASR across 14 datasets and five targets. Coverage is the fraction of timestamps where perturbations are allowed, counting overlaps once; global attacks have 100% support. ASR is averaged equally over the 70 dataset–target pairs.

C2FDA combines the highest mean ASR (68.61%) with the smallest coverage (14.64%). Against full-sequence AutoZOOM (59.04% ASR), it improves ASR by 9.6 percentage points without target queries. Compared with local baselines, C2FDA uses less than half of SGAN's support (32.83%) and approximately one third of TSadv's (43.46%), while achieving higher ASR than both (38.56% and 26.80%, respectively). Shapelet intervals restrict *where* attacks occur, but can still allocate support to less useful timestamps. C2FDA instead coordinates timestamp selection with perturbation synthesis through shared coarse-to-fine representations. This joint *where–what* learning improves attack effectiveness while confining perturbations to a smaller temporal support. Coverage captures temporal locality, which low MSE alone does not ensure.

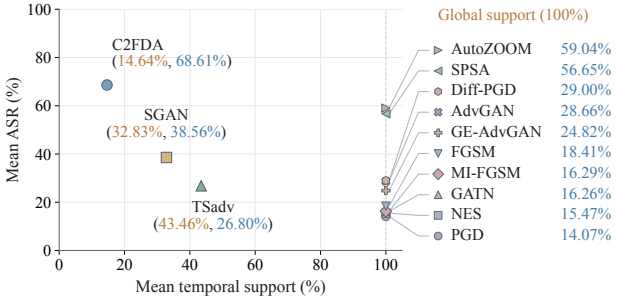


Figure 6: Mean ASR versus temporal support. Blue and orange values denote ASR and support, respectively.

5 Conclusion

We presented C2FDA, a diffusion-based framework for local black-box attacks that couples *where* to perturb with *what* perturbation to generate. Shared coarse-to-fine attention selects exact-budget temporal support and shapes perturbation content, while a mask-aware hybrid boundary objective jointly optimizes both decisions through the applied perturbation. Experiments across 14 datasets and five target architectures demonstrate strong transferability, low distortion, and efficient generation without target-model queries. Future work will explore surrogate ensembles to improve cross-model transfer, sample-adaptive support and magnitude budgets, and evaluation against adaptive defenses.

AI USE STATEMENT

Generative AI tools were used solely for grammar checking and language polishing of the manuscript. They were not used in the design or conduct of experiments, analysis of results, or any other part of the research.

ETHICS STATEMENT

This work studies localized adversarial attacks on time series classifiers to support robustness evaluation and defense research. Experiments use benchmark datasets and offline classifier evaluations. The proposed method has dual-use implications: it can help identify model vulnerabilities but could also be misused to manipulate predictions. We advocate its use only for authorized robustness assessment and defense development. Applications involving health or safety require additional evaluation under realistic operating conditions.

REPRODUCIBILITY STATEMENT

Section 3 describes the C2FDA architecture, perturbation and support coupling, and training objective. Appendix A provides additional method details, including Algorithm 1 and the adversarial generation procedure. Appendix B defines the evaluation metrics and aggregation rules. Appendix C reports dataset splits, preprocessing, classifier and attack settings, and the software and hardware environment. Appendix D provides additional experimental results and analyses.

REFERENCES

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 15–26, 2017.

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.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193, 2018.

Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Adversarial attacks on deep neural networks for time series classification. In *International Joint Conference on Neural Networks*, pp. 1–8, 2019.

Navid Mohammadi Foumani, Lynn Miller, Chang Wei Tan, Geoffrey I. Webb, Germain Forestier, and Mahsa Salehi. Deep learning for time series classification and extrinsic regression: A current survey. *ACM Computing Surveys*, 56(9):217:1–217:45, 2024.

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

Nan Huang, Haishuai Wang, Zihuai He, Marinka Zitnik, and Xiang Zhang. Repurposing foundation model for generalizable medical time series classification. In *International Conference on Learning Representations*, 2026.

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

Yoo-Min Jung and Leekyung Kim. MambaSL: Exploring single-layer mamba for time series classification. In *International Conference on Learning Representations*, 2026.

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.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J. Inman. 1D convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing*, 151:107398, 2021.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*, pp. 99–112. Chapman and Hall/CRC, 2018.

Pranpaveen Laykaviriyakul and Ekachai Phaisangittisagul. Collaborative defense-GAN for protecting adversarial attacks on classification system. *Expert Systems with Applications*, 214:118957, 2023.

Htet Myet Lynn, Sung Bum Pan, and Pankoo Kim. A deep bidirectional GRU network model for biometric electrocardiogram classification based on recurrent neural networks. *IEEE Access*, 7: 145395–145405, 2019.

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.

Gautier Pialla, Hassan Ismail Fawaz, Maxime Devanne, Jonathan Weber, Lhassane Idoumghar, Pierre-Alain Muller, Christoph Bergmeir, Daniel F. Schmidt, Geoffrey I. Webb, and Germain Forestier. Time series adversarial attacks: An investigation of smooth perturbations and defense approaches. *International Journal of Data Science and Analytics*, 19(1):129–139, 2025.

Pradeep Rathore, Arghya Basak, Sri Harsha Nistala, and Venkataramana Runkana. Untargeted, targeted and universal adversarial attacks and defenses on time series. In *International Joint Conference on Neural Networks*, pp. 1–8. IEEE, 2020.

Alex Sherstinsky. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: Removing noise by adding noise, 2017.

Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. AutoZOOM: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 742–749, 2019.

Jonathan Uesato, Brendan O'Donoghue, Pushmeet Kohli, and Aaron van den Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *International Conference on Machine Learning*, pp. 5025–5034. PMLR, 2018.

Felix Wu, Angela Fan, Alexei Baevski, Yann N. Dauphin, and Michael Auli. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*, 2019.

Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 3905–3911, 2018.

Shifeng Xie, Vasilii Feofanov, Jianfeng Zhang, Themis Palpanas, and Ievgen Redko. CauKer: Classification time series foundation models can be pretrained on synthetic data. In *International Conference on Learning Representations*, 2026.

Haotian Xue, Alexandre Araujo, Bin Hu, and Yongxin Chen. Diffusion-based adversarial sample generation for improved stealthiness and controllability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 2894–2921, 2023.

Kaiyu Yang, Jidong Yuan, Haoyu Yan, Haiyang Liu, and Yongqi Sun. SGAN: Shapelet-based GAN for local time series black-box attack. *Pattern Recognition*, 179:113635, 2026.

Wenbo Yang, Jidong Yuan, Xiaokang Wang, and Peixiang Zhao. TSadv: Black-box adversarial attack on time series with local perturbations. *Engineering Applications of Artificial Intelligence*, 114:105218, 2022.

Lexiang Ye and Eamonn Keogh. Time series shapelets: A new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 947–956, 2009.

Zhiyu Zhu, Huaming Chen, Xinyi Wang, Jiayu Zhang, Zhibo Jin, Kim-Kwang Raymond Choo, Jun Shen, and Dong Yuan. GE-AdvGAN: Improving the transferability of adversarial samples by gradient editing-based adversarial generative model. In *SIAM International Conference on Data Mining*, pp. 706–714. SIAM, 2024.

A Additional Method Details

Conditional Diffusion Generation

C2FDA uses a conditional one-dimensional diffusion model. At diffusion step τ, the clean sequence is corrupted as

$$\mathbf{x}^{(\tau)} = \sqrt{\bar{\alpha}_\tau}\,\mathbf{x} + \sqrt{1-\bar{\alpha}_\tau}\,\boldsymbol{\eta}, \qquad \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{10}$$

where $\alpha_\tau = 1 - \beta_\tau$ and $\bar{\alpha}_\tau = \prod_{s=0}^{\tau} \alpha_s$.

The conditional U-Net receives $[\mathbf{x}; \mathbf{x}^{(\tau)}]$ together with an embedding of τ. Its encoder extracts multi-resolution temporal features and produces the low-resolution bottleneck representation $\mathbf{b}$, while the decoder restores full-resolution features for noise prediction.

Given the predicted noise $\hat{\boldsymbol{\eta}}_\theta$, the corresponding clean-state estimate is

$$\hat{\mathbf{x}}_0 = \frac{\mathbf{x}^{(\tau)} - \sqrt{1-\bar{\alpha}_\tau}\,\hat{\boldsymbol{\eta}}_\theta}{\sqrt{\bar{\alpha}_\tau}}. \tag{11}$$

The residual $\hat{\mathbf{x}}_0 - \mathbf{x}$ is used to construct the bounded perturbation content in Section 3.3.

Surrogate-Derived Temporal Guidance

The smoothed quantities in equation 6 are estimated with M_{sg} Gaussian samples as

$$\mu_0 \approx \frac{1}{M_{\mathrm{sg}}} \sum_{i=1}^{M_{\mathrm{sg}}} \mathcal{M}_{\mathrm{s}}\left(\mathbf{x} + \boldsymbol{\xi}^{(i)}, y\right), \qquad \boldsymbol{\xi}^{(i)} \sim \mathcal{N}\left(\mathbf{0}, \sigma_{\mathrm{sg}}^2 \mathbf{I}\right), \tag{12}$$

and

$$\mathbf{g}_{\mathrm{sg}} \approx \frac{1}{M_{\mathrm{sg}}} \sum_{i=1}^{M_{\mathrm{sg}}} \nabla_{\mathbf{x}} \mathcal{M}_{\mathrm{s}}\left(\mathbf{x} + \boldsymbol{\xi}^{(i)}, y\right). \tag{13}$$

To normalize gradient scales across channels, we define the timestamp-level gradient sensitivity as

$$r_t = \left[\frac{1}{C}\sum_{c=1}^{C}\left(\frac{g_{\mathrm{sg},t,c}}{\max\left(\max_u |g_{\mathrm{sg},u,c}|, \varepsilon_g\right)}\right)^2\right]^{1/2}. \tag{14}$$

Local temporal variation is measured by

$$\begin{aligned} \bar{v}_t &= \frac{1}{C}\sum_{c=1}^{C} \mathrm{Var}_{u\in\mathcal{N}(t)}(x_{u,c}), \\ v_t &= \frac{\bar{v}_t}{\max\left(\max_j \bar{v}_j, \varepsilon_v\right)}, \end{aligned} \tag{15}$$

where $\mathcal{N}(t)$ denotes a local temporal neighborhood around timestamp t. The two signals are combined as

$$\psi_t = (1-\lambda_a) r_t + \lambda_a v_t, \qquad \lambda_a \in [0,1]. \tag{16}$$

For coarse attention, the guidance signal is standardized and pooled to segment resolution:

$$\boldsymbol{\Psi}_K = \mathrm{Pool}_K\left(\mathrm{Standardize}(\boldsymbol{\psi})\right) \in \mathbb{R}^{K\times 1}. \tag{17}$$

For coarse attention, the query input is

$$\mathbf{U}_{\mathrm{c}} = \begin{cases} \mathrm{Concat}_{\mathrm{feat}}\left(\mathbf{S}, \boldsymbol{\Psi}_K\right), & \text{with temporal guidance,} \\ \mathbf{S}, & \text{otherwise.} \end{cases} \tag{18}$$

The same temporal prior provides an additive bias to the fine relevance logits:

$$\ell_{\mathrm{f}}(t) \leftarrow \ell_{\mathrm{f}}(t) + \mathrm{softplus}(b_{\mathrm{sal}})\left[\mathrm{MinMax}(\boldsymbol{\psi})\right]_t, \tag{19}$$

where

$$[\mathrm{MinMax}(\boldsymbol{\psi})]_t = \frac{\psi_t - \min_u \psi_u}{\max\left(\max_u \psi_u - \min_u \psi_u, \varepsilon_\psi\right)}. \tag{20}$$

Here, b_{sal} controls the fine-level guidance strength, while ε_g, ε_v, and ε_ψ provide numerical stability.

Both μ_0 and $\mathbf{g}_{\mathrm{sg}}$ are detached before generator optimization. During adversarial example generation, the clean surrogate prediction $\hat{y}_{\mathrm{s}}(\mathbf{x})$ replaces the ground-truth label when constructing the temporal guidance prior.

Coarse Conditioning and Differentiable Support Selection

Intermediate coarse conditioning. Let $\mathbf{e}_1$ denote an intermediate encoder skip feature with temporal length L_1. Its temporal position u is mapped to a coarse segment by

$$\pi_1(u) = \min\left\{K, \left\lfloor \frac{uK}{L_1} \right\rfloor + 1\right\}, \qquad u = 0, \ldots, L_1 - 1. \tag{21}$$

The expanded coarse relevance is $\omega_{\mathrm{c}}^{(1)}(u) = w_{\mathrm{c},\pi_1(u)}$, and the conditioned skip feature is

$$\tilde{\mathbf{e}}_1(u) = \mathbf{e}_1(u)\left[1 + \alpha_{\mathrm{c}} \tanh\left(\frac{\omega_{\mathrm{c}}^{(1)}(u)}{\max\left(L_1^{-1}\sum_{v=0}^{L_1-1}\omega_{\mathrm{c}}^{(1)}(v), \varepsilon_{\mathrm{c}}\right)}\right)\right], \tag{22}$$

where $\alpha_{\mathrm{c}} \geq 0$ controls the conditioning strength and ε_{c} provides numerical stability. The full-resolution skip feature $\mathbf{e}_0$ remains unchanged.

Timestamp mapping and fine-logit smoothing. Using zero-based timestamp indexing, timestamp t is mapped to its corresponding coarse segment as

$$\pi(t) = \min\left\{K, \left\lfloor \frac{tK}{T} \right\rfloor + 1\right\}, \qquad t = 0, \ldots, T - 1. \tag{23}$$

For temporal support selection, the fine logits are optionally smoothed along the temporal axis:

$$\mathcal{S}_q(\boldsymbol{\ell}_{\mathrm{f}}) = \begin{cases} \mathrm{AvgPool1D}_q(\boldsymbol{\ell}_{\mathrm{f}}), & q > 1, \\ \boldsymbol{\ell}_{\mathrm{f}}, & q = 1, \end{cases} \tag{24}$$

where q is an odd kernel width. The support-selection logits are

$$\boldsymbol{\ell}_{\mathrm{f}}^{\mathrm{mask}} = \mathcal{S}_q(\boldsymbol{\ell}_{\mathrm{f}}). \tag{25}$$

The original logits $\boldsymbol{\ell}_{\mathrm{f}}$ are retained for perturbation synthesis.

Straight-through exact-budget selection. The binary Top-k mask in equation 5 is used in the forward computation. For gradient propagation, the joint temporal score is projected onto the capped simplex

$$\mathcal{C}_k = \left\{\mathbf{u} \in [0,1]^T : \sum_{t=0}^{T-1} u_t = k\right\}. \tag{26}$$

The continuous relaxation is

$$\widetilde{\mathbf{m}} = \Pi_{\mathcal{C}_k}\left(\mathbf{s}_{\mathrm{joint}}\right), \tag{27}$$

and the straight-through estimator is

$$\mathbf{m}_{\mathrm{ST}} = \mathbf{m} + \widetilde{\mathbf{m}} - \mathrm{stopgrad}\left(\widetilde{\mathbf{m}}\right). \tag{28}$$

During generator optimization, $\mathbf{m}_{\mathrm{ST}}$ is used in place of $\mathbf{m}$ when applying $\boldsymbol{\delta} = \mathbf{m} \odot \boldsymbol{\Delta}$. During adversarial example generation, the binary mask $\mathbf{m}$ is used directly. The forward value of $\mathbf{m}_{\mathrm{ST}}$ equals $\mathbf{m}$, while its backward gradient follows $\widetilde{\mathbf{m}}$.

Optimization and Gradient Routes

The attack loss in equation 8 is evaluated on the final masked perturbation defined in Section 3.3.

Along the perturbation-synthesis path, $\mathcal{L}_{\text{attack}}$ propagates through equation 4, updating the fine feature modulation, predicted diffusion noise, and perturbation content $\mathbf{\Delta}$.

Along the temporal support-selection path, gradients propagate through the straight-through estimator in equation 28. The coarse stage is updated through intermediate coarse conditioning, fine cross-attention, and the coarse component of s_{joint}.

The smoothed quantities μ_0 and $\mathbf{g}_{\text{sg}}$ are treated as constants during generator optimization. The diffusion loss $\mathcal{L}_{\text{diff}}$ supervises noise prediction, while the perturbation regularization term $\mathcal{L}_{\text{hide}}$ penalizes the magnitude of the final masked perturbation.

Training and Adversarial Example Generation

Algorithm 1 summarizes one generator update.

Algorithm 1 One C2FDA Training Step

Require: Clean series $\mathbf{x}$, label y, frozen surrogate f_{s}, generator G_θ
1: Sample τ and $\boldsymbol{\eta}$; construct $\mathbf{x}^{(\tau)}$ using equation 10.
2: If $\gamma > 0$ or temporal guidance is enabled, estimate μ_0 and $\mathbf{g}_{\text{sg}}$.
3: If temporal guidance is enabled, construct the temporal prior $\boldsymbol{\psi}$.
4: Encode $[\mathbf{x}; \mathbf{x}^{(\tau)}]$ and compute $\mathbf{Z}_{\text{c}}$ and $\mathbf{w}_{\text{c}}$.
5: Apply coarse conditioning and decode the full-resolution representation $\mathbf{h}$.
6: Compute the fine logits $\boldsymbol{\ell}_{\text{f}}$.
7: Modulate $\mathbf{h}$ using the fine logits and predict $\hat{\boldsymbol{\eta}}_\theta$.
8: Reconstruct $\hat{\mathbf{x}}_0$ and compute the bounded perturbation content $\mathbf{\Delta}$.
9: Compute $\mathbf{s}_{\text{joint}}$, the binary Top-$k$ mask $\mathbf{m}$, and $\mathbf{m}_{\text{ST}}$.
10: Apply $\boldsymbol{\delta} = \mathbf{m}_{\text{ST}} \odot \mathbf{\Delta}$ and obtain $\mathbf{x}_{\text{adv}} = \mathbf{x} + \boldsymbol{\delta}$.
11: Compute losses $\mathcal{L}_{\text{diff}}$ and $\mathcal{L}_{\text{attack}}$ and regularizer $\mathcal{L}_{\text{hide}}$; update θ using equation 9.

Input-anchored DDIM generation. Adversarial example generation uses deterministic DDIM sampling conditioned on the clean sequence. Let $\tau_1 > \tau_2 > \cdots > \tau_L = 0$ denote the selected reverse timesteps. The initial noisy state is

$$\mathbf{x}^{(\tau_1)} = \sqrt{\bar{\alpha}_{\tau_1}}\mathbf{x} + \sqrt{1 - \bar{\alpha}_{\tau_1}}\boldsymbol{\eta}_s, \qquad \boldsymbol{\eta}_s \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{29}$$

At each reverse step τ_i, the network recomputes the coarse tokens $\mathbf{Z}_{\text{c}}$, coarse relevance weights $\mathbf{w}_{\text{c}}$, fine logits $\boldsymbol{\ell}_{\text{f}}$, and feature-modulation gain before predicting $\hat{\boldsymbol{\eta}}_\theta^{(\tau_i)}$ and $\hat{\mathbf{x}}_0^{(\tau_i)}$.

For $i < L$, the deterministic DDIM update is

$$\mathbf{x}^{(\tau_{i+1})} = \sqrt{\bar{\alpha}_{\tau_{i+1}}}\hat{\mathbf{x}}_0^{(\tau_i)} + \sqrt{1 - \bar{\alpha}_{\tau_{i+1}}}\hat{\boldsymbol{\eta}}_\theta^{(\tau_i)}. \tag{30}$$

When temporal guidance is enabled, the clean surrogate prediction $\hat{y}_{\text{s}}(\mathbf{x})$ is used as the pseudo-label. The temporal prior is computed once for each input and kept fixed over the reverse trajectory.

After the final reverse step, the latest coarse relevance and fine logits are used to compute the joint temporal score and the binary Top-k mask $\mathbf{m}$. The final perturbation is obtained from $\boldsymbol{\delta} = \mathbf{m} \odot \mathbf{\Delta}$, giving $\mathbf{x}_{\text{adv}} = \mathbf{x} + \boldsymbol{\delta}$.

B Evaluation Metric Definitions

Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ denote the set of samples used for attack evaluation, where $\mathbf{x}_i \in \mathbb{R}^{T \times C}$. For a target classifier f_{t}, we define the subset of clean samples that are correctly classified as

$$\mathcal{C}_{\text{t}} = \{i : \hat{y}_{\text{t}}(\mathbf{x}_i) = y_i\}. \tag{31}$$

The attack success rate is defined as

$$\mathrm{ASR}_{\mathrm{t}} = \frac{1}{|\mathcal{C}_{\mathrm{t}}|} \sum_{i \in \mathcal{C}_{\mathrm{t}}} \mathbb{I}\left[\hat{y}_{\mathrm{t}}(\mathbf{x}_{i,\mathrm{adv}}) \neq y_i\right], \tag{32}$$

where $\mathbb{I}[\cdot]$ is the indicator function.

For purification,

$$\mathrm{DSR} = \frac{\mathrm{ASR}_{\mathrm{pre}} - \mathrm{ASR}_{\mathrm{post}}}{\mathrm{ASR}_{\mathrm{pre}}} \times 100\%, \tag{33}$$

where pre- and post-purification ASRs use the same set $\mathcal{C}_{\mathrm{t}}$. DSR is rounded to one decimal and reported as "–" when $\mathrm{ASR}_{\mathrm{pre}} = 0$.

Mean squared error measures the average pointwise distortion between the clean and adversarial series:

$$\mathrm{MSE} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{TC} \sum_{p=1}^{T} \sum_{c=1}^{C} \left(x_{i,p,c} - x_{i,\mathrm{adv},p,c}\right)^2. \tag{34}$$

For DTW, each channel is aligned independently. For sample i and channel c, the accumulated alignment cost is recursively defined as

$$D_{i,c}(p,q) = d(x_{i,p,c}, x_{i,\mathrm{adv},q,c}) + \min\left\{D_{i,c}(p-1,q), D_{i,c}(p,q-1), D_{i,c}(p-1,q-1)\right\}, \tag{35}$$

where $d(a,b) = |a-b|$ is the absolute pointwise distance between scalar channel values. Accumulated costs are used without taking a square root or normalizing by path length. The boundary conditions are $D_{i,c}(0,0) = 0$ and $D_{i,c}(p,0) = D_{i,c}(0,q) = +\infty$ for $p, q > 0$.

Channel-wise costs are summed as $\mathrm{DTW}_i = \sum_{c=1}^{C} D_{i,c}(T,T)$ and averaged over all evaluated samples:

$$\mathrm{DTW} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{DTW}_i. \tag{36}$$

In the result tables, MSE and DTW are marked "–" when ASR is zero. Avg ASR includes all five targets; Avg MSE and DTW average only reported values and remain "–" if none are available.

C Additional Experimental Details

Dataset Splits and Preprocessing

Table 5 reports split sizes, sequence lengths, numbers of variables, and class counts for all 14 datasets. Each dataset uses a 7:1:2 train–validation–test split. The training set is used for classifier and attack training, the validation set for model selection, and the test set for final evaluation. Each variable is standardized independently to zero mean and unit variance.

Classifier and Attack Configuration

The pretrained one-dimensional ResNet surrogate has channel widths $(64, 128, 128, 128, 128)$ and kernel sizes $(7, 5, 3, 5, 3)$. The five target classifiers—LSTM, BiRNN, CNN, TCN, and DynamicConv—each use a hidden size of 20. All classifiers use a learning rate of 0.005. ResNet, CNN, and DynamicConv are trained with Adam (Kingma & Ba, 2015), while the recurrent models use RMSProp. The surrogate remains frozen during C2FDA training, and target-model parameters and gradients are never accessed by transfer attacks.

Table 6 lists the default C2FDA configuration unless an experiment explicitly varies a setting. The hybrid boundary objective blends the exact and first-order margins using γ; the objective weights balance the diffusion and attack losses with perturbation regularization.

Runtime environment. Experiments are conducted on Ubuntu 20.04.4 LTS using an NVIDIA GeForce RTX 5090 GPU with 32 GB of memory and an Intel Xeon Platinum 8470Q CPU. The implementation uses PyTorch 2.11.0 with CUDA 12.8.

Table 5: Dataset statistics and train–validation–test splits.

Dataset	Train/Val/Test	$T \times C$	Classes
BasicMotions	56/8/16	100×6	4
HandMovementDirection	164/23/47	400×10	4
UWaveGestureLibrary	308/44/88	315×3	8
Car	84/12/24	577×1	4
CinCECGTorso	994/142/284	$1,639 \times 1$	4
ECG5000	3,500/500/1,000	140×1	5
ECGFiveDays	619/88/177	136×1	2
Earthquakes	323/46/92	512×1	2
FordA	3,445/492/984	500×1	2
FordB	3,112/445/889	500×1	2
Trace	140/20/40	275×1	4
ItalyPowerDemand	767/110/219	24×1	2
SonyAIBORobotSurface1	435/62/124	70×1	2
SonyAIBORobotSurface2	686/98/196	65×1	2

Table 6: Experimental hyperparameter settings.

Hyperparameter	Value
Perturbation budget ϵ	0.5
Diffusion steps	100
DDIM reverse steps	20
Taylor blend γ	0.4
Diffusion weight λ_{diff}	1.0
Attack weight λ_{attack}	5.0
Perturbation regularization weight λ_{hide}	1.0

Comparison of Time Efficiency

To evaluate computational efficiency, Table 7 compares trainable attack-generator parameters, one-time training cost, and complete-test-set generation time. C2FDA contains 2.14M trainable parameters, smaller than AdvGAN (3.40M) and Diff-PGD (3.62M), and requires 4.66 minutes for one-time training. Once trained, it generates a complete test set in 1.58 seconds. Compared with query- and search-based attacks, C2FDA avoids repeated target-model queries and per-input search while retaining second-scale generation. Although SGAN is more compact and GE-AdvGAN achieves faster generation, C2FDA provides a stronger overall trade-off between efficiency and attack performance, achieving high ASR together with low MSE and DTW. It therefore combines moderate model size and training cost with target-query-free generation, explicit local support, and low-distortion transferable attacks.

Parameter Analysis

We further analyze the sensitivity of C2FDA to three key hyperparameters: the perturbation budget ϵ, the attack-loss weight λ_{attack}, and the perturbation-regularization weight λ_{hide}. The analysis is conducted on five representative datasets using CNN as the target classifier. In each sweep, only the corresponding hyperparameter is varied, while all remaining settings follow Table 6. Both ASR and MSE are reported to characterize the trade-off between attack effectiveness and perturbation distortion.

Effect of the perturbation budget. Figure 7 shows that increasing ϵ initially improves attack success substantially, since a larger amplitude budget provides greater freedom for the generated perturbation to cross the decision boundary. The most pronounced improvement occurs between $\epsilon = 0.2$ and $\epsilon = 0.5$. For example, ASR increases from 33.8% to 97.4% on UWaveGestureLibrary

Table 7: Comparison of model parameters and time consumption.

Method	Params (M)	Initial process	Avg. complete-test-set generation time
FGSM	–	–	**0.17 s**†
PGD	–	–	2.48 s†
MI-FGSM	–	–	2.80 s†
NES	–	–	0.18 s†
SPSA	–	–	27.17 min†*
AutoZOOM	–	–	24.57 min†*
GATN	1.46	6.02 min	0.72 s
TSadv	–	–	5.46 h†
AdvGAN	3.40	4.96 min	0.79 s
Diff-PGD	3.62	25.81 min‡	3.74 min‡
GE-AdvGAN	1.13	5.83 min‡	0.53 s‡
SGAN	0.10	**2.74 min**‡	0.98 s‡
C2FDA	2.14	4.66 min	1.58 s†

Params count persistent trainable attack-generator weights; frozen surrogate and target classifiers are excluded. "–" denotes that no trainable attack generator is used or that the corresponding stage is not applicable. †Average complete-test-set generation time. C2FDA, FGSM, PGD, MI-FGSM, and NES use generation-only timing, while TSadv excludes shapelet discovery. *Historical full-command wall-clock time, including loading, metric computation, and serialization, and therefore represents a conservative upper bound. ‡Values reported by the SGAN study. C2FDA's initial process corresponds to its one-time generator training.

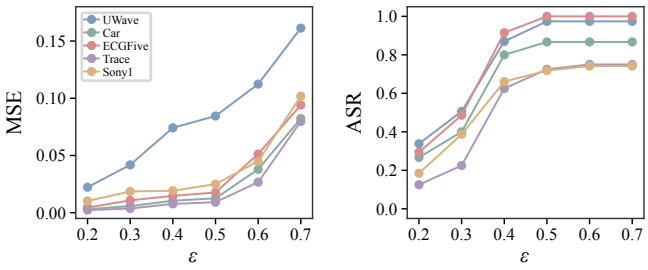


Figure 7: Effect of the perturbation budget ϵ.

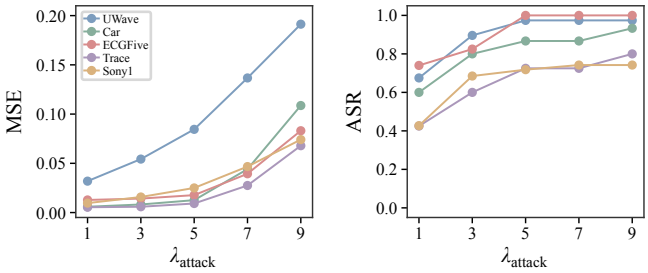


Figure 8: Effect of the attack weight λ_{attack}.

and from 12.5% to 72.5% on Trace, while ECGFiveDays reaches 100% ASR at $\epsilon = 0.5$. Beyond $\epsilon = 0.5$, however, ASR tends to saturate while MSE continues to increase. On UWaveGestureLibrary, increasing ϵ from 0.5 to 0.7 leaves ASR unchanged at 97.4%, whereas MSE rises from 0.0845 to 0.1613. Similar increases in distortion are observed on Car and SonyAIBORobotSurface1. These results indicate that further enlarging the amplitude budget mainly increases perturbation distortion once sufficient boundary-crossing capability has been reached. We therefore use $\epsilon = 0.5$ as the default setting.

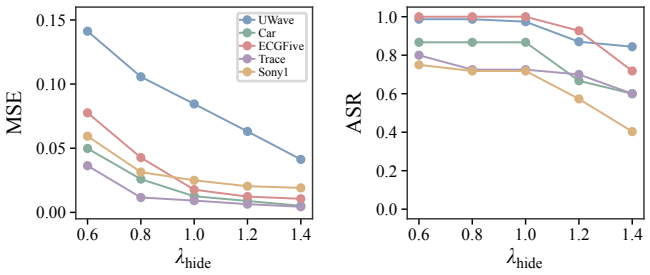


Figure 9: Effect of the perturbation-regularization weight λ_{hide}.

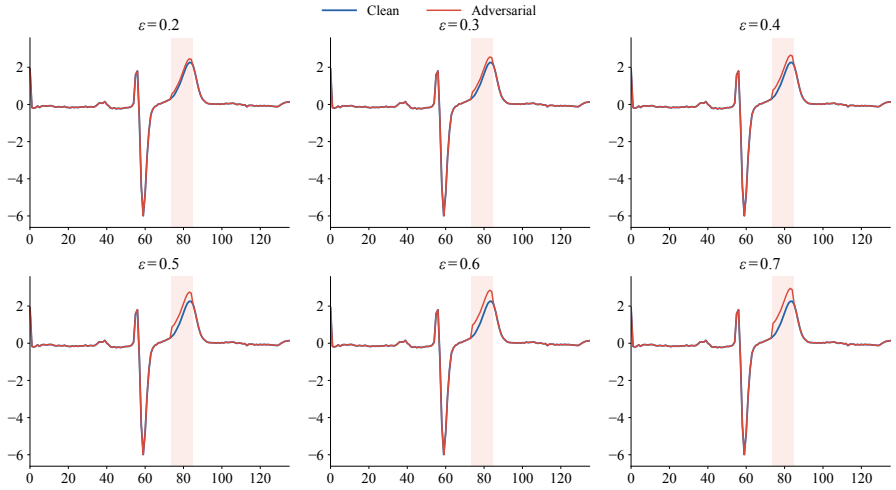


Figure 10: Effect of the perturbation budget on ECGFiveDays.

Figure 10 keeps the ECGFiveDays sample and learned support fixed while varying ϵ. The curves remain concentrated in the same selected region as the budget increases: ϵ changes the strength of the local displacement, while Joint Coarse-and-Fine attention determines where that displacement occurs. This separation explains the parameter analysis in Figure 7: ASR largely saturates near $\epsilon = 0.5$, but MSE continues to grow when the amplitude budget is increased further.

Effect of the attack-loss weight. Figure 8 examines λ_{attack}, which controls the contribution of the hybrid boundary objective during generator optimization. Increasing λ_{attack} from 1.0 to 5.0 generally strengthens boundary crossing and improves ASR across the evaluated datasets. A larger attack weight places greater emphasis on generating perturbations that cross the surrogate decision boundary. The benefit becomes limited once λ_{attack} exceeds 5.0. For example, increasing it to 9.0 yields 97.4% ASR on UWaveGestureLibrary and 100% on ECGFiveDays, while producing noticeably larger MSE. Thus, an excessively large attack weight mainly increases perturbation distortion after the boundary objective has already been sufficiently optimized. We therefore set $\lambda_{\text{attack}} = 5.0$ by default.

Effect of the perturbation regularization. Figure 9 studies the regularization weight λ_{hide}, which controls the penalty on the final masked perturbation magnitude. A smaller value weakens regularization and allows stronger perturbations, generally improving attack success at the cost of larger MSE. Increasing λ_{hide} suppresses perturbation magnitude, but excessive regularization can prevent the generated perturbation from crossing the decision boundary. At $\lambda_{\text{hide}} = 1.4$, ASR decreases to 84.4% on UWaveGestureLibrary, 60.0% on Car, and 40.3% on SonyAIBORobotSurface1. The attack-loss and regularization weights play complementary roles: λ_{attack} encourages boundary crossing, whereas λ_{hide} constrains perturbation magnitude. Setting $\lambda_{\text{hide}} = 1.0$ provides a balanced compromise between these objectives.

Overall, the three sensitivity studies exhibit a consistent effectiveness–distortion trade-off. Increasing the perturbation budget or attack weight improves attack success only up to a saturation region, whereas stronger perturbation regularization reduces distortion but can weaken the attack. Based on these observations, we use $\epsilon = 0.5$, $\lambda_{\text{attack}} = 5.0$, and $\lambda_{\text{hide}} = 1.0$ as the default configuration.

D Complete Experimental Results

Appendix Table 8 reports the 11 datasets omitted from the compact body table, Appendix Table 9 provides the corresponding ablation rows and their 11-dataset average, and Appendix Table 10 contains the complete additional defense results. Together with the three representative datasets in the main text, these tables form the complete evaluation.

Attack Performance on the Additional Datasets

Table 8 extends the comparison to 11 univariate datasets. C2FDA achieves the highest dataset-average ASR on CinCECGTorso (61.6%), ECGFiveDays (74.3%), Earthquakes (54.8%), ItalyPowerDemand (73.1%), and SonyAIBORobotSurface2 (55.6%). Together with the three multivariate datasets in the main text, it leads on eight of the 14 datasets. FordA and FordB both show nonzero transfer to all five targets, with average ASRs of 84.2% and 82.3%, respectively.

The comparison with AutoZOOM illustrates the success–distortion trade-off. On Car and Trace, C2FDA trails AutoZOOM in average ASR by 1.5 and 2.3 percentage points, respectively, but reduces MSE from 0.0167 to 0.0136 and from 0.0146 to 0.0127, and DTW from 41.40 to 23.13 and from 20.61 to 11.77. These results show that learned local support can retain high attack effectiveness with less waveform distortion while avoiding target-side query optimization.

Transferability still depends on surrogate–target alignment. C2FDA reaches 100% CNN ASR on ECG5000 but averages 50.7% over all five targets, below GATN at 53.7%; on Earthquakes, zero transfer to TCN reduces its average ASR to 54.8%. The local *where–what* design therefore does not eliminate variation across target architectures.

Ablation Results on the Additional Datasets

Table 9 reports the 11 additional datasets and their average results. The full model achieves 83.30% ASR, compared with 73.08% without coarse attention and 60.19% without fine attention. Removing the full attention mechanism retains 77.27% ASR but increases MSE from 0.0263 to 0.1509 and DTW from 20.90 to 102.13. Thus, the shared temporal representation supports both attack effectiveness and distortion control, consistent with coupling *where* and *what*.

Random intervals reduce average ASR to 37.70%, despite higher MSE (0.0411) and DTW (29.72) than the full model, reinforcing the importance of sample-specific support selection. Using only the exact margin yields 71.53% ASR, 0.0427 MSE, and 27.95 DTW. The hybrid boundary objective improves all three averages, supporting the complementary contribution of neighborhood-level boundary information.

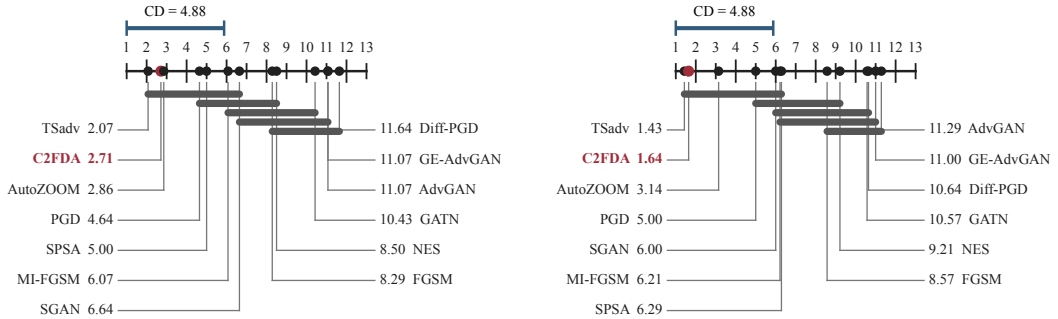
Statistical Analysis of Perturbation Distortion

Figure 11 compares MSE and DTW ranks across all 14 datasets. For each dataset, the reported Avg values are ranked in ascending order, with average ranks assigned to ties. Entries marked "–" receive the tied-worst rank; numerical entries are retained as reported. The Friedman tests reject equal performance for both metrics ($p < 10^{-15}$). We apply the two-sided Nemenyi test at $\alpha = 0.05$, with a critical difference of 4.88. Methods connected by a horizontal bar are not significantly different under this criterion.

C2FDA achieves the second-best average ranks for both MSE (2.71) and DTW (1.64). TSadv attains slightly better distortion ranks, but its weaker ASR rank highlights the need to consider whether the perturbation content remains effective at the selected locations. Together with the leading ASR rank, these results support the goal of coupling *where* and *what*: generate effective perturbation content within restricted temporal support while preserving the original waveform.

Table 8: Adversarial attack results on the additional datasets.

Dataset	Attack Method	BiRNN ASR↑	BiRNN MSE↓	BiRNN DTW↓	LSTM ASR↑	LSTM MSE↓	LSTM DTW↓	CNN ASR↑	CNN MSE↓	CNN DTW↓	TCN ASR↑	TCN MSE↓	TCN DTW↓	DynamicConv ASR↑	DynamicConv MSE↓	DynamicConv DTW↓	Avg ASR↑	Avg MSE↓	Avg DTW↓
Car	FGSM	46.2%	0.0900	148.93	77.8%	0.0900	147.90	53.3%	0.0900	148.65	0.0%	–	–	0.0%	–	–	35.5%	0.0900	148.49
	PGD	30.8%	0.0156	39.45	44.4%	0.0157	38.57	6.7%	0.0157	39.19	0.0%	–	–	0.0%	–	–	16.4%	0.0157	39.07
	MI-FGSM	30.8%	0.0320	60.38	66.7%	0.0327	60.30	13.3%	0.0337	63.65	0.0%	–	–	0.0%	–	–	22.2%	0.0328	61.44
	NES	69.2%	0.0920	154.54	77.8%	0.0900	154.58	53.3%	0.0900	155.08	8.3%	0.0900	154.10	9.1%	0.0900	153.69	43.6%	0.0904	154.40
	SPSA	**100.0%**	0.0547	96.18	**100.0%**	0.0531	95.31	**100.0%**	0.0483	91.75	25.0%	0.0526	92.50	**100.0%**	0.0520	87.56	85.0%	0.0521	92.66
	AutoZOOM	**100.0%**	**0.0109**	37.23	**100.0%**	**0.0108**	38.58	**100.0%**	0.0160	45.35	33.3%	0.0167	46.21	**100.0%**	0.0292	39.61	**86.7%**	0.0167	41.40
	GATN	7.7%	0.1185	195.78	11.1%	0.1074	171.39	13.3%	0.1130	191.65	8.3%	0.1295	200.14	9.1%	0.1345	187.18	9.9%	0.1206	189.23
	TSadv	0.0%	–	–	0.0%	–	–	80.0%	**0.0017**	**7.07**	33.3%	**0.0013**	**5.07**	0.0%	–	–	22.7%	**0.0015**	**6.07**
	AdvGAN	38.5%	0.2588	192.69	44.4%	0.2655	216.81	80.0%	0.2460	205.36	41.7%	0.2303	190.53	36.4%	0.2679	200.57	48.2%	0.2537	201.19
	Diff-PGD	30.8%	0.2709	217.23	33.3%	0.2743	204.79	60.0%	0.2349	186.04	8.3%	0.2516	198.51	27.3%	0.3242	271.29	31.9%	0.2712	215.57
	GE-AdvGAN	46.2%	0.1913	160.63	33.3%	0.1852	170.32	53.3%	0.1793	165.41	41.7%	0.1631	179.62	36.4%	0.2046	214.26	42.2%	0.1847	178.05
	SGAN	76.9%	0.0295	65.19	33.3%	0.0304	66.70	66.7%	0.0286	62.91	41.7%	0.0279	61.32	63.6%	0.0253	56.08	56.4%	0.0283	62.44
	C2FDA	92.3%	0.0129	**27.12**	88.9%	0.0148	**19.43**	86.7%	0.0126	26.67	**58.3%**	0.0120	30.21	**100.0%**	**0.0157**	**12.21**	85.2%	0.0136	23.13
CinCECGTorso	FGSM	25.6%	0.0900	465.05	23.6%	0.0900	447.89	11.5%	0.0900	439.61	5.1%	0.0900	447.20	23.4%	0.0900	456.94	17.8%	0.0900	451.34
	PGD	14.6%	0.0166	134.51	16.9%	0.0123	105.98	9.0%	0.0160	122.35	3.7%	0.0150	121.76	7.8%	0.0166	122.43	10.4%	0.0153	121.41
	MI-FGSM	24.6%	0.0332	202.81	22.2%	0.0299	184.25	9.4%	0.0356	203.72	4.4%	0.0308	187.38	36.7%	0.0343	196.87	19.4%	0.0328	195.01
	NES	32.7%	0.0915	460.25	23.1%	0.0900	460.19	10.4%	0.0900	460.41	4.4%	0.0900	460.49	15.6%	0.0900	460.26	17.3%	0.0903	460.32
	SPSA	81.5%	0.0526	288.39	83.6%	0.0524	288.49	50.4%	0.0518	285.37	21.6%	0.0517	285.22	32.1%	0.0530	287.93	53.8%	0.0523	287.08
	AutoZOOM	84.7%	**0.0105**	116.61	88.4%	**0.0099**	114.45	53.6%	0.0126	127.02	22.0%	**0.0124**	126.68	37.2%	**0.0113**	106.25	57.2%	**0.0113**	118.20
	GATN	23.8%	0.1636	268.97	21.8%	0.1831	300.59	28.1%	0.1890	320.52	10.3%	0.1640	280.23	17.4%	0.1985	326.68	20.3%	0.1796	299.40
	TSadv	53.4%	0.0256	130.75	10.2%	0.0140	**72.53**	46.4%	0.0315	160.78	**22.7%**	0.0224	116.59	7.3%	0.0157	84.02	28.0%	0.0218	112.93
	AdvGAN	37.4%	0.2023	338.85	40.4%	0.1902	313.37	75.5%	0.1380	292.96	17.6%	0.2078	291.83	28.0%	0.2036	328.81	39.8%	0.1884	313.16
	Diff-PGD	26.7%	0.2150	226.33	32.0%	0.2159	227.13	69.8%	0.1625	185.92	13.9%	0.2438	207.95	22.5%	0.2064	216.62	33.0%	0.2087	212.79
	GE-AdvGAN	28.8%	0.1491	249.99	28.0%	0.1513	256.95	62.9%	0.0976	263.92	13.9%	0.1229	226.68	23.4%	0.1200	249.57	31.4%	0.1282	249.42
	SGAN	39.5%	0.0299	172.08	38.2%	0.0344	189.18	**79.9%**	0.0368	164.53	15.4%	0.0354	142.06	26.6%	0.0319	144.84	39.9%	0.0337	162.54
	C2FDA	**86.1%**	0.0143	**87.34**	**89.3%**	0.0110	83.76	72.3%	**0.0104**	**70.82**	22.3%	0.0134	**73.53**	**38.1%**	0.0139	**78.73**	**61.6%**	0.0126	**78.84**
ECG5000	FGSM	1.3%	0.0900	36.46	0.7%	0.0900	36.88	3.9%	0.0900	34.87	1.7%	0.0900	35.00	2.0%	0.0900	35.25	1.9%	0.0900	35.69
	PGD	0.9%	0.0289	15.81	0.5%	0.0252	15.37	2.2%	0.0284	14.86	0.9%	0.0321	16.34	1.9%	0.0413	18.78	1.3%	0.0312	16.23
	MI-FGSM	1.0%	0.0425	19.66	0.6%	0.0380	19.47	3.2%	0.0424	18.71	1.0%	0.0465	20.33	2.2%	0.0550	22.32	1.6%	0.0449	20.10
	NES	1.8%	0.0876	36.57	0.6%	0.0900	36.58	1.9%	0.0900	36.56	3.1%	0.0900	36.58	1.3%	0.0900	36.54	1.7%	0.0895	36.57
	SPSA	29.3%	0.0568	23.39	21.5%	0.0553	23.38	53.6%	0.0561	24.20	35.1%	0.0570	23.48	5.4%	0.0528	23.51	29.0%	0.0556	23.59
	AutoZOOM	35.7%	0.0351	16.04	29.6%	0.0266	13.48	60.3%	0.0263	14.71	37.8%	0.0369	16.59	6.6%	0.0199	12.53	34.0%	0.0290	14.67
	GATN	24.3%	0.1456	29.75	**67.2%**	0.1354	25.04	**100.0%**	0.1530	23.66	54.5%	0.1690	23.15	22.4%	0.1456	27.37	**53.7%**	0.1497	25.79
	TSadv	8.2%	**0.0099**	**5.68**	12.0%	**0.0102**	5.72	6.7%	**0.0100**	**6.16**	31.8%	**0.0099**	**5.72**	3.2%	**0.0095**	5.55	12.4%	**0.0099**	**5.77**
	AdvGAN	21.8%	0.1976	69.03	31.9%	0.1986	71.89	86.1%	0.1796	66.15	25.8%	0.2040	71.67	**24.2%**	0.1842	31.10	38.0%	0.1928	61.97
	Diff-PGD	36.8%	0.2127	72.46	24.2%	0.2547	76.06	63.7%	0.1874	25.86	**71.4%**	0.2304	70.79	8.6%	0.2554	76.84	40.9%	0.2281	64.40
	GE-AdvGAN	28.9%	0.2065	74.16	27.7%	0.1547	61.02	64.8%	0.1383	64.98	20.4%	0.1942	74.83	10.5%	0.1702	69.79	30.5%	0.1728	68.96
	SGAN	33.9%	0.0635	23.67	44.8%	0.0893	22.53	73.7%	0.0783	19.82	34.8%	0.0693	18.13	10.5%	0.0876	24.11	39.5%	0.0776	21.65
	C2FDA	**39.7%**	0.0191	7.95	50.5%	0.0198	**4.65**	**100.0%**	0.0221	13.38	46.1%	0.0194	7.56	17.3%	0.0134	**5.44**	50.7%	0.0188	7.80
ECGFiveDays	FGSM	12.4%	0.0900	36.30	13.6%	0.0900	37.05	5.1%	0.0900	36.71	2.8%	0.0900	36.02	28.7%	0.0900	36.43	12.5%	0.0900	36.50
	PGD	10.7%	0.0355	17.58	6.2%	0.0285	16.04	4.0%	0.0299	16.68	5.1%	0.0299	15.49	16.8%	0.0295	15.66	8.6%	0.0307	16.29
	MI-FGSM	11.3%	0.0477	21.34	13.1%	0.0426	20.76	4.5%	0.0444	21.31	5.1%	0.0450	20.26	19.2%	0.0429	20.26	10.6%	0.0445	20.79
	NES	16.9%	0.0925	37.62	11.9%	0.0900	37.52	7.9%	0.0900	37.59	2.8%	0.0900	37.64	16.8%	0.0900	37.58	11.3%	0.0905	37.59
	SPSA	83.1%	0.0569	24.79	**83.5%**	0.0563	23.70	48.0%	0.0555	24.71	31.6%	0.0552	24.45	**91.6%**	0.0516	23.60	67.6%	0.0551	24.25
	AutoZOOM	**87.6%**	0.0266	13.72	83.0%	0.0236	12.36	52.5%	0.0272	14.07	35.6%	0.0244	13.17	**91.6%**	0.0175	11.67	70.1%	0.0239	13.00
	GATN	9.6%	0.0878	62.06	10.8%	0.0927	65.99	11.9%	0.0830	64.23	5.6%	0.0695	60.63	11.4%	0.0738	62.89	9.9%	0.0814	63.16
	TSadv	2.8%	0.0171	5.82	2.8%	0.0191	6.33	53.7%	0.0185	**6.97**	**41.2%**	0.0165	**5.75**	12.6%	**0.0131**	5.88	22.6%	0.0169	**6.15**
	AdvGAN	21.5%	0.1318	80.85	19.3%	0.1425	86.08	52.5%	0.0931	53.42	10.2%	0.1137	66.11	21.6%	0.1118	64.01	25.0%	0.1186	70.09
	Diff-PGD	22.0%	0.119	53.76	23.3%	0.1282	55.67	45.8%	0.1173	54.47	10.7%	0.1335	60.65	27.5%	0.1122	53.62	25.9%	0.1220	55.63
	GE-AdvGAN	13.6%	0.1361	73.27	12.5%	0.1286	74.15	24.3%	0.1374	75.55	5.6%	0.1293	68.55	13.8%	0.1299	70.57	14.0%	0.1323	72.42
	SGAN	29.4%	0.0169	23.23	35.2%	**0.0178**	24.88	35.6%	**0.0172**	26.55	8.5%	**0.0154**	25.59	16.2%	0.0186	24.44	25.0%	0.0172	24.94
	C2FDA	79.7%	**0.0157**	**3.62**	69.3%	0.0184	**6.27**	**100.0%**	0.0177	11.04	35.6%	0.0160	8.31	86.8%	0.0156	**4.62**	**74.3%**	**0.0167**	6.77



(a) MSE↓ (b) DTW↓

Figure 11: Critical-difference diagrams for MSE and DTW across all 14 datasets. Lower average rank indicates better overall performance.

Table 8: Adversarial attack results (continued).

Dataset	Attack	BiRNN			LSTM			CNN			TCN			DynamicConv			Avg		
	Method	ASR↑	MSE↓	DTW↓	ASR↑	MSE↓	DTW↓	ASR↑	MSE↓	DTW↓	ASR↑	MSE↓	DTW↓	ASR↑	MSE↓	DTW↓	ASR↑	MSE↓	DTW↓
Earthquakes	FGSM	10.5%	0.0900	153.51	1.4%	0.0900	153.51	1.3%	0.0900	153.51	0.0%	–	–	7.9%	0.0900	153.50	4.2%	0.0900	153.51
	PGD	6.6%	0.0600	116.20	1.4%	0.0505	103.64	1.3%	0.0464	98.37	0.0%	–	–	6.6%	0.0549	108.15	3.2%	0.0530	106.59
	MI-FGSM	7.9%	0.0708	129.07	1.4%	0.0619	117.53	2.6%	0.0604	115.61	0.0%	–	–	6.6%	0.0618	116.69	3.7%	0.0637	119.73
	NES	21.1%	0.0920	153.49	1.4%	0.0900	153.50	7.8%	0.0900	153.50	0.0%	–	–	6.6%	0.0900	153.51	7.4%	0.0905	153.50
	SPSA	75.0%	0.0553	110.63	2.7%	0.0521	108.48	93.5%	0.0544	110.63	0.0%	–	–	57.9%	0.0514	106.81	45.8%	0.0533	109.14
	AutoZOOM	81.6%	**0.0125**	**43.81**	4.1%	**0.0126**	44.85	**94.8%**	0.0231	61.35	0.0%	–	–	57.9%	0.0168	52.59	47.7%	0.0163	50.65
	GATN	1.3%	0.1419	196.79	0.0%	–	–	3.9%	0.1220	184.14	0.0%	–	–	1.3%	0.1126	200.51	1.3%	0.1255	193.81
	TSadv	0.0%	–	–	0.0%	–	–	23.4%	**0.0026**	**9.19**	0.0%	–	–	5.3%	**0.0027**	**9.85**	5.7%	**0.0027**	**9.52**
	AdvGAN	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–
	Diff-PGD	1.3%	0.1841	191.46	0.0%	–	–	3.9%	0.2179	211.84	0.0%	–	–	1.3%	0.3400	297.25	1.3%	0.2473	233.52
	GE-AdvGAN	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–
	SGAN	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–
	C2FDA	**90.8%**	0.0986	45.28	**16.2%**	0.0889	**43.93**	**94.8%**	0.0944	44.45	0.0%	–	–	**72.4%**	0.0877	43.56	**54.8%**	0.0924	44.31
FordA	FGSM	4.9%	0.0900	116.39	15.8%	0.0900	120.48	27.1%	0.0900	121.99	7.2%	0.0900	121.22	46.2%	0.0900	117.91	20.2%	0.0900	119.60
	PGD	1.8%	0.0454	75.27	13.0%	0.0310	64.00	17.1%	0.0261	57.99	3.6%	0.0258	56.22	33.3%	0.0382	69.81	13.7%	0.0333	64.66
	MI-FGSM	1.8%	0.0552	82.92	14.4%	0.0440	76.70	22.0%	0.0399	72.07	5.0%	0.0395	69.99	36.6%	0.0504	80.72	15.9%	0.0458	76.48
	NES	5.7%	0.0917	120.46	19.2%	0.0900	120.43	25.9%	0.0900	120.46	5.2%	0.0900	120.45	59.0%	0.0900	120.50	23.0%	0.0903	120.46
	SPSA	83.2%	0.0524	82.17	81.2%	0.0530	86.25	99.7%	0.0494	83.00	90.1%	0.0519	81.53	95.0%	0.0552	87.64	89.8%	0.0524	84.12
	AutoZOOM	**89.1%**	0.0238	50.57	**85.0%**	0.0121	38.52	**99.9%**	0.0155	44.47	**97.6%**	0.0304	57.73	**98.1%**	0.0098	34.76	**93.9%**	0.0183	45.21
	GATN	5.3%	0.1032	153.88	8.2%	0.1039	150.79	10.0%	0.1130	154.10	5.5%	0.1356	163.96	10.6%	0.1125	152.55	7.9%	0.1136	155.06
	TSadv	27.8%	**0.0079**	**20.44**	0.4%	**0.0046**	**12.23**	93.9%	**0.0079**	**20.28**	27.2%	**0.0077**	**19.89**	72.5%	**0.0062**	**15.82**	44.4%	**0.0069**	**17.73**
	AdvGAN	38.2%	0.1779	153.01	52.7%	0.1837	161.03	64.8%	0.2281	190.91	31.9%	0.1975	176.59	76.2%	0.1736	155.91	52.8%	0.1922	167.49
	Diff-PGD	39.4%	0.1255	142.01	53.8%	0.1397	153.68	62.3%	0.1343	148.20	50.2%	0.1381	131.37	74.6%	0.1034	124.04	56.1%	0.1282	139.86
	GE-AdvGAN	18.3%	0.1729	170.12	34.3%	0.2157	184.80	60.8%	0.1831	161.95	41.9%	0.1668	190.08	67.5%	0.2358	216.41	44.6%	0.1949	184.67
	SGAN	70.1%	0.0894	67.03	63.5%	0.0917	73.39	71.6%	0.0935	71.23	62.0%	0.1049	76.66	84.3%	0.1085	77.81	70.3%	0.0976	73.22
	C2FDA	78.6%	0.0196	20.46	80.6%	0.0085	18.49	87.3%	0.0127	20.63	83.7%	0.0186	25.24	90.9%	0.0077	27.01	84.2%	0.0134	22.37
FordB	FGSM	10.8%	0.0900	116.19	78.4%	0.0900	120.15	54.1%	0.0900	121.01	19.4%	0.0900	120.31	65.7%	0.0900	117.42	45.7%	0.0900	119.02
	PGD	9.6%	0.0443	74.06	17.0%	0.0291	61.71	23.5%	0.0244	55.89	7.3%	0.0242	54.21	58.1%	0.0360	66.91	23.1%	0.0316	62.56
	MI-FGSM	9.4%	0.0544	81.86	37.2%	0.0425	75.02	33.3%	0.0385	70.17	11.3%	0.0381	68.33	64.1%	0.0485	78.20	31.1%	0.0444	74.72
	NES	8.2%	0.0883	119.71	61.8%	0.0900	119.83	53.2%	0.0900	119.90	19.6%	0.0900	119.72	65.7%	0.0900	119.61	41.7%	0.0897	119.75
	SPSA	**75.1%**	0.0527	83.03	90.9%	0.0554	87.45	**99.7%**	0.0497	82.95	96.7%	0.0507	81.35	97.4%	0.0536	86.09	91.9%	0.0524	84.17
	AutoZOOM	71.3%	0.0245	51.12	**91.3%**	0.0105	35.58	**99.7%**	0.0140	42.32	**98.4%**	0.0235	51.91	**100.0%**	0.0126	40.81	**92.2%**	0.0170	44.35
	GATN	58.9%	0.1182	157.70	44.6%	0.1065	148.02	60.3%	0.1190	155.20	54.9%	0.1226	160.38	42.3%	0.1425	162.51	52.2%	0.1218	156.76
	TSadv	48.9%	**0.0114**	**29.22**	88.2%	**0.0088**	21.90	99.3%	**0.0111**	27.63	66.7%	**0.0111**	**28.25**	88.6%	**0.0086**	21.78	78.3%	**0.0102**	25.76
	AdvGAN	42.1%	0.2032	153.22	64.7%	0.2148	160.04	80.6%	0.2466	193.09	77.6%	0.2246	189.13	67.5%	0.1954	187.82	66.5%	0.2169	176.66
	Diff-PGD	55.4%	0.2278	187.61	60.9%	0.1931	153.90	74.3%	0.2594	233.59	63.9%	0.2498	189.59	74.7%	0.2412	178.78	65.8%	0.2343	188.69
	GE-AdvGAN	36.2%	0.2019	148.64	54.6%	0.2390	163.26	71.4%	0.1823	157.76	49.6%	0.1861	163.79	41.9%	0.1569	151.98	50.7%	0.1932	157.09
	SGAN	66.5%	0.0593	74.09	87.4%	0.0817	82.03	84.1%	0.0693	87.44	72.3%	0.0691	89.25	75.1%	0.0581	94.31	77.1%	0.0675	85.42
	C2FDA	64.7%	0.0198	37.09	79.3%	0.0097	**16.94**	90.3%	0.0127	**22.43**	85.2%	0.0184	32.54	92.2%	0.0104	**16.28**	82.3%	0.0142	**25.06**
Trace	FGSM	2.5%	0.0900	78.27	4.2%	0.0900	77.17	10.0%	0.0900	75.21	7.5%	0.0900	76.73	62.5%	0.0900	75.18	17.3%	0.0900	76.51
	PGD	20.0%	0.0205	26.78	4.2%	0.0218	26.84	0.0%	–	–	17.5%	0.0248	26.64	50.0%	0.0308	30.26	18.3%	0.0245	27.63
	MI-FGSM	17.5%	0.0345	36.86	8.3%	0.0367	36.73	2.5%	0.0414	38.29	20.0%	0.0418	36.65	57.5%	0.0459	39.67	21.2%	0.0401	37.64
	NES	10.0%	0.0916	76.90	29.2%	0.0900	77.00	17.5%	0.0900	76.94	27.5%	0.0900	76.98	65.0%	0.0900	77.24	29.8%	0.0903	77.01
	SPSA	37.5%	0.0550	49.26	**87.5%**	0.0537	48.08	40.0%	0.0520	47.28	**75.0%**	0.0530	46.82	**100.0%**	0.0516	47.91	68.0%	0.0531	47.87
	AutoZOOM	**45.0%**	**0.0173**	23.02	**87.5%**	**0.0111**	17.69	42.5%	0.0174	22.85	**75.0%**	0.0169	21.90	**100.0%**	0.0103	17.57	**70.0%**	0.0146	20.61
	GATN	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–	0.0%	–	–
	TSadv	0.0%	–	–	0.0%	–	–	5.0%	**0.0033**	**6.29**	25.0%	**0.0033**	**6.18**	40.0%	**0.0025**	**4.55**	14.0%	**0.0030**	**5.67**
	AdvGAN	12.5%	0.1148	92.39	41.7%	0.1209	92.12	47.5%	0.1359	96.70	22.5%	0.1036	87.03	25.0%	0.1252	99.91	29.8%	0.1201	93.63
	Diff-PGD	12.5%	0.1045	80.04	20.8%	0.1394	95.41	50.0%	0.1286	81.42	50.0%	0.1160	89.14	47.5%	0.1362	89.21	36.2%	0.1249	87.04
	GE-AdvGAN	10.0%	0.1535	129.27	41.7%	0.1340	104.57	42.5%	0.1229	102.12	40.0%	0.1099	91.22	32.5%	0.1222	94.82	33.3%	0.1285	104.40
	SGAN	25.0%	0.0435	55.72	66.7%	0.0350	47.79	60.0%	0.0417	51.50	57.5%	0.0476	48.01	72.5%	0.0442	49.74	56.3%	0.0424	50.55
	C2FDA	30.0%	0.0192	**18.48**	83.3%	0.0152	**9.76**	**72.5%**	0.0092	13.31	**75.0%**	0.0146	8.65	77.5%	0.0055	8.66	67.7%	0.0127	11.77

Additional Defense Results

Table 10 reports pre-purification ASR / DSR (%) on CNN for the remaining 11 datasets. After purification, C2FDA retains higher ASR than AutoZOOM and SGAN on all 11 datasets and TSadv on nine. On ECGFiveDays, residual ASR is 88.7% for C2FDA, compared with 23.7% for AutoZOOM, 46.3% for TSadv, and 18.6% for SGAN. C2FDA also has lower DSR than AutoZOOM on all 11 datasets and SGAN on all 10 datasets with nonzero initial SGAN ASR.

Qualitative Comparison

Figure 12 compares a random contiguous interval with C2FDA on an ECG5000 sample. The random interval lies in an early segment, whereas C2FDA selects late-sequence transitions and the terminal waveform change (*where*) and concentrates its generated deviations within these regions (*what*). The clean and adversarial series remain closely aligned elsewhere.

Table 8: Adversarial attack results (continued).

Dataset	Attack Method	BiRNN ASR↑	BiRNN MSE↓	BiRNN DTW↓	LSTM ASR↑	LSTM MSE↓	LSTM DTW↓	CNN ASR↑	CNN MSE↓	CNN DTW↓	TCN ASR↑	TCN MSE↓	TCN DTW↓	DynamicConv ASR↑	DynamicConv MSE↓	DynamicConv DTW↓	Avg ASR↑	Avg MSE↓	Avg DTW↓
ItalyPowerDemand	FGSM	31.5%	0.0900	6.21	34.4%	0.0900	6.21	45.1%	0.0900	6.21	50.2%	0.0900	6.21	30.8%	0.0900	6.21	38.4%	0.0900	6.21
	PGD	33.8%	0.0712	5.27	36.8%	0.0712	5.27	46.0%	0.0712	5.27	53.1%	0.0712	5.27	35.5%	0.0712	5.27	41.0%	0.0712	5.27
	MI-FGSM	32.4%	0.0769	5.57	35.4%	0.0769	5.57	44.6%	0.0769	5.57	52.6%	0.0769	5.57	34.1%	0.0769	5.57	39.8%	0.0769	5.57
	NES	8.0%	0.0924	6.35	7.5%	0.0900	6.37	8.0%	0.0900	6.36	7.0%	0.0900	6.39	14.2%	0.0900	6.37	9.0%	0.0905	6.37
	SPSA	60.6%	0.0697	5.12	**64.6%**	0.0667	5.03	**88.3%**	0.0709	5.27	72.8%	0.0750	5.23	55.9%	0.0588	4.88	68.4%	0.0682	5.11
	AutoZOOM	60.1%	0.0565	4.24	61.8%	0.0518	3.97	85.9%	0.0724	5.30	**77.0%**	0.0786	5.29	63.0%	0.0836	3.91	69.6%	0.0686	4.54
	GATN	8.9%	0.1260	13.99	8.5%	0.1189	12.65	12.2%	0.1232	12.72	9.4%	0.1261	12.46	9.5%	0.1306	13.49	9.7%	0.1250	13.06
	TSadv	10.3%	**0.0070**	**1.30**	13.7%	**0.0075**	1.37	18.8%	**0.0078**	1.43	17.8%	**0.0074**	1.36	3.3%	**0.0059**	**1.17**	12.8%	**0.0071**	1.33
	AdvGAN	20.2%	0.1406	18.01	21.2%	0.1182	17.87	27.7%	0.1192	15.72	23.9%	0.1095	16.08	22.3%	0.1260	18.13	23.1%	0.1227	17.16
	Diff-PGD	12.7%	0.1373	12.52	12.7%	0.1333	12.47	16.9%	0.1267	11.46	11.7%	0.1252	10.41	12.3%	0.1013	9.57	13.3%	0.1248	11.29
	GE-AdvGAN	17.4%	0.1206	12.69	17.5%	0.1094	9.86	24.9%	0.1067	10.32	19.7%	0.1122	11.60	17.5%	0.1321	13.82	19.4%	0.1162	11.66
	SGAN	14.6%	0.0633	4.64	13.2%	0.0693	5.90	20.2%	0.0787	4.60	16.4%	0.0728	4.56	13.3%	0.0825	6.00	15.5%	0.0733	5.14
	C2FDA	**71.4%**	0.0377	1.35	62.7%	0.0373	**1.13**	87.3%	0.0403	**1.24**	75.1%	0.0531	**1.07**	**68.7%**	0.0436	1.30	**73.1%**	0.0424	**1.22**
SonyAIBORobotSurface1	FGSM	25.0%	0.0900	18.50	39.0%	0.0900	18.50	23.4%	0.0900	18.50	19.4%	0.0900	18.50	44.6%	0.0900	18.50	30.3%	0.0900	18.50
	PGD	25.0%	0.0692	15.47	40.7%	0.0692	15.47	25.0%	0.0692	15.47	24.2%	0.0692	15.47	43.0%	0.0692	15.47	31.6%	0.0692	15.47
	MI-FGSM	25.0%	0.0725	15.85	40.7%	0.0725	15.85	24.2%	0.0725	15.85	23.4%	0.0725	15.85	43.8%	0.0725	15.85	31.4%	0.0725	15.85
	NES	10.5%	0.0878	18.57	19.5%	0.0900	18.56	8.1%	0.0900	18.53	4.0%	0.0900	18.56	15.7%	0.0900	18.55	11.6%	0.0896	18.55
	SPSA	**52.4%**	0.0561	13.63	**82.1%**	0.0566	13.57	67.7%	0.0607	14.33	44.4%	0.0596	14.35	69.4%	0.0557	13.54	63.2%	0.0577	13.88
	AutoZOOM	51.6%	0.0766	10.49	81.3%	0.0529	8.38	**75.8%**	0.0451	11.82	**49.2%**	0.0391	10.96	73.6%	0.0465	9.31	**66.3%**	0.0520	10.19
	GATN	25.8%	0.1421	30.85	38.2%	0.1358	28.56	43.5%	0.1480	35.55	27.4%	0.1582	37.18	39.7%	0.1559	38.54	34.9%	0.1480	34.14
	TSadv	2.4%	**0.0068**	**2.63**	21.1%	**0.0063**	**2.43**	22.6%	**0.0086**	3.24	8.9%	**0.0101**	3.78	12.4%	**0.0055**	**2.04**	13.5%	**0.0075**	**2.82**
	AdvGAN	10.5%	0.1589	32.34	16.3%	0.1463	31.19	19.4%	0.1341	32.35	11.3%	0.1597	36.65	18.2%	0.1317	30.35	15.1%	0.1461	32.58
	Diff-PGD	12.1%	0.1114	27.78	17.1%	0.1458	30.63	23.4%	0.1315	28.95	13.7%	0.1391	25.79	19.0%	0.1243	23.16	17.1%	0.1304	27.26
	GE-AdvGAN	13.7%	0.1161	21.17	21.1%	0.1146	20.67	21.8%	0.1310	24.68	13.7%	0.1160	25.90	19.8%	0.1223	27.93	18.0%	0.1200	24.07
	SGAN	12.1%	0.0645	13.50	18.7%	0.0658	13.38	22.6%	0.0625	11.43	12.9%	0.0545	12.33	21.5%	0.0646	10.74	17.6%	0.0624	12.28
	C2FDA	50.8%	0.0281	3.71	80.5%	0.0207	4.04	71.8%	0.0250	**2.86**	39.5%	0.0233	**3.72**	**78.5%**	0.0137	2.11	64.2%	0.0222	3.29
SonyAIBORobotSurface2	FGSM	6.2%	0.0900	16.62	19.6%	0.0900	16.62	24.6%	0.0900	16.62	13.3%	0.0900	16.62	8.1%	0.0900	16.62	14.3%	0.0900	16.62
	PGD	6.2%	0.0700	14.54	21.6%	0.0700	14.54	26.2%	0.0700	14.54	13.8%	0.0700	14.54	8.6%	0.0700	14.54	15.3%	0.0700	14.54
	MI-FGSM	6.7%	0.0743	14.95	22.2%	0.0743	14.95	26.2%	0.0743	14.95	13.3%	0.0743	14.95	7.6%	0.0743	14.95	15.2%	0.0743	14.95
	NES	5.6%	0.0926	17.48	10.3%	0.0900	17.54	5.1%	0.0900	17.50	6.1%	0.0900	17.49	17.3%	0.0900	17.54	8.9%	0.0905	17.51
	SPSA	**36.4%**	0.0582	13.36	**70.1%**	0.0575	13.03	51.3%	0.0623	13.95	28.6%	0.0572	13.23	**77.8%**	0.0572	13.13	52.8%	0.0585	13.34
	AutoZOOM	35.4%	0.0791	11.76	69.1%	0.0870	11.70	**55.4%**	0.0458	11.34	37.8%	0.0948	14.35	77.3%	0.0515	9.05	55.0%	0.0716	11.64
	GATN	7.2%	0.2104	25.71	16.5%	0.1654	22.40	15.9%	0.1790	23.22	10.2%	0.1989	23.66	15.7%	0.2014	25.08	13.1%	0.1910	24.01
	TSadv	1.5%	**0.0085**	**2.75**	44.8%	**0.0085**	**2.71**	12.8%	**0.0117**	3.67	7.7%	**0.0096**	**3.12**	3.2%	**0.0070**	**2.23**	14.0%	**0.0091**	**2.90**
	AdvGAN	9.7%	0.1009	15.73	14.4%	0.1241	18.28	17.9%	0.1115	17.51	11.2%	0.1282	18.71	17.8%	0.1268	20.94	14.2%	0.1183	18.23
	Diff-PGD	9.7%	0.1173	18.63	17.0%	0.1287	19.14	18.5%	0.1251	20.56	10.7%	0.1207	19.03	18.9%	0.1132	17.90	15.0%	0.1210	19.05
	GE-AdvGAN	9.2%	0.1274	23.73	21.6%	0.1320	26.74	22.1%	0.1153	21.87	12.8%	0.1215	23.20	22.2%	0.1207	24.38	17.6%	0.1234	23.98
	SGAN	12.3%	0.0594	11.61	20.6%	0.0636	10.04	26.7%	0.0787	10.31	17.3%	0.0749	10.93	25.9%	0.0595	12.98	20.6%	0.0672	11.17
	C2FDA	33.8%	0.0471	3.71	68.0%	0.0396	3.29	53.3%	0.0319	**3.03**	**46.4%**	0.0109	3.15	76.2%	0.0265	3.25	**55.6%**	0.0312	3.29

Table 9: Ablation results on the additional datasets. Each entry reports ASR (%) / MSE / DTW.

Dataset	Full Model	w/o Coarse	w/o Fine	w/o C2F Attn.	Random Intervals	Exact Margin Only
Car	**86.7** / 0.0126 / **26.67**	73.3 / 0.0218 / 33.14	66.7 / **0.0116** / 30.71	80.0 / 0.2035 / 152.76	40.0 / 0.0244 / 36.15	73.3 / 0.0177 / 37.42
CinCECGTorso	72.3 / **0.0104** / **70.82**	61.2 / 0.0239 / 87.23	49.6 / 0.0129 / 76.48	**74.1** / 0.1587 / 246.13	40.3 / 0.0208 / 83.74	73.4 / 0.0241 / 84.55
ECG5000	**100.0** / 0.0221 / **13.38**	74.3 / 0.0403 / 20.21	75.4 / **0.0208** / 16.84	91.9 / 0.1854 / 70.52	44.2 / 0.0407 / 30.37	89.6 / 0.0355 / 18.46
ECGFiveDays	**100.0** / 0.0177 / **11.04**	98.3 / 0.0286 / 12.86	68.9 / **0.0142** / 12.79	94.4 / 0.1798 / 69.84	47.5 / 0.0354 / 23.18	88.70 / 0.0412 / 26.37
Earthquakes	**94.8** / 0.0944 / 44.45	87.0 / 0.1262 / 55.48	67.5 / **0.0848** / **43.27**	81.8 / 0.1614 / 144.96	54.5 / 0.1038 / 50.83	84.4 / 0.1072 / 49.71
FordA	87.3 / 0.0127 / **20.63**	77.3 / 0.0431 / 30.94	65.1 / **0.0105** / 24.56	**90.7** / 0.1459 / 146.42	40.7 / 0.0315 / 36.14	72.4 / 0.0271 / 27.84
FordB	**90.3** / 0.0127 / **22.43**	80.7 / 0.0248 / 31.18	62.9 / **0.0116** / 28.25	85.5 / 0.1527 / 153.18	37.6 / 0.0149 / 33.51	65.7 / 0.0271 / 29.42
Trace	**72.5** / **0.0092** / **13.31**	65.0 / 0.0364 / 26.47	47.5 / 0.0108 / 18.38	**72.5** / 0.1242 / 90.79	27.5 / 0.0187 / 20.25	60.0 / 0.0204 / 17.61
ItalyPowerDemand	**87.3** / 0.0403 / **1.24**	78.4 / 0.0659 / 4.51	69.0 / **0.0315** / 1.67	59.2 / 0.1096 / 9.17	41.3 / 0.0604 / 4.17	73.2 / 0.0736 / 6.52
SonyAIBORobotSurface1	**71.8** / 0.0250 / **2.86**	62.1 / 0.0475 / 4.24	52.4 / **0.0227** / 3.68	64.5 / 0.1169 / 20.26	16.9 / 0.0518 / 3.85	60.5 / 0.0443 / 4.67
SonyAIBORobotSurface2	53.3 / 0.0319 / **3.03**	46.2 / 0.0598 / 4.88	36.9 / **0.0281** / 4.09	**55.4** / 0.1213 / 19.37	24.1 / 0.0499 / 4.72	45.6 / 0.0512 / 4.89
Average (11 datasets)	**83.30** / 0.0263 / **20.90**	73.08 / 0.0471 / 28.29	60.19 / **0.0236** / 23.70	77.27 / 0.1509 / 102.13	37.70 / 0.0411 / 29.72	71.53 / 0.0427 / 27.95

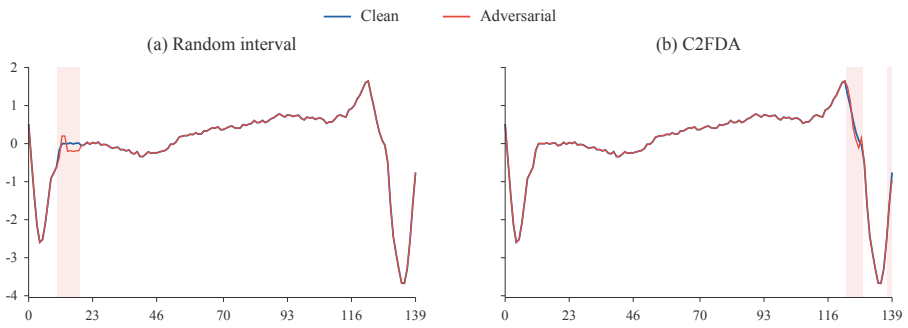


Figure 12: Random and learned temporal support on an ECG5000 sample.

Table 10: Defense results on the additional datasets (ASR / DSR, %).

Dataset	FGSM ASR/DSR	PGD ASR/DSR	MI-FGSM ASR/DSR	NES ASR/DSR	SPSA ASR/DSR	AutoZOOM ASR/DSR
Car	53.3 / 87.4	6.7 / 100.0	13.3 / 49.6	53.3 / 87.4	100.0 / 80.0	100.0 / 40.0
CinCECGTorso	11.5 / 87.8	9.0 / 75.6	9.4 / 76.6	10.4 / 100.0	50.4 / 80.8	53.6 / 52.4
ECG5000	3.9 / 66.7	2.2 / 63.6	3.2 / 59.4	1.9 / 100.0	53.6 / 60.3	60.3 / 45.8
ECGFiveDays	5.1 / 66.7	4.0 / 72.5	4.5 / 62.2	7.9 / 92.4	48.0 / 64.8	52.5 / 54.9
Earthquakes	1.3 / 100.0	1.3 / 100.0	2.6 / 100.0	7.8 / 83.3	93.5 / 76.4	94.8 / 53.4
FordA	27.1 / 74.9	17.1 / 49.1	22.0 / 62.3	25.9 / 74.1	99.7 / 63.5	99.9 / 40.3
FordB	54.1 / 72.5	23.5 / 43.0	33.3 / 45.6	53.2 / 72.2	99.7 / 70.5	99.7 / 51.8
Trace	10.0 / 50.0	0.0 / –	2.5 / 100.0	17.5 / 71.4	40.0 / 62.5	42.5 / 41.2
ItalyPowerDemand	45.1 / 64.5	46.0 / 54.1	44.6 / 51.6	8.0 / 41.3	88.3 / 54.8	85.9 / 48.7
SonyAIBORobotSurface1	23.4 / 55.1	25.0 / 38.8	24.2 / 46.7	8.1 / 40.7	67.7 / 54.8	75.8 / 37.2
SonyAIBORobotSurface2	24.6 / 60.6	26.2 / 47.3	26.2 / 47.3	5.1 / 29.4	51.3 / 53.0	55.4 / 46.4

Dataset	GATN ASR/DSR	TSadv ASR/DSR	AdvGAN ASR/DSR	Diff-PGD ASR/DSR	GE-AdvGAN ASR/DSR	SGAN ASR/DSR	C2FDA ASR/DSR
Car	13.3 / 100.0	80.0 / 8.4	80.0 / 91.6	60.0 / 88.8	53.3 / 87.4	66.7 / 30.0	86.7 / 15.5
CinCECGTorso	28.1 / 80.8	46.4 / 54.3	75.5 / 78.1	69.8 / 79.9	62.9 / 85.1	79.9 / 50.1	72.3 / 19.4
ECG5000	100.0 / 96.8	6.7 / 20.9	86.1 / 78.3	63.7 / 86.2	64.8 / 72.7	73.7 / 42.6	100.0 / 15.3
ECGFiveDays	11.9 / 66.4	53.7 / 13.8	52.5 / 73.1	45.8 / 86.5	24.3 / 88.5	35.6 / 47.8	100.0 / 11.3
Earthquakes	3.9 / 100.0	23.4 / 50.0	0.0 / –	3.9 / 100.0	0.0 / –	0.0 / –	94.8 / 39.8
FordA	10.0 / 82.0	93.9 / 19.0	64.8 / 73.0	62.3 / 75.0	60.8 / 69.2	71.6 / 43.2	87.3 / 11.1
FordB	60.3 / 78.1	99.3 / 12.5	80.6 / 66.5	74.3 / 73.5	71.4 / 81.5	84.1 / 27.9	90.3 / 16.3
Trace	0.0 / –	5.0 / 50.0	47.5 / 78.9	50.0 / 70.0	42.5 / 64.7	60.0 / 54.2	72.5 / 20.7
ItalyPowerDemand	12.2 / 54.1	18.8 / 17.6	27.7 / 57.8	16.9 / 44.4	24.9 / 49.0	20.2 / 39.6	87.3 / 8.0
SonyAIBORobotSurface1	43.5 / 64.8	22.6 / 21.7	19.4 / 66.5	23.4 / 44.9	21.8 / 55.5	22.6 / 39.4	71.8 / 15.7
SonyAIBORobotSurface2	15.9 / 83.6	12.8 / 28.1	17.9 / 89.4	18.5 / 53.0	22.1 / 56.1	26.7 / 27.0	53.3 / 15.4

These plots should be interpreted together with ASR, MSE, and DTW: MSE measures pointwise squared distortion, DTW captures temporal-shape displacement, and neither alone measures whether the perturbed support is decision relevant. The figures are therefore qualitative illustrations of the complete quantitative results rather than standalone perceptual evidence.

Visualization of the reverse-generation process. Figure 13 visualizes input-anchored DDIM generation on a UWaveGestureLibrary sample. At selected reverse steps, we visualize the clean-state estimate $\hat{\mathbf{x}}_0^{(\tau)}$, its residual $\hat{\mathbf{x}}_0^{(\tau)} - \mathbf{x}$, and the candidate temporal support computed from the current coarse-to-fine scores. The bottom rows show the final adversarial example and its applied support.

Across the displayed steps, the residuals change in amplitude and temporal structure, while the candidate support shifts along the sequence. The candidate masks are diagnostic outputs and are not applied to the DDIM state updates in equation 30, so the intermediate estimates are not a sequence of constrained adversarial examples. The final binary mask restricts the applied perturbation to 16 of 315 timestamps. The final output preserves the clean waveform outside the selected support, with local deviations around descending and rising portions of the sequence. The trajectory illustrates the coupling of *where* and *what*: perturbation content and temporal relevance are recomputed within the same generator, allowing support selection to evolve alongside synthesis before the final local perturbation is formed.

Hybrid Margin Consistency and Cross-Model Transfer

Figure 14 compares exact-only ($\gamma = 0$) and hybrid ($\gamma = 0.4$) boundary objectives on the same 177 ECGFiveDays samples, all initially classified correctly by the ResNet surrogate and target CNN. The horizontal axis shows the objective margin μ_γ, and the vertical axis shows the exact surrogate margin μ_{adv}. Negative values indicate boundary crossing under the corresponding margin criterion; marker type separately indicates attack success on the target CNN.

With $\gamma = 0$, the two margins coincide by definition, so all points lie on the diagonal. Although all 177 samples cross the surrogate boundary, only 157 transfer successfully to CNN (88.70% ASR). The remaining 20 samples illustrate that optimizing the exact surrogate margin to cross its boundary does not necessarily produce a successful cross-model attack. With $\gamma = 0.4$, all 177 points lie in

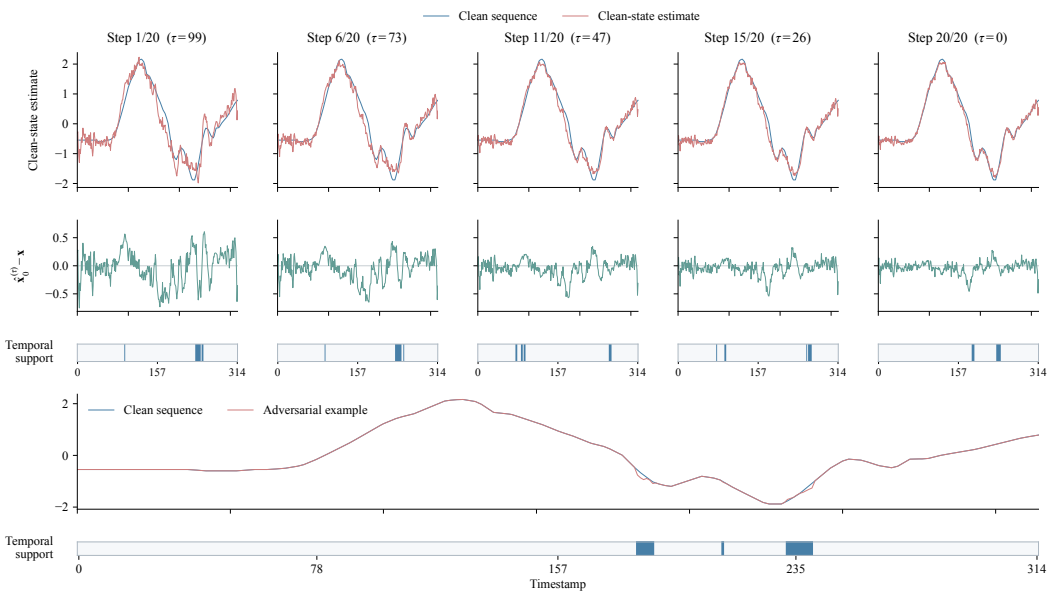
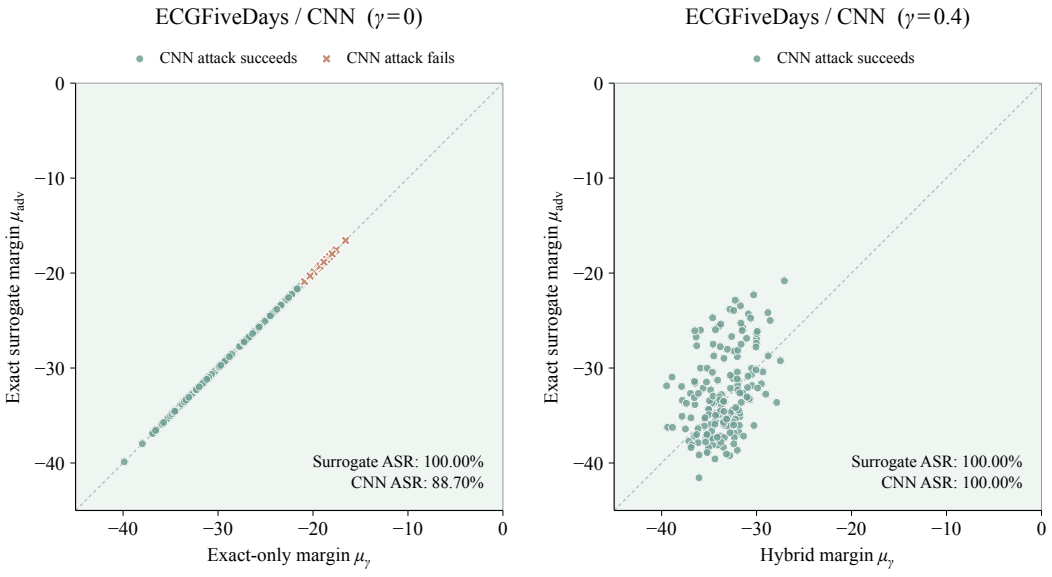


Figure 13: Reverse-generation process on a UWaveGestureLibrary sample. Intermediate rows show clean-state estimates, residuals relative to the clean input, and candidate temporal support; bottom rows show the final adversarial example and applied support.



(a) Exact-only objective ($\gamma = 0$). (b) Hybrid objective ($\gamma = 0.4$).

Figure 14: Objective margin versus exact surrogate margin on ECGFiveDays. Each panel contains the same 177 samples. Green circles denote successful transfer to CNN and orange crosses denote failed transfer; the dashed line marks equality of the two margins. Both objectives achieve 100% surrogate ASR, while CNN ASR is 88.70% for $\gamma = 0$ and 100% for $\gamma = 0.4$.

the lower-left quadrant: the hybrid and exact margins agree on boundary crossing, with no sample having a negative hybrid margin but a nonnegative exact margin. Their spread away from the diagonal indicates agreement in sign rather than numerical equality. All 177 samples also succeed on CNN (100% ASR), giving an 11.30-percentage-point increase over the exact-only objective. CNN transfer is measured separately by target evaluation.