SAURON: A Zero-Shot Network Digital Twin for 5G Traffic and Threat Generation

Alejandro Calvillo-Fernandez
Dept. of Telematic Engineering
Universidad Carlos III de Madrid
Leganes, Spain
acalvill@pa.uc3m.es

Noelia Pérez Palma
Dept. of Information and Comm. Eng.
Universidad de Murcia
Murcia, Spain
noelia.perez3@um.es

Constantine Ayimba
Dept. of Telematic Engineering
Universidad Carlos III de Madrid
Leganes, Spain
ayconsta@it.uc3m.es

Carlos J. Bernardos
Dept. of Telematic Engineering
Universidad Carlos III de Madrid
Leganes, Spain
cjbc@it.uc3m.es

Antonio Skármeta
Dept. of Information and Comm. Eng.
Universidad de Murcia
Murcia, Spain
skarmeta@um.es

***Abstract*—As modern 5G networks evolve, network management has become increasingly complex, making reliable operations using only raw telemetry impractical. Network digital twins (NDT) are emerging as a solution to model complex scenarios without compromising the real network. Unfortunately, historical datasets are limited, and in most cases private, restricting data-driven network management. We present SAURON (Synthetic Autoregressive Unified Replica for Operator Networks), a zero-shot NDT that synthesizes statistically faithful 5G packet traces without the need for retraining. Built on Chronos-2, a pre-trained autoregressive transformer, SAURON augments its conditional-mean forecast with a spectral diversity step that injects controlled phase noise, followed by quantile mapping that anchors every feature's marginal distribution to the empirical reference and enforces physical network constraints. Then, SAURON reconstructs the packet trace, preserving microscopic burstiness, to generate synthetic traces indistinguishable from real ones. We evaluate SAURON on real 5G captures, spanning from a DNS-flooding attack to a nominal multi-user baseline. Results shows that SAURON almost reach the theoretical lower bound of residual distributional distances, outperforming literature solutions.**

***Index Terms*—Network Digital Twin, 5G, Zero-Shot Traffic Generation, Synthetic Traces.**

I. Introduction

As cellular networks evolve beyond 5G, their capabilities, as well as the services they support, have grown in number and diversity. Consequently, network management has become more complex. Therefore, achieving reliable network management using only raw telemetry is quickly becoming impractical: the management decision space is larger, the data is partial, and most consequential events; volumetric denial-of-service campaigns, advanced persistent threats, configuration drifts that cascade into outages, are rarely included in any historical captures that form the basis of data-driven approaches. AI-driven network digital twins (NDTs) have emerged as a remedy: by replicating the network behavior in a separate computing space, an network digital twin (NDT) enables what-if analysis, root-cause investigation and proactive defense without affecting the underlying network.

However, the mainstay, and counterpart, of an NDT is data. Its fidelity is limited by the quality of the traces it ingests, and high-fidelity traces are hard to obtain for two reasons: (i) deep packet capture is intrusive: it requires probes to be placed at key points in the network to facilitate telemetry, introducing a significant overhead, and (ii) full traces are commercially and legally sensitive, so operators rarely release more than randomly subsampled headers. For routine dimensioning of resources and for the resolution of ordinary faults this is usually sufficient. However, to better predict resource demand during rare events and to detect sophisticated indicators of compromise such as those originating from advanced persistent threats, entire flows may be needed.

With this in mind, we propose **SAURON** (**S**ynthetic **A**utoregressive **U**nified **R**eplica for **O**perator **N**etworks), a zero-shot NDT that calibrates on real traffic traces and generates synthetic packet traces statistically indistinguishable from the captured reference. We adopt, as our foundational model, Chronos-2 [1], an autoregressive, $\sim$120M-parameter encoder-only transformer pre-trained on a large heterogeneous corpus. It enables zero-shot probabilistic forecasting on unseen traces, providing a strong prior of the temporal structure.

Despite being mostly reliable, Chronos-2 suffers from the following limitations when used on its own: (i) the median-quantile forecast regresses toward the conditional mean and under-estimates heavy tails, and (ii) the model is feature-agnostic and can violate physical constraints. Our pipeline addresses both shortcomings without retraining the backbone: we introduce a spectral step that injects controlled high-frequency variations, followed by a quantile-mapping step to align every feature with the empirical reference.

Comparison to state of the art: The adaptability of transformers [2] to different domains has led to their increased use in addressing a myriad of network management problems. We group these approaches into three architectural categories:

This work has been partially funded by the European Commission Horizon Europe SNS JU PRIME-6G (GA 101272485) Project, 6GINSPIRE PID2022-137329OB-C42, funded by MCIN/AEI/10.13039/501100011033/. Thanks to SLICES-Madrid (https://slices-madrid.eu/), from ESFRI SLICES-RI project. Thanks to Carlos Barroso-Fernandez for the valuable comments.

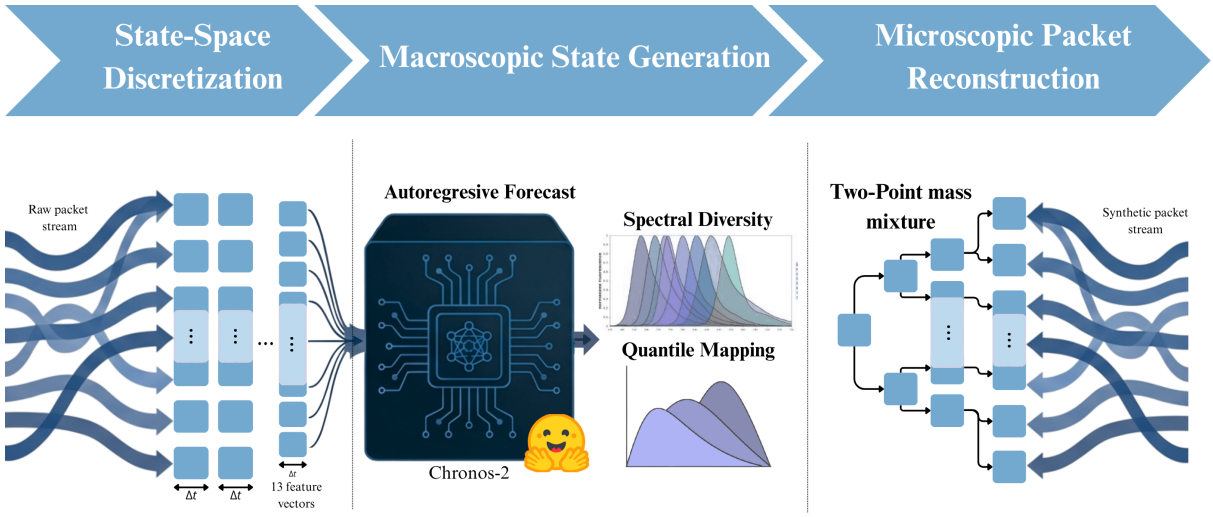


Fig. 1. SAURON's design over the three stages pipeline

(i) GAN-based generators such as DoppelGANger [3] that employs attribute-level synthesis and NetShare [4] that generates packet and flow-header traces, (ii) diffusion models such as NetDiffus [5] that treats trace generation as a score-matching problem, and (iii) autoregressive sequence models like NetSSM [6] that achieves byte-level fidelity. None of these baselines, however, addresses the structural limitations of standard deep generative models when applied to network telemetry. In contrast, SAURON bypasses the need for re-training through its zero-shot calibration. Furthermore, by introducing exact quantile mapping it prevents the statistical hallucinations and physical constraint violations typical of unconstrained architectures.

Main Contribution: Coupled with a microscopic trace reconstruction step, SAURON is, to the best of our knowledge, the first framework that most closely approaches the theoretical lower bound of the residual distributional distance, outperforming existing baselines.

The rest of the article is structured as follows: Section II describes the characteristics of our proposed framework. Section III presents the real deployment from which the traffic traces are gathered. Section IV evaluates the framework and compares the performance against literature solutions. Finally, Section V summarizes and concludes the article.

II. SAURON DESIGN

SAURON's three-stage pipeline designs as shown in Fig. 1. First, it discretizes a captured packet stream on a network edge into a sequence of macroscopic state vectors. Then, its calibrated three-step pipeline produces a synthetic stream of state vectors that preserves the marginal, spectral, and second-order statistics of the reference. Finally, each synthetic state vector is mapped back to a continuous packet sequence whose microscopic characteristics match the empirical capture.

In the descriptions that follow, real quantities are denoted plainly (e.g. x_k, r_k) whereas synthetic quantities bear a tilde (e.g. $\tilde{x}_k$). The indices for the given variables follow the convention: k for time bin, d for feature and i for sample. Table I summarizes the main notation used in the paper.

A. State-Space Discretization

Let a continuous-time 5G network edge E packet capture be defined as $\mathcal{P}^{(E)} = \{p_j\}_{j=1}^{N_p}$, where N_p is the number of packets of the capture, with $p_j \coloneqq (t_j, L_j, s_j, d_j, \pi_j)$ where t_j is the arrival timestamp, L_j is the payload length, s_j is the source IP, d_j is the destination IP, and π_j is the protocol-tag set. The data collection window $\Delta t > 0$ partitions the time axis into bins $I_k = [k\Delta t, (k+1)\Delta t)$ for $k = 1, \ldots, K_r$, where K_r is the reference stream length. We define an extraction function $\Phi : \mathcal{P}^{(E)} \to \mathbb{R}^D$ to aggregate the packet features into a sequence of macroscopic state vectors $x_k \in \mathbb{R}^D$ with $D = 13$ coordinates categorized as:

- **Volumetric**: packet count $N_k = |\{j : t_j \in I_k\}|$, byte volume $V_k = \sum_{t_j \in I_k} L_j$, and mean payload size $\mu_{L,k} = V_k / N_k$.
- **Topological**: unique sources $U_{\mathrm{src},k}$ and destinations $U_{\mathrm{dst},k}$.
- **Protocol Spectrum**: counts the protocol tag for $p \in \{\mathrm{DNS}, \mathrm{TCP}, \mathrm{UDP}, \mathrm{SCTP}, \mathrm{NGAP}, \mathrm{GTP}\}$.
- **Temporal**: the mean inter-arrival time $\mu_{\mathrm{IAT},k}$ and its standard deviation $\sigma_{\mathrm{IAT},k}$.

We write $x_k^{(d)}$ for the d-th coordinate $(d = 1, \ldots, D)$ whenever the per-feature distinction matters.

B. Macroscopic State Generation

The macroscopic pipeline produces a synthetic stream $\tilde{x}_1, \ldots, \tilde{x}_{K_s}$ from an empirical reference $r_1, \ldots, r_{K_r}$ by composing three stages applied per feature:

a) Autoregressive forecast: Underneath the framework, we integrate Chronos-2 transformer [1] as a modular forecasting component. Instead of retraining with our data, it ingests a given set of x_k and returns the median-quantile autoregressive forecast $x_{\mathrm{chronos}}[k]$ of a future window for each of the D features in parallel, tokenized through the model's mean-absolute scaler. This component provides the conditional-mean trajectory of the synthetic stream.

b) Spectral diversity: For any per-feature reference sequence $x[k]$ of length K_r we compute its discrete Fast Fourier Transform (FFT) with amplitude $A[f] = |\mathcal{F}\{x[k]\}|$

TABLE I
MAIN NOTATION USED IN THE PAPER

Notation	Parameter
$\mathcal{P}^{(E)}$	Real packet stream on edge E
p_j	Packet tuple: arrival time, payload length, source IP, destination IP, protocol-tag set
Δt	Time-bin width (s)
I_k	Time-bin $[k\Delta t, (k+1)\Delta t]$ with k index
D	Macroscopic-state dimensionality
Φ	State-extraction function $\mathcal{P}^{(E)} \to \mathbb{R}^D$
x_k	Real (captured) state vector at bin k, $\in \mathbb{R}^D$
$\tilde{x}_k$	Calibrated synthetic state vector at bin k
r	Empirical reference sample, used as the target distribution of the quantile-mapping operator
N_k	Packet count in bin k
$\mu_{\mathrm{IAT},k}$, $\sigma_{\mathrm{IAT},k}$	Mean and standard deviation of inter-arrival times in bin k
w_c	Blend weight Chronos-2 vs. spectral, $w_c \in [0,1]$
σ_ϕ	Standard deviation of the spectral phase noise
T	Quantile-mapping operator
$n_{\mathrm{spike},k}$	Inter-burst gap count of the two-point microscopic reconstruction in bin k
b_k	Gap duration of the two point microscopic reconstruction in bin k

and phase $\phi[f] = \angle\mathcal{F}\{x[k]\}$, where f indexes the frequency bins. Frequency-scaled Gaussian phase noise $\eta[f] \sim \mathcal{N}\big(0,\ \sigma_\phi^2\,(f/K_r)\big)$ is injected into every non-DC bin, preserving the amplitude spectrum while randomizing the temporal sequence,

$$x_{\text{spectral}}[k] \;=\; \max\big(\mathcal{F}^{-1}\{A[f]\,e^{j(\phi[f]+\eta[f])}\},\ 0\big). \qquad (1)$$

Where $\mathcal{F}^{-1}$ is the inverse of the FFT. The autoregressive and spectral streams are then blended with a fixed weight $w_c \in [0,1]$,

$$x_{\text{raw}}[k] \;=\; w_c \cdot x_{\text{chronos}}[k] \;+\; (1-w_c)\cdot x_{\text{spectral}}[k]. \qquad (2)$$

c) Quantile Mapping: One well-known problem of transformers sequence generator are hallucinations [7]. To prevent this issue and mimic the reality of the forecasted sequence, we introduce a distributional calibration based on quantile-mapping (QM).

Let $\hat{F}_{\text{raw}}$ and $\hat{F}_r$ be the empirical Cumulative Distribution Functions (CDFs) of x_{raw} and of the reference r, and let $\hat{F}_r^{-1} : (0,1) \to \mathbb{R}$ be the empirical reference quantile function, defined for any quantile level $u \in (0,1)$ as $\hat{F}_r^{-1}(u) = \inf\{x \in \mathbb{R} : \hat{F}_r(x) \geq u\}$. The QM operator is the composition,

$$T(x) \;=\; \hat{F}_r^{-1}\big(\hat{F}_{\text{raw}}(x)\big), \qquad (3)$$

applied per feature. In the discrete implementation, the empirical quantile level of the i-th synthetic sample is $u_i = \mathrm{rank}(x_{\text{raw},i})/(K_s+1) \in (0,1)$, and the calibrated value is read from the order statistics $r_{(1)} \leq \cdots \leq r_{(K_r)}$ of the reference,

$$\tilde{x}_i \;=\; r_{\left(\lfloor u_i \cdot K_r \rfloor + 1\right)}, \qquad i = 1, \ldots, K_s. \qquad (4)$$

Two properties of T follow directly. *(P1) Marginal preservation*: $\sup_x |\hat{F}_{\tilde{x}}(x) - \hat{F}_r(x)| \leq 1/\min(K_s, K_r)$, so any distributional distance between $\tilde{x}$ and a fresh sample from the population of r is bounded above by the natural drift between two finite samples of that population, independently of the autoregressive backbone. *(P2) Rank preservation*: T is monotonically non-decreasing, so the temporal ordering of x_{raw} is preserved, and the autocorrelation envelope generated by Eq. (1) survives the mapping.

Therefore, if the physical network only allows $N_k = 1,000$ packets per time bin, and the autoregressive forecast predicts $N_k = 1080$, the value will be clipped to $N_k = 1000$. The QM is applied independently per feature and is therefore agnostic to joint constraints. The state vector contains one hard constraint imposed by the bin geometry, $\mu_{\mathrm{IAT},k} \cdot N_k \leq \Delta t$, which a naive per-feature QM occasionally violates because the upper tails of $\mu_{\mathrm{IAT},k}$ and N_k are sampled independently. We close this with a final deterministic clip

$$\tilde{\mu}_{\mathrm{IAT},k} \;=\; \min\!\left(T(\mu_{\mathrm{IAT},k}),\ \frac{0.999\,\Delta t}{N_k}\right), \qquad (5)$$

where the 0.999 factor reserves head-room for the second-moment reconstruction described next.

C. Microscopic Packet Reconstruction

Given a calibrated state vector $\tilde{x}_k$ at bin k, the microscopic step maps the triple $(N_k,\, \tilde{\mu}_{\mathrm{IAT},k},\, \tilde{\sigma}_{\mathrm{IAT},k})$ back to a sequence of N_k inter-arrival times whose sample mean and standard deviation match $\tilde{\mu}_{\mathrm{IAT},k}$ and $\tilde{\sigma}_{\mathrm{IAT},k}$ exactly. To avoid the sampling-noise destruction that continuous draws cause on bins with small N_k, we use a deterministic two-point mass mixture: $n_{\mathrm{spike},k}$ inter-burst gaps of duration b_k and $n_{\mathrm{base},k} = N_k - n_{\mathrm{spike},k}$ intra-burst gaps of duration $\epsilon \to 0^+$. Let $\widetilde{CV}_k = \tilde{\sigma}_{\mathrm{IAT},k}/\tilde{\mu}_{\mathrm{IAT},k}$ be the target coefficient of variation. The spike count is fixed by the second-moment constraint $\widetilde{CV}_k^{\,2} = \mathrm{Var}/\mathrm{mean}^2$ applied to a two-point mass,

$$n_{\mathrm{spike},k} \;=\; \left\lceil \frac{N_k^2}{(N_k-1)\,\widetilde{CV}_k^{\,2} + N_k} \right\rceil, \qquad (6)$$

and the spike duration b_k is fixed by the first-moment constraint $\mathbb{E}[\mathrm{IAT}] = \tilde{\mu}_{\mathrm{IAT},k}$,

$$b_k \;=\; \frac{N_k\,\tilde{\mu}_{\mathrm{IAT},k} \;-\; n_{\mathrm{base},k}\,\epsilon}{n_{\mathrm{spike},k}}. \qquad (7)$$

Both moments are matched analytically by construction: the reconstruction has zero sampling variance and is therefore lossless on the $(N_k, \tilde{\mu}, \tilde{\sigma})$ round-trip, a property we verify empirically in Section IV-B. The N_k inter-arrival times are then placed inside I_k in a random order and the corresponding packets are generated with the protocol-tag distribution dictated by the calibrated counters $C_{p,k}$, completing the mapping from the discrete state space back to a continuous packet trace ready for downstream detector evaluation.

III. TOPOLOGY AND DATA ACQUISITION

The real network on which the NDT is based comes from a scalable 5G testing infrastructure that generates traffic from virtual 5G User Equipments (UEs). This ensures that the scenarios can be scaled up to mimic real-world deployments and not be limited to a few physical UEs. Our topology enables non-5G devices to connect to a real 5G Radio Access Network (RAN) through a reactive virtualization layer (i.e., a gateway that spawns a virtual UE on-demand in response to each

new client's Dynamic Host Configuration Protocol (DHCP) request), ensuring the resulting datasets capture authentic protocol behaviors. This topology is more detailed in [8].

A. Infrastructure Architecture

The framework is built upon a multi-layer stack that integrates general-purpose devices into a 3GPP-compliant environment. The layout of the testbed components, including the 5G RAN and the core processing units, is depicted in Fig. 2:

- Entry Point: Heterogeneous non-5G devices, such as smartphones or Personal Computers (PCs), connect via standard Wi-Fi or Ethernet to a Reactive Multi-Technology 5G Network Gateway.
- Dynamic Virtualization: Upon detecting a client through DHCP inspection, the gateway instantiates a Virtual UE (vUE) isolated within a Linux network namespace, providing dedicated 5G connectivity for that specific client.
- 5G RAN and Core: The vUEs interact with a physical gNodeB (Amarisoft Callbox, 3GPP R16) transmitting in the 5G NR band over-the-air through its RF front-end and an Open5Gs core [8], establishing standard 5G interfaces such as N2 for control signaling and N3 for GPRS Tunneling Protocol (GTP)-encapsulated user data. The radio link is not emulated: the gNB operates over a real 10 MHz NR channel, with the RF hardware setup documented in [8].

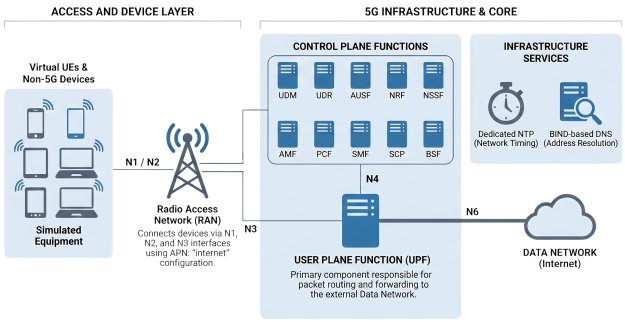


Fig. 2. 5G Network Topology and Data Acquisition Architecture.

B. Data Extraction Methodology

To provide the necessary traces for the synthetic engine, the framework uses a Remote Traffic Capture Probe. This tool facilitates the synchronized collection of packets across multiple observation points:

- Control Plane (CP) Capture: Traffic is monitored at the N1/N2 interfaces, recording Non-Access Stratum (NAS) and NG Application Protocol (NGAP) signaling over Stream Control Transmission Protocol (SCTP). This data allows for the modeling of temporal patterns related to registration, session establishment, and radio resource management.
- User Plane (UP) Capture: Data is extracted at the N3 (encapsulated) and N6 (decapsulated) interfaces. This process captures the volumetric and microscopic variance of application traffic, such as video streaming or file transfers, as it traverses the 5G infrastructure.

C. Data Acquisition for Digital Twin Modeling

How faithfully the NDT reflects the reality of the network it models depends on the quality of the data used during the state-space discretization phase. Using the testbed of Section III-A and the methodology of Section III-B, we collected two primary categories of traces to model both adversarial traffic (cyberthreats) and innocuous traffic from legitimate 5G users.

1) Adversarial DNS Flooding Scenarios: We collected Packet Capture (PCAP) traces from controlled experiments of a Domain Name System (DNS) flooding attack and a corresponding mitigation mechanism based on DNS rate limiting. These traces represent different operational conditions of the DNS infrastructure with respect to the activation of the mitigation mechanism. All captures correspond to observation intervals of approximately one minute, enabling a consistent comparison of traffic characteristics across scenarios. In Table II, we summarize the main characteristics of the traces we used for NDT calibration and for behavioral modeling.

- Attack Scenario (No Mitigation): This trace captures an unmitigated DNS flooding attack. The target server receives a continuous stream of DNS queries generated at a high rate by the attacking source(s). It contains 1,208,016 packets captured over 64.53 seconds (approx. 18,700 packets per second). The high packet rate is typical of this type of attack aiming to overload the DNS server. This trace serves as the reference for peak-load saturation.
- Rate-Limited Attack (Mitigated): This trace captures the same DNS flooding attack, but with a mitigation mechanism based on DNS rate limiting at the authoritative DNS server using the functionality available in BIND9 [9]. The mechanism constrains the rate at which DNS queries are processed when repetitive query patterns exceed predefined thresholds. As a result, the effective query processing rate is significantly reduced. This trace contains 944,996 packets captured over 68.23 seconds (approx. 13,800 pps). The reduction in volume reflects the impact of rate limiting.
- Transitional Attack Scenario: This trace records the moment when the mitigation mechanism is activated during the traffic capture. Initially, the DNS server operates without defensive mechanisms, producing traffic characteristics consistent with the unmitigated flooding scenario. At a specific point during the capture interval, the rate limiting mechanism is enabled, resulting in a noticeable change in the traffic dynamics. The trace contains 1,051,451 packets over a duration of 65.71 seconds, with an average packet rate of approximately 16.0 kpps. This combined trace allows the model to learn the dynamics of system recovery across both operational states within a single continuous measurement.

2) Nominal Behavioral Baselines: To ensure the high-fidelity calibration of the NDT, two distinct baseline datasets were collected to capture the non-stationary periodicity of nominal 5G traffic under different operational scales:

a) Synchronized Single-UE Baseline (June 25th, 2024): This dataset provides a high-resolution snapshot of the causal

relationship between the 5G control and user planes. Captured on June 25th, 2024, it records a 2-hour window of non-intensive activity from a commercial smartphone connected to the testbed:

- Control Plane (N2 Interface): This component documents the underlying NGAP and NAS signaling overhead required for Protocol Data Unit (PDU) session maintenance and UE context management.
- User Plane (N3 Interface): Captured over GTP-U, the traffic reflects standard mobile behavior, including messaging (WhatsApp), web browsing with HyperText Transfer Protocol Secure (HTTPS), and media streaming via YouTube and Spotify.

b) Scalable Multi-User Baseline (June 18th, 2024): To model aggregate network behavior, an extended 4-hour trace was collected involving eight simultaneous vUEs. In this setup, traffic from real users was redirected through a reactive virtualization layer to the 5G core.

- User Plane (N3 Interface): Captured at the gNB-UPF boundary, the User Plane Function (UPF) manages the assigned vUE address range as defined in the network topology (see Fig. 2).
- Protocol Diversity: The dataset features a high prevalence of Quick UDP Internet Connections (QUIC) and HTTPS for Google services, User Datagram Protocol (UDP) for Instagram and Jitsi video conferencing, and Transmission Control Protocol (TCP)/Session Traversal Utilities for NAT (STUN) for instant messaging platforms like Telegram and WhatsApp.
- Microscopic Integrity: Because the capture chain is end-to-end real (over-the-air transmission from the Amarisoft Callbox gNB, full 3GPP user-plane stack (RLC/PDCP/SDAP/GTP-U), and TCP segments generated by the actual kernel of the client device) channel-induced artifacts such as RLC-triggered TCP out-of-order segments, retransmissions, and inter-arrival jitter appear naturally at the N3 capture point. No artifact is injected or modeled, the probe records the wire timing produced by the real radio scheduler and the real client congestion-control loop, which is what the Synthetic Engine consumes to derive empirical inter-arrival moments and the per-bin burstiness profile.

By incorporating both single-device signaling and multi-user volumetric patterns, this combined baseline allows the NDT to maintain structural and statistical integrity during the subsequent injection of adversarial DNS flooding profiles.

The significant variance in traffic density between adversarial and nominal scenarios is summarized in Table II. As observed, the unmitigated attack exhibits a packet rate several orders of magnitude higher than the multi-user baseline, providing a clear volumetric boundary for the NDT calibration. Furthermore, the inclusion of long-term nominal traces ensures the model captures the heavy-tailed distributions typical of human-driven 5G traffic.

TABLE II
CHARACTERISTICS OF BASELINE DATASETS

Category	Scenario / Interface	Packets	Duration	Rate (pkt/s)
Adversarial	Attack (no mitigation)	1,208,016	64.53 s	18,720
	Rate-limited attack	944,996	68.23 s	13,850
	Attack + mitigation (Transition)	1,051,451	65.71 s	16,000
Nominal	Multi-User Baseline (N3)	4,835,726	4.00 h	335.8
	Single-UE Baseline (N2+N3)	∼124,000	2.00 h	∼17.2

TABLE III
AGGREGATE MACROSCOPIC FIDELITY ACROSS THE 13 FEATURES OF Φ ON THE HELD-OUT WINDOW

	Comparison	KS_{mean}	JS_{mean}	r	KS@0.05
Adversarial	Synth. vs. test	0.117	0.271	0.837	13/13
	Irreducible drift	0.110	0.263	0.834	13/13
Nominal	Synth. vs. test	0.078	0.123	0.983	5/13
	Irreducible Drift	0.075	0.121	0.981	6/13
Cross-trace	no calibration	0.696	0.563	0.738	3/13

IV. SYSTEM VALIDATION

The full workflow of SAURON is deployed end-to-end (E2E) on an *NVIDIA DGX Spark*, which also hosts the Chronos-2 agent engine. Real packet captures are ingested, aggregated and stored on a separate companion server with an AMD Ryzen 9 7900 (12-core, 3.7 GHz) CPU and 32 GB of RAM. This setup mirrors a realistic deployment of a telco operator in which raw traffic collection is decoupled from generative inference.

For the evaluation, we split each of the captured traces in time. The first half is used as the calibration context for Chronos-2 and the second half serves as the held-out window for validating the predicted series; SAURON then generates a 20s synthetic stream. We set the deployment hyperparameter to $w_c = 0.80$. The resulting synthetic stream is then compared with the held-out window. Notice that the held-out window is never observed during calibration. We evaluate two real captures spanning the operating regime of the framework. The *adversarial* case is a DNS-flooding attack (detailed in Section III); we restrict the evaluation to the $40\,\mathrm{s}$ steady state plateau (80 bins of $\Delta t = 0.5\,s$, 40 calibration bins, 40 held-out bins) so that the comparison is not contaminated by the onset and ramp-down transients. The *nominal* case is a 16 minute window ($1{,}920$ bins of $\Delta t = 0.5\,s$; 960 calibration, 960 held-out) of the multi-user N3 baseline of Section III.

A. Macroscopic Fidelity

For every feature in the extraction function Φ, we compute the four metrics of distributional and temporal agreement: Kolmogorov–Smirnov (KS), Jensen–Shannon (JS), the Pearson correlation, and the fraction of features that pass the two sample KS test at the $\alpha = 0.05$ level.

Notice that the quantile-mapping step of Section II forces the empirical marginal of the synthetic stream to coincide with that of the calibration window by construction. Consequently, $\mathrm{KS}(\tilde{x}_{\text{synth}}, x_{\text{held-out}}) \geq \mathrm{KS}(x_{\text{cal}}, x_{\text{held-out}})$. The right-hand side is the irreducible drift of the captured trace between the two windows. This bound holds not only for KS, but also for JS and the Pearson correlation. Thus, this is a property of the data, not of the pipeline. We therefore report both quantities, and the pipeline operates at its theoretical optimum when the

two coincide. As an absolute reference, we additionally report the same statistic between two unrelated real captures of the same length (the adversarial held-out window and a matching nominal slice); this is the "no-calibration" baseline.

Table III reports the aggregate fidelity. The synthetic forecast is within $\Delta\mathrm{KS} = 0.007$ of the irreducible drift on the adversarial regime, and within $\Delta\mathrm{KS} = 0.003$ on the nominal regime; the same gap on JS is below 0.01 in both cases. SAURON therefore attains near-optimum performance regarding these two metrics.

The Pearson correlation r is also at the bound in the adversarial regime (0.837 vs. 0.834 drift) and within 0.002 on the nominal regime (0.983 vs. 0.981); the small residual is the information loss introduced by the spectral phase perturbation $\eta[k]$ (Eq.1). The KS@0.05 on the adversarial regime is $13/13$ for both comparisons. The lower nominal pass-rate ($5/13$) is purely a statistical effect of the much larger sample size. With 960 test bins the KS test is sensitive to sub-percent distributional shifts that the irreducible drift also exhibits ($6/13$), confirming that the failures originate in the trace's multi-minute heterogeneity rather than in the NDT. However, this phenomenon warrants further study and we intend to tackle it in future work.

Overall, adversarial and nominal regimes are nearly an order of magnitude below the cross-trace baseline ($\mathrm{KS} = 0.696$, $\mathrm{JS} = 0.563$): the calibrated pipeline closes the distributional gap to the held-out trace by a factor of $5.9\times$ in the adversarial regime and $7.2\times$ in the nominal regime in KS, and by $2.1\times$ and $4.6\times$ in JS, relative to an uncalibrated control.

In order to compare our solution with the state-of-the-art, we retrain three multivariate-time-series generators that the synthetic traffic literature uses as standard baselines: (i) a TimeGAN [10], (ii) a Wasserstein-GAN with gradient penalty (WGAN-GP) [11], and (iii) a Long Short-Term Memory (LSTM) Time Series Variational Autoencoder (TVAE) [12], on the same calibration window SAURON consumes, and evaluate them on the same held-out window.

To eliminate hyperparameter under-tuning, we run a multi-seed grid search over learning rate, hidden dimensions, sequence length, latent dimensions, training epochs and, for WGAN-GP, the critic-to-generator ratio described in [11] ($\sim 10^2$ configurations per baseline, 5 seeds at the best configuration), and report the best mean$\pm$std attained by each model. NetShare [4] and NetSSM [6] are not retrained because their public implementations are target IP-byte-level. We restrict head-to-head evaluation to baselines that operate on the same feature space.

Fig. 3 shows the comparison between our solution and the state-of-the-art in terms of KS. Table IV expands this comparison using the same metrics used in Table III. Our framework outperforms all baselines by $\Delta\mathrm{KS} = 0.2$. This is because the quantile mapping step described in Section II forces the marginal of the synthetic stream to coincide with the physical limitations of the system it tries to mimic; consequently, SAURON exhibits an upper bound on its KS despite having an autoregressive component. The three generative baselines,

TABLE IV
HEAD-TO-HEAD AGAINST STATE-OF-THE-ART GENERATIVE BASELINES ON THE HELD-OUT WINDOWS

Adversarial (cal=40 bins)				
Method	KS	JS	r	KS@0.05
TimeGAN [10]	$0.39\pm.11$	0.54	0.75	$6/13$
WGAN-GP [11]	$0.32\pm.04$	0.56	0.79	$8/13$
TVAE [12]	$0.31\pm.04$	0.50	0.80	$8/13$
SAURON	**0.117**	**0.27**	**0.84**	**13/13**
Drift bound	0.110	0.26	0.83	$13/13$
Nominal (cal=960 bins)				
Method	KS	JS	r	KS@0.05
TimeGAN [10]	$0.56\pm.18$	0.45	0.78	$1/13$
WGAN-GP [11]	$0.36\pm.01$	0.37	0.95	$0/13$
TVAE [12]	$0.41\pm.02$	0.34	0.91	$0/13$
SAURON	**0.078**	**0.12**	**0.98**	**5/13**
Drift bound	0.075	0.12	0.98	$6/13$

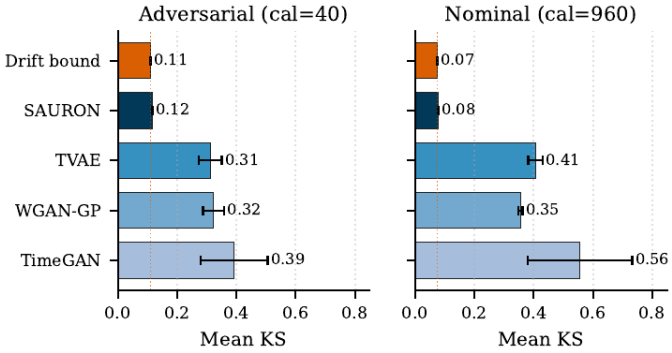


Fig. 3. Head-to-head KS and JS on the held-out windows for the three retrained baselines, the DT, and the irreducible drift bound.

in contrast, learn an unconditional distribution over the 13-dimensional state and have no explicit mechanism that anchors their marginals to the calibration: their KS is bounded only by training quality and model capacity. As a consequence, no amount of additional hyperparameter tuning or training compute would allow them to reduce the drift bound, without introducing the QM step SAURON has; the gap reported is therefore not a tuning artifact but a structural property of our solution.

Even after the grid search, on the adversarial scenario the three solutions close at $\mathrm{KS} \in [0.31, 0.39]$, $2.7\times$ to $3.4\times$ further from the held-out reference. The KS pass-rate (KS@0.05) of TimeGAN, WGAN-GP and TVAE remains between 0 and 1 out of 13, while their JS divergence ranges from 0.34 to 0.45. SAURON almost achieves the theoretical optimum for both cases. This is because deep baselines learn an unconditional distribution over a 13-dimensional time series whose components include sparse protocol counts (SCTP, NGAP) that are zero in most bins, while our solution inherits Chronos-2's pre-trained temporal prior and anchors the marginal of every feature to the empirical calibration marginal via the quantile mapping step (see Section II), which none of the other solutions provides.

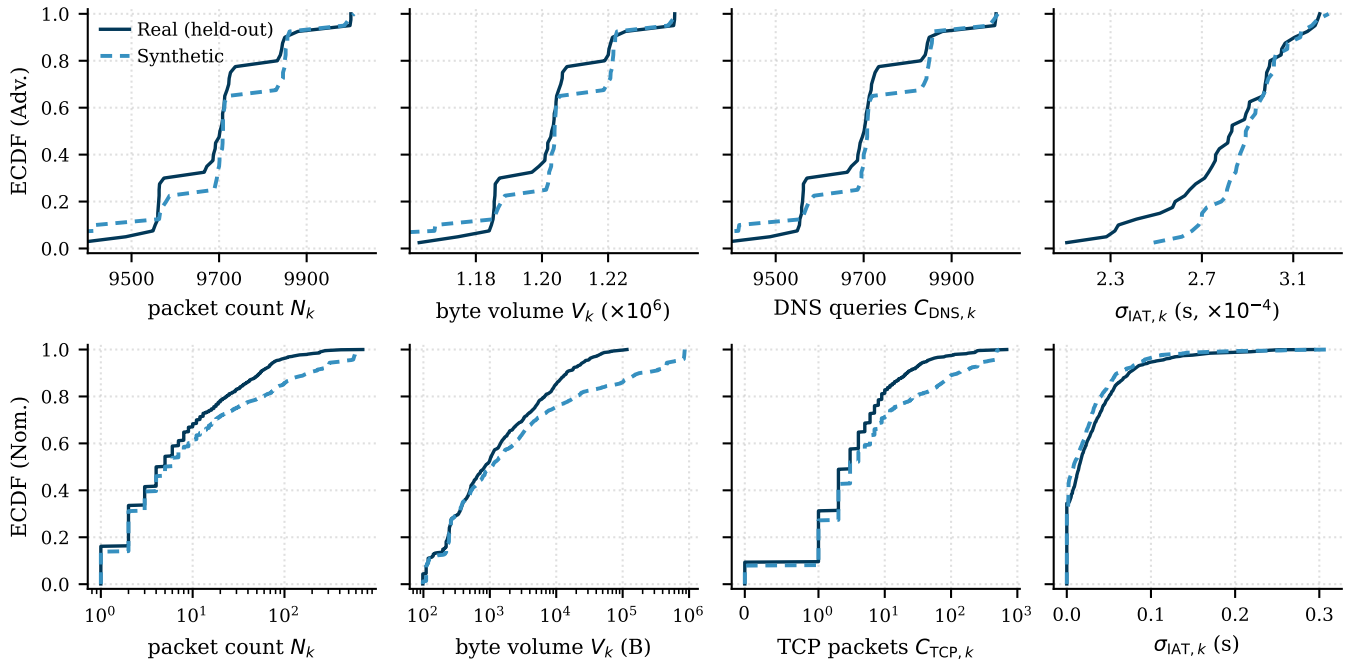


Fig. 4. Empirical CDF overlays of four representative features on the adversarial and nominal held-out window.

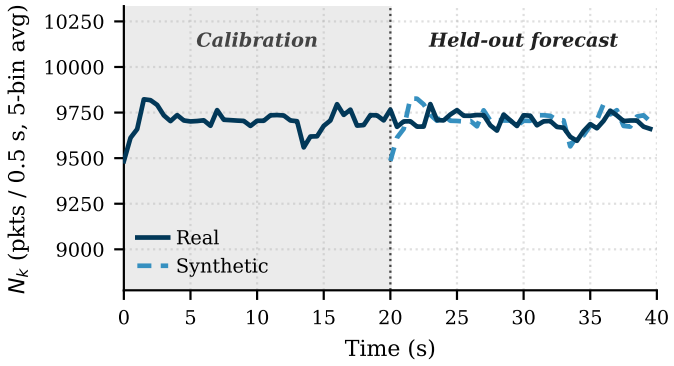


Fig. 5. Per-bin packet count N_k over the 40 s adversarial plateau, 5-bin moving average. The grey region marks the 20 s calibration prefix the NDT observes; the white region marks the 20 s held-out forecast window which the NDT never sees during calibration.

For reference, NetShare's [4] reports JS values of 0.16–0.18 and NetSSM [6] reaches ∼0.02 at byte granularity. Although not a strict comparison, the NDT's nominal JS = 0.12 falls below NetShare's reported range. NetSSM's tighter byte-level fidelity comes at the cost of a narrower feature space and a specialized pre-trained corpus.

Fig. 4 unpacks the aggregate KS and JS numbers of Table III along four representative features of the macro-state vector. The top row corresponds to the adversarial scenario. In this case, the synthetic forecast reproduces the three inflection points of the real CDF and the correct quantile locations for all four features. The residual divergence is concentrated in the second and third quartiles (visible as the slight horizontal offset around the 0.5–0.7 Empirical Cumulative Distribution Function (ECDF) band) and is on the order of KS ≈ 0.15 per feature, exactly the irreducible drift between the two halves of the captured trace. The $\sigma_{IAT,k}$panel, whose support spans the narrowband of $2.2–3.2 \times 10^{-4}$ s, imposed by the per-bin packet budget of Section II, is reproduced to within $\sim 1.5 \times 10^{-5}$ of the empirical CDF, i.e., a fraction of the bin discretization step.

The bottom row of Fig. 4 depicts the nominal held-out window. Unlike the adversarial case, the empirical N_k ranges between 1 and ∼80 pkts/bin from the 5th to the 95th percentile, yet its support extends two orders of magnitude further into the tail, with isolated bursty bins reaching $\sim 5 \times 10^2$ pkts/bin. The synthetic forecast tracks the empirical reference on the lower three decades — which contain ∼90% of the probability mass — with a barely visible offset; it over-estimates the upper tail by roughly 5–8% of the ECDF in the top decade, a direct consequence of the calibration prefix containing slightly more high-burst minutes than the held-out suffix. The synthetic marginal is anchored to the calibration distribution by construction (Section II-B), so any calibration-vs-test drift in the upper tail propagates directly into the synthetic CDF; the magnitude of that offset matches the drift bound KS = 0.075 reported in Table III for the nominal scenario, confirming that the residual is data-bound, not pipeline-bound. The $\sigma_{\mathrm{IAT},k}$ panel is reproduced to within $\sim 10^{-3}$ s of the empirical CDF over the entire 0–0.3 s support, closing the loop on the temporal-scale fidelity of the spectral calibration. Both rows therefore confirm the same picture as the aggregate table: the residual gap visible in the figure is dominated by the irreducible drift of the captured trace (a property of the data, not of the pipeline).

Fig. 5 translates the distributional agreement back to the time domain. The empirical capture maintains a plateau mean of $\overline{N}_{\mathrm{real}} = 9{,}699$ pkts/bin with a standard deviation of $\sigma_{N,\mathrm{real}} = 121$ pkts/bin across the held-out window; the synthetic forecast reproduces both to within 0.02% on the mean ($\overline{N}_{\mathrm{synth}} = 9{,}700$ pkts/bin) and 11% on the standard deviation ($\sigma_{N,\mathrm{synth}} = 134$ pkts/bin), with the 5-bin moving averages of the two streams remaining within 30 pkts/bin of each other

TABLE V
MICROSCOPIC FIDELITY OF THE MOMENT-MATCHING RECONSTRUCTION VERSUS A POISSON SURROGATE, COMPUTED BIN BY BIN

	Estimator	Adv. (mean)	Nom. (P95)
CV^2 rel. err.	Moment-match	$1.0{\cdot}10^{-3}$	0.41
CV^2 rel. err.	Poisson	0.96	$\gg 1$
B abs. err.	Moment-match	$1.2{\cdot}10^{-4}$	0.092
B abs. err.	Poisson	0.683	0.301

over the entire 20 s held-out horizon. The two streams clearly do not coincide bin-by-bin, but they are not designed to: the phase perturbation $\eta[k]$ of eq. (1) introduces a controlled desynchronization that guarantees the output is a genuinely new realization rather than a replay of the calibration prefix. A synthetic output that tracked the empirical bin sequence inside the held-out window would, in fact, indicate that the calibration step had memorized the trace.

B. Microscopic Fidelity

The reconstruction algorithm of Section II-C maps a per-bin forecast $(N_k, \mu_{IAT,k}, \sigma_{IAT,k})$ back to a sequence of N_k inter-arrival times by solving the two-point mass mixture of Eqs. (6)–(7). To validate that this round-trip preserves the microscopic burstiness profile of the empirical capture (a property that prior synthetic-traffic generators do not explicitly evaluate [3], [4], [6]) we instantiate the algorithm, bin by bin, on the held-out windows and compare two scale-invariant statistics of the reconstructed packet sequence against the same statistics computed on the empirical packet sequence of the same bin: the squared coefficient of variation $CV^2 = (\sigma_{IAT}/\mu_{IAT})^2$ and the Goh–Barabási burstiness index $B = (\sigma_{IAT} - \mu_{IAT})/(\sigma_{IAT} + \mu_{IAT}) \in [-1, 1]$ ($B = 0$ for a Poisson process, $B \to 1$ for an extreme burst limit). To put the moment-matching algorithm in perspective, we contrast it with a Poisson surrogate that samples N_k inter-arrivals from $\mathrm{Exp}(1/\mu_{IAT,k})$, the standard benchmark in synthetic packet traffic literature.

Table V reports the results. In the adversarial regime, with a large per-bin packet count ($\overline{N} \approx 9{,}600$), the moment-matching reconstruction preserves CV^2 to within 0.1% of the empirical value and the burstiness index B to within 1.2×10^{-4}, three orders of magnitude below the trace's own burstiness ($\overline{B}_{\text{real}} = 0.68$). The Poisson surrogate yields a 96% relative error on CV^2 and an absolute error of 0.68 on B, i.e., it cannot reproduce the bursty geometry of the DNS flood.

In the nominal regime, where the per-bin packet count is two orders of magnitude smaller, the integer rounding of the spike count introduces a quantization error that increases the P95 relative error of CV^2 to 0.41; even in this demanding regime, the moment-matching reconstruction reduces the burstiness error by a factor of $\sim 3.3\times$ over the Poisson surrogate (0.029 vs. 0.301 in mean) and recovers the empirical burstiness $\overline{B}_{\text{real}} = 0.16$ to within one decimal digit in 95% of the bins.

Fig. 6 shows the same comparison per bin: the moment-matching points follow the identity line over the full empirical range of B in both regimes, while the Poisson surrogate yields an isotropic cloud near $B = 0$. With the macroscopic results

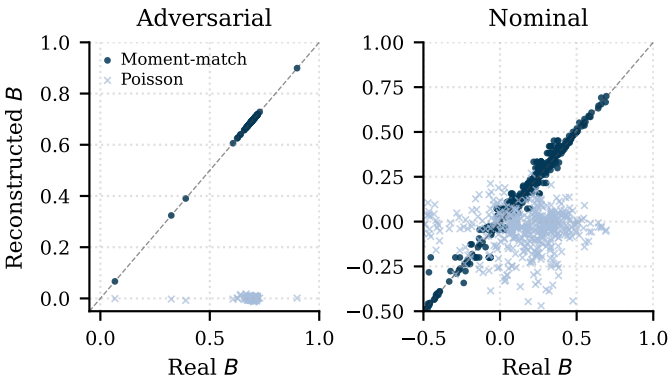


Fig. 6. Per-bin reconstructed burstiness B versus empirical B. Left: adversarial regime. Right: nominal regime.

of Section IV-A, this closes the fidelity loop from packets to state space and back under both distributional and moment tests, on a genuine out-of-sample forecast rather than a replay of the calibration context.

V. CONCLUSIONS

SAURON demonstrates that a zero-shot, calibration-only pipeline can synthesize 5G packet traces (spanning both adversarial and nominal regimes) at near-theoretical fidelity without the need for training. By composing a pre-trained autoregressive forecast with spectral diversity and exact quantile mapping, the framework closes the distributional gap to the held-out reference by up to 7.2× over an uncalibrated control, consistently outperforming state-of-the-art generative baselines.

Future work will include extending the microscopic reconstruction to handle correlated protocol counters, and dynamically adapting the calibration window to non-stationary captures of real captured traces.

REFERENCES

[1] A. F. Ansari *et al.*, "Chronos-2: From univariate to universal forecasting," 2025. [Online]. Available: https://arxiv.org/abs/2510.15821

[2] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[3] Z. Lin *et al.*, "Using gans for sharing networked time series data: Challenges, initial promise, and open questions," in *Proceedings of the ACM internet measurement conference*, 2020, pp. 464–483.

[4] Y. Yin *et al.*, "Practical gan-based synthetic ip header trace generation using netshare," in *Proceedings of the ACM SIGCOMM 2022 Conference*, 2022, pp. 458–472.

[5] N. Sivaroopan *et al.*, "Netdiffus: Network traffic generation by diffusion models through time-series imaging," *Computer Networks*, vol. 251, p. 110616, 2024.

[6] A. Chu *et al.*, "Netssm: Multi-flow and state-aware network trace generation using state-space models," *Proceedings of the ACM on Networking*, vol. 4, no. CoNEXT1, pp. 1–24, 2026.

[7] A. T. Kalai *et al.*, "Evaluating large language models for accuracy incentivizes hallucinations," *Nature*, pp. 1–3, 2026.

[8] E. G. de la Calera Molina *et al.*, "Automating 5g traffic generation with virtual ues: A scalable network testing infrastructure," *IEEE Transactions on Network and Service Management*, vol. 23, pp. 299–316, 2025.

[9] Internet Systems Consortium, *BIND 9*, Internet Systems Consortium, 2024, version 9.20.x. [Online]. Available: https://www.isc.org/bind/

[10] J. Yoon *et al.*, "Time-series generative adversarial networks," *Advances in neural information processing systems*, vol. 32, 2019.

[11] M. Arjovsky *et al.*, "Wasserstein gan," 2017. [Online]. Available: https://arxiv.org/abs/1701.07875

[12] H. Ishfaq *et al.*, "Tvae: Triplet-based variational autoencoder using metric learning," 2023. [Online]. Available: https://arxiv.org/abs/1802.04403