

Data Can Speak for Itself: Quality-guided Utilization of Wireless Synthetic Data

Chen Gong^ℙ, Bo Liang^ℙ, Wei Gao^ℤ, Chenren Xu^{ℙℕ✉*}

^ℙSchool of Computer Science, Peking University; ^ℤUniversity of Pittsburgh

^ℕKey Laboratory of High Confidence Software Technologies, Ministry of Education (PKU)

Abstract

Generative models have gained significant attention for their ability to produce realistic synthetic data that supplements the *quantity* of real-world datasets. While recent studies show performance improvements in wireless sensing tasks by incorporating *all* synthetic data into training sets, the *quality* of synthetic data remains unpredictable and the resulting performance gains are not guaranteed. To address this gap, we propose tractable and generalizable metrics to quantify quality attributes of synthetic data—affinity and diversity. Our assessment reveals prevalent affinity limitation in current wireless synthetic data, leading to mislabeled data and degraded task performance. We attribute the quality limitation to generative models’ lack of awareness of untrained conditions and domain-specific processing. To mitigate these issues, we introduce SYNCHECK, a quality-guided synthetic data utilization scheme that refines synthetic data quality during task model training. Our evaluation demonstrates that SYNCHECK consistently outperforms quality-oblivious utilization of synthetic data, and achieves 4.3% performance improvement even when the previous utilization degrades performance by 13.4%.

CCS Concepts

• Computing methodologies → Artificial intelligence.

Keywords

Generative Model, Data Quality, Wireless Sensing

ACM Reference Format:

Chen Gong, Bo Liang, Wei Gao, Chenren Xu. 2025. Data Can Speak for Itself: Quality-guided Utilization of Wireless Synthetic Data. In *The 23rd Annual International Conference on Mobile Systems, Applications and Services (MobiSys ’25)*, June 23–27, 2025, Anaheim, CA, USA. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3711875.3729123>

1 Introduction

Recent technical advances in generative AI models have shed light on the possibility to generate realistic synthetic data [1–3]. One important variant is the conditional generative models that can

guide the generative process with input class labels and produce output from different specified types [4], such as the generation of handwritten words based on text condition to facilitate tasks like word recognition [5]. Following this trend, the wireless research community, long facing data scarcity, has adopted various conditional generative models [6–10] to augment the availability of training data with synthetic samples produced with task-specific conditions, such as human action and antenna location in part (a) of Fig. 1. The *labeled synthetic data* has proved effective to improve the performance of data-hungry wireless applications such as action recognition [11] and indoor localization [12].

While most work in data synthesis for data scarcity alleviation approves its success in supplementing *quantity* [13–15], prior studies have highlighted a remaining deficiency of synthetic data *quality* in capturing domain-specific knowledge, such as that required in medical applications [16, 17]. To bridge this gap, human expert feedback is frequently employed for calibration [18]. However, wireless signals are not understandable by humans, making human-driven quality evaluation impractical. Given that the nonselective utilization of all synthetic data has been shown to yield suboptimal task performance [19], we ask a question in this paper: *How to evaluate the quality of wireless synthetic data and how can the quality assessment guide the utilization of synthetic data?*

Answering this question is not easy, as there are several technique challenges of conducting and leveraging the synthetic data quality assessment in the wireless domain. The first challenge is to **define tractable and generalizable quality quantification metrics (C1)**. In established synthetic data quality assessment [19, 20], affinity and diversity are recognized key attributes: *Affinity captures the realism of synthetic data in its similarity to real data, while diversity emphasizes the data variability to extend the original distribution of real data*. Current quantification metrics for the two attributes often rely on empirical comparisons of synthetic data generated from the same training set [21, 22], which is problematic for wireless generative models tailored to specific hardware and signal properties [10, 23] with customized datasets, making it difficult to perform unified comparisons on common datasets. Therefore, wireless synthetic data quality assessment requires generalizable metrics, with tractable theoretical support, to enable fair cross-dataset comparisons.

The second challenge is to evaluate **heterogeneous data generation techniques (C2)**. Wireless researchers have developed advanced generation techniques across diverse applications [24–26]. Given the breadth of research, systematic evaluation within a reasonable computational budget is essential to draw reliable and universal conclusions about the quality of wireless synthetic data.

To address the challenges of tractable and generalizable quality assessment (C1), we propose quality quantification metrics

*✉: chenren@pku.edu.cn

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MobiSys ’25, Anaheim, CA, USA

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1453-5/25/06
<https://doi.org/10.1145/3711875.3729123>

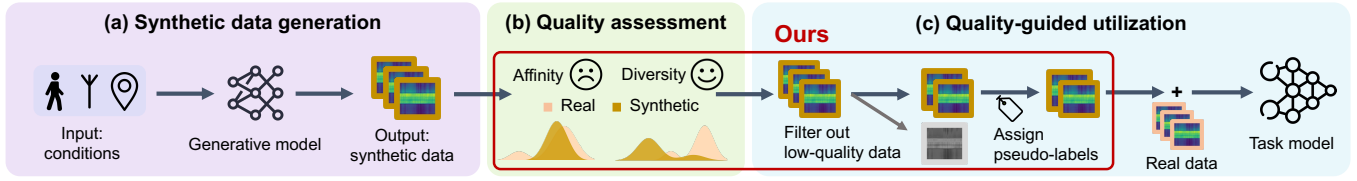


Figure 1: Our quality assessment and quality-guided utilization of wireless synthetic data. Generative models produce synthetic data from conditions to supplement data *quantity* for wireless applications. Compared to previous quality-oblivious utilization using *all* synthetic data with *conditions as labels*, we assess synthetic data quality, reveal its affinity limitation, and propose a quality-guided utilization scheme incorporating *filtered* synthetic samples with *assigned pseudo-labels* for better data quality and task performance.

with both *theory justification and cross-dataset comparability*. Our approach incorporates Bayesian analysis and performance indicators that can be consistently referenced across datasets, effectively overcoming the limitations of previous task-oriented performance metrics [27] in theoretical support and generalization. Specifically, we interpret affinity and diversity as conditional probability distribution between samples and labels, and establish the connection between these attributes and task performance across different training sets. Additionally, we propose to employ the overall classification confidence distribution as performance indicator [28, 29], where training set serves as reference for each dataset. This approach enables tractable and generalizable quality assessment for each wireless generative model.

To efficiently account for the diverse generation techniques of wireless synthetic data (C2), We categorize the synthetic data according to its source and target domains of generation, and conduct a *rigorous and systematic quality quantification of synthetic data* from representative generative models with our metrics. Our quality assessment reveals that current wireless synthetic data suffer from prevalent *limited affinity*, yet possess *good diversity* for classes in the training set, as shown by the part (b) of Fig. 1. Consequently, prior quality-oblivious utilization of synthetic data with limited affinity results in training data with incorrect labels. The mislabeled data can distort training process and degrade performance [30, 31], and we observe performance degradation when trained with synthetic data deficient in both affinity and diversity in §7.4.

Having identified the quality deficiencies, we propose a **quality-guided utilization scheme** designed to *enhance synthetic data affinity while preserving its diversity*. This solution addresses a critical gap in existing work within the wireless domain, which has focused on advancing data generation techniques but overlooks effective utilization of synthetic data for downstream tasks [9, 32]. Our key rationale is to treat synthetic data as unlabeled and real data as labeled, combining both data sources with a *semi-supervised learning* framework. Specifically, our approach, SYNCHECK, filters out synthetic samples with low affinity and assigns pseudo-labels to remaining samples during iterative model training, as illustrated in part (c) of Fig. 1. This utilization scheme is general post-processing step without modifying the training or inference of generative models, making it adaptable to various generation processes.

We implement SYNCHECK and conduct extensive experiments across various categories of wireless synthetic data. Our results show that SYNCHECK improves task performance by 12.9% to 4.3%, while performance gain with quality-oblivious utilization of synthetic data is 4.1%. Even in worst-case scenarios, SYNCHECK yields

a 4.3% improvement, while previous utilization degrades performance by 13.4%. Furthermore, synthetic data filtered by SYNCHECK exhibits enhanced affinity and comparable diversity to unfiltered raw data. The alignment between task performance and our quality metrics validates the effectiveness of our quality metrics and the necessity of quality-guided utilization of synthetic data¹.

Our contributions in this paper are summarized as follows:

- We introduce the first tractable and generalizable metrics for quantifying affinity and diversity in wireless synthetic data. Our data quality evaluation of representative generative models reveals limited affinity but good diversity.
- We introduce SYNCHECK, to the best of our knowledge, the first quality-guided scheme for wireless synthetic data utilization, improving task performance with better data affinity.
- We conduct extensive experiments and show that SYNCHECK boosts task performance compared to nonselective utilization of real and synthetic data.

2 Background

2.1 Wireless Data Generation

The scarcity of high-quality large-scale training datasets has long been a challenge in wireless community [35, 36]. To address this limitation and alleviate the burden of data collection, data generation techniques have been developed to expand quantity through augmentation or synthesis [37].

Data Augmentation. Augmentation techniques extend existing data by modifying the outer representations while maintaining the inner semantics, informed by domain knowledge. It is widely used across various domains and proves effective in boosting performance. In the wireless domain, data augmentation models the variation of channels and transceivers. Examples include multi-scale frequency resolution and motion-aware random erasing and shifting [38], random phase shifts and amplitude fluctuations [39], and random circular shifts and re-scaling [40].

Data Synthesis. Synthesis techniques generate new data either by adhering to physical laws [10, 41, 42] or through deep learning models [43]. The former approach typically involves environment modeling and simulation based on physical rules to replicate real data, while the latter learns the data distribution of the training set and synthesizes new data that follows this underlying distribution. Given the complexity and over-idealization of accurate physical modeling, which is often limited to a single scenario, this study

¹Codes are available at <https://github.com/MobiSys25AE/SynCheck>.

Table 1: Comparison of existing quantification metrics on synthetic data quality.

Literature	Quality Metrics		Quantification Rationale	Properties	
	Affinity	Diversity		Tractable	Generalizable
QAGAN [33]	SSIM	FID	Visual similarity	✗	✓
GoodGAN [27]	TRTS accuracy	TSTR accuracy	Task accuracy	✗	✗
CAS [34]	(missing)	TSTR w. $p_\theta(y x)$	Bayesian analysis	✓	✗
Faithful [19]	α -precision	β -recall	Hypersphere distribution	✗	✓
Ours	TR margin w. $p_\theta(x y)$	TS margin w. $p_\theta(y x)$	Task confidence w. Bayesian analysis	✓	✓

focuses on *data-driven generative models* and *aims to advance our understanding of the quality of wireless synthetic data and enhance its utilization to improve task performance*.

2.2 Synthetic Data Quality Assessment

The quality of synthetic data has been extensively studied in generative model research [44, 45]. However, existing metrics fail to simultaneously achieve tractability for theoretical grounding and generalizability for task- and dataset-agnostic evaluation, as they either rely on visual heuristics or are designed for controlled tests on the same dataset. To better understand these limitations, we summarize relevant work on affinity and diversity metrics in Tab. 1.

Deficiencies of Existing Metric Designs. *Visual similarity* metrics are widely adopted for synthetic image evaluation, including per-sample structural similarity (SSIM) [46] and overall distribution diversity (FID) [47]. However, these similarity-based metrics rely on heuristic assumptions and fail to align with human perception [21, 22], limiting their theoretical tractability. *Task-oriented* metrics evaluate real-world task performance in *train-real-test-synthetic* (TRTS) and *train-synthetic-test-real* (TSTR) setups [27]. The former trains a model on real data and tests it on synthetic data, and captures synthetic data affinity; the latter trains the model on synthetic data and tests it on real data, and reflects diversity. While subsequent work has analyzed TSTR from a Bayesian perspective [34], theoretical support for measuring affinity remains underdeveloped. Overall, these empirical metrics require comparisons within the same dataset, limiting their generalizability. Another line of research assumes a *hypersphere distribution* of data [19] and calculates precision and recall of synthetic data relative to different support sets determined by the sphere radius. However, this hypersphere assumption and its dependence on a predefined distance function [48] restrict its tractability.

Challenges in the Wireless Domain. Unlike image data, wireless data lacks semantic interpretability, precluding human visual validation. This necessitates theoretically grounded quality metrics to ensure reliability. Furthermore, wireless data exhibits diverse structural complexity, including spatial dependencies and dynamic channel variations. Since most wireless generative models are optimized for task- or dataset-specific requirements, fair quality assessment demands references for each model, in order to mitigate biases of data and task dependencies and maintain generalizability.

3 Our Quality Metrics

To address the limitations of existing evaluation metrics, we propose quality quantification metrics that balance tractability and

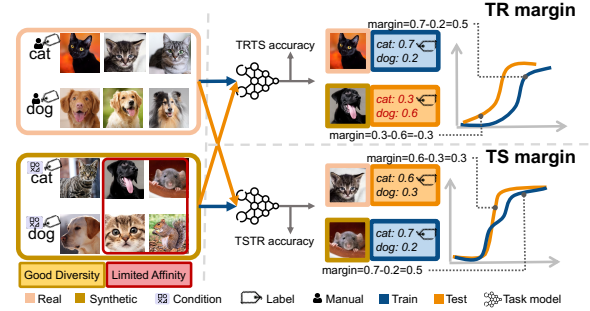


Figure 2: Our tractable and generalizable metrics: TR margin and TS margin. Since wireless signals cannot be easily classified by humans, we use animal pictures to illustrate the basic ideas of confidence distribution and synthetic label errors.

generalizability. Specifically, our metrics inherit the idea of task-oriented evaluation with different train set choices, and complete Bayesian analysis [34] to guarantee tractability. Furthermore, we repurpose the model decision confidence concept, known as *margin* [30], to assess the quality of synthetic data, where the margin of the training set serves as a readily available reference. The quantification process is illustrated in Fig. 2. In the following subsections, we provide a detailed explanation of the Bayesian framework and margin definition, along with empirical evidence demonstrating the advantages of our margin-based metrics.

3.1 Metric Definition

To establish the Bayesian theoretical foundation, we begin by formulating the two quality attributes, affinity and diversity, with conditional distributions between samples x and their corresponding labels y , where the label of synthetic data corresponds to its generation condition. We use y to represent both labels and generation conditions interchangeably. Given the real data distribution $p(x, y)$, a generative model approximates this distribution using its parameters θ , resulting in a synthetic data distribution $p_\theta(x, y)$.

Definition 1 (Affinity). Affinity measures similarity between synthetic samples and real samples, denoted by consistency between $p_\theta(x|y)$ and $p(x|y)$.

Definition 2 (Diversity). Diversity measures the extent that synthetic data covers the range of real data, denoted by consistency of $p_\theta(y|x)$ and $p(y|x)$.

Definition Explanation. Part (c) of Fig. 1 illustrates the synthetic data distribution in terms of affinity and diversity. For high-affinity synthetic samples within one specific generation condition y_0 , they

resembles real samples of the same class y_0 . Therefore, synthetic and real samples of class y_0 share similar distribution. In this way, affinity can be captured by similarity between $p_\theta(x|y)$ and $p(x|y)$. For high-diversity synthetic data, it should cover the range of real data across various generation conditions while maintaining the inter-class boundary. Such value coverage can be described as correlation between $p_\theta(y|x)$ and $p(y|x)$. Based on the definition above, we further establish the connection between quality attributes task performance with different training set.

Theorem 1 (Affinity from train-real setup). When a task model is trained on real data, its test performance on synthetic data reflects affinity of synthetic data. High synthetic data affinity is a sufficient condition for good test performance in the train-real setup.

Theorem 2 (Diversity from train-synthetic setup). When a task model is trained on synthetic data, its test performance on real data reflects the diversity of the synthetic data. High synthetic data diversity is a sufficient condition for achieving strong test performance in the train-synthetic setup².

A potential concerns emerges when synthetic data deviates from the real label space. While this issue is precluded by the affinity requirement of $p_\theta(x|y) = p(x|y)$, it remains unresolved by diversity requirement of $p_\theta(y|x) = p(y|x)$. By Bayes' theorem [49], with known condition distribution $p(y)$, the diversity requirement implies $\frac{p_\theta(x|y)}{p_\theta(x)} = \frac{p(x|y)}{p(x)}$, preserving the relative likelihoods of data conditioned on labels. As the ratio remains invariant despite the presence of label-mismatched samples, such synthetic data will maintain high test accuracy in the train-synthetic setup [34]³. Thus, conditional distribution alignment for diversity remains theoretically sound, and synthetic data remains effective for improving model performance provided that outlier samples are effectively identified and managed.

Having established the Bayesian foundation for using task performance as quality quantification metrics to ensure tractability, we further investigate task performance indicators to facilitate generalizable quality assessment. Previous metrics, such as TSTR and TRTS accuracies [27], can only compare synthetic data quality across multiple generative models on the same dataset, as the single-valued metric fails to indicate the absolute levels of affinity or diversity *without a standard reference*. However, wireless generative models are tailored to specific hardware and signal properties, making unified comparisons across datasets challenging. To address this issue, we adopt a confidence distribution concept called margin to enable fair comparisons across datasets. The formal definition of margin along its generalization with standard reference is:

Definition 3 (Margin). For a classification model, margin is the classification confidence difference between the label and the largest among other classes:

$$\text{Margin}(x, y) = z_y(x) - \max_{i \neq y} z_i(x) \quad (1)$$

where $z(x)$ is the classification confidence of sample x . The margin calculation is demonstrated with concrete numerical and human-comprehensible examples in Fig. 2.

²Proofs of the two theorems are included in Appendix §A

³A generative model blending noise (with probability p) and real-distribution data (with probability $1 - p$) still enables classifiers to generalize effectively to real data.

The margin distribution has been shown predictive of model generalization [28, 29]. For a well-generalized model, the margin distribution on test set closely resembles that of training set. Thus, the margin of training set serves as a *natural standard reference*, enabling generalizable quality quantification: If the margin distribution of test set composed of synthetic data deviates significantly from that of real data training set, it indicates poor affinity of synthetic data. The same principle applies to the diversity property in train-synthetic setup.

Conclusion 1 (Affinity with TR margin). The TR margin of test set, composed of synthetic data, reflects synthetic data affinity.

Conclusion 2 (Diversity with TS margin). The TS margin of test set, composed of real data, reflects synthetic data diversity.

3.2 Metric Advantage

Given the fine granularity and natural reference points inherent in our quality metrics, we empirically investigate their correlation with task performance. To assess this relationship, we conduct controlled experiments by systematically varying the affinity and diversity of synthetic data, then evaluate model performance trained on real data and artificially modified synthetic data.

To examine affinity, we perturb synthetic data with Gaussian noise at varying standard deviations from 0.05 to 0.3 in increments of 0.05. The results reveal a strong positive correlation between the TR-margin metric and task performance (Pearson $r = 0.900$, $p = 0.0146$), outperforming the TRTS accuracy metric ($r = 0.864$, $p = 0.0265$). This suggests that margin-based metrics exhibits stronger correlation due to its fine-grained characterization of model behavior. For diversity analysis, we vary the quantity of synthetic training data from 10% to 100% of the real dataset size in 10% increments. Here, the TS-margin metric maintains a robust correlation with task performance ($r = 0.864$, $p = 0.0013$), while TSTR accuracy shows a weaker ($r = 0.540$) and statistically nonsignificant ($p = 0.1147$) relationship. These results demonstrate that margin-based metrics are more sensitive to variations in both affinity and diversity, making them more reliable for predicting task performance.

4 Quality Assessment of Wireless Synthetic Data

To systematically assess the quality of wireless synthetic data, we conduct a comprehensive literature review of wireless generative models, aligning them to corresponding techniques in the AI research and categorizing them into two distinct groups. We reimplement representative works from each category using a unified experimental framework to ensure a fair comparison, and evaluate the affinity and diversity of synthetic data from each model.

4.1 Wireless Generative Model Taxonomy

We review and categorize recent publications that augment training sets with synthetic data in Tab. 2. They are categorized based on source and target data domains for synthesis in Fig. 3:

Cross-domain Sample Transfer. This category uses data from one domain as input and produce synthetic data in a different domain. Here, a 'domain' refers to application-specific aspects such as individuals, environments, or device configurations. The goal is

Table 2: A taxonomy of existing data-driven wireless generative models is provided, with representative works highlighted for further assessment in affinity and diversity attributes in this paper.

Literature	Approach	Technique Category	Availability
CsiGAN [6]	Synthetic data cross persons	Cross-domain transfer [50]	Open-source code and data
DFAR [51], LearnSense [15]	Synthetic data cross scenarios		Close-source
SCTRFL [52], AutoFi [53]	Synthetic data cross time		Close-source
RF-Diffusion [9]	Training set enrichment with synthetic samples	In-domain amplification [54]	Open-source code
RF-EATS [43], AF-DCGAN [55]			Close-source

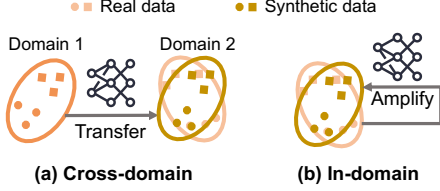


Figure 3: Categories of generative model techniques.

to transfer real samples from source domain into synthetic samples in the target domain and complement the target domain data.

In-domain Sample Amplification. This category focuses on generating additional samples within the same domain. While similar to data augmentation, this approach replaces heuristic rules with data-driven generative models to create new synthetic samples and enhance dataset diversity.

4.2 Quality Quantification and Analysis

Based on our quantification metrics and categories of existing wireless generative models, we conduct a series of experiments to answer the following research question:

RQ1: In terms of affinity and diversity, how is the quality of existing wireless synthetic data?

Representative Generative Models. We aim to benchmark all existing wireless generative models shown in Fig. 3. However, the accessibility of both models and datasets varies significantly, which influenced our selection process. To ensure a fair and unbiased evaluation, we prioritize recent models that provide open-source code and publicly available dataset partitions, as this helps mitigate implementation biases and enhances the reliability of our results.

For cross-domain transfer, we select CsiGAN [6], which transfers gesture samples from trained users to unseen users. The training data consists of only a subset of gesture classes, while the testing data includes all classes. This setup, with incomplete class coverage during training, reflects real-world deployment scenarios where samples from new users are often limited. For in-domain amplification, we select RF-Diffusion [9], a diffusion model conditioned on human orientation, activity, and antenna configuration to generate synthetic samples. It utilizes all available activity classes for training to ensure comprehensive coverage of the conditions.

Task model and dataset. We employ a simple yet effective ResNet34 architecture [56] as the task model, which is proven effective across various wireless tasks and benchmarks [57]. The selected generative models are tested on the datasets used in their original studies:

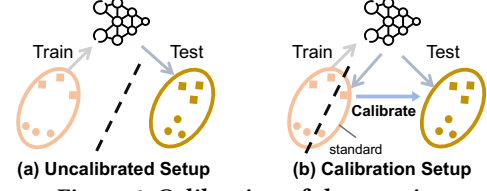


Figure 4: Calibration of the margin.

the SignFi dataset [58] for CsiGAN[6], and the Widar dataset [59] for RF-Diffusion [9].

Margin Calibration. Our metrics employ margin comparisons between the test set and the train set to enable generalizable quality quantification across datasets, where the margin of train set serves as a reference for each dataset. However, inherent performance discrepancies between the train and test sets [60, 61] can disrupt the margin comparison. To address this issue, we introduce a calibration procedure shown in Fig. 4. Specifically, in both the TR and TS setups, we pre-select a subset of the corresponding training data and designate it as the standard test set, ensuring it follows the same distribution as the train set. The calibration process involves calculating the mean margin values for both the train and standard test sets, and then adjusting the margins of the actual test set based on the difference between these two mean margin values. The data quality is quantified by measuring the distribution gap between the training set and calibrated test set. One suitable measure for this gap is Jensen-Shannon (JS) Divergence [62], which assesses the discrepancy between two distributions, and a smaller value indicates greater similarity. The results of quality assessment with JS Divergence, are presented in Tab. 3.

Necessity of Calibration. To validate the importance of calibration, we evaluate TR-margin and TS-margin across different model architectures (ResNet18, ResNet34) and training epochs, involving model accuracy variations exceeding 20%. The analysis reveals that uncalibrated margin distribution gap show 13.5% variations and calibration achieves better consistency of 5.9% variation. These findings demonstrate that calibration mitigates model dependency to some extent, enabling more stable and reliable quality assessment.

Data Quality Results. The synthetic data quality results are shown in Fig. 5: A stronger correlation between the margins of the train set and test set in the train-real and train-synthetic setups indicates better affinity and diversity, respectively. Based on these quality quantification results, we highlight the following key observations:

- **Prevalent Affinity Limitation.** In both cross-domain (Fig. 5a) and in-domain (Fig. 5c) generation, the calibrated test margins for synthetic data show significant discrepancies from the train

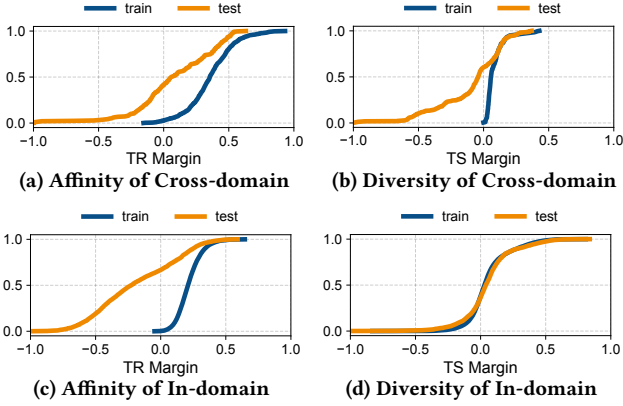


Figure 5: Quality quantification of synthetic data.

margin. This margin difference indicates that the synthetic data is *affinity-deficient* and *misabeled*, which can mislead the task model’s training.

- **Imbalanced Diversity in Cross-domain Generation.** In the TS margin (Fig. 5b) for cross-domain generation, the calibrated test margin closely aligns with the train margin for the top 40% of synthetic samples, but diverges significantly in the remaining 60%. This discrepancy indicates inadequate and imbalanced diversity in the cross-domain synthetic data.
- **Good Diversity in In-domain Generation.** The TS test margin (Fig. 5d) for in-domain generation closely matches the train margin, indicating good diversity of in-domain synthetic data.
- **Abnormal Train Margins.** An additional observation in the train-synthetic setup is the abnormal distribution of train margins for in-domain generation. In other setups, train margins are predominantly greater than zero, which is reasonable for a model well-behaved on the training set. However, in Fig. 5d approximately 20% of the train margins are negative. This abnormal distribution suggests label confusion in the in-domain synthetic data, which disrupts the model’s training process.

Answer to RQ1: Our quantification results reveal that existing wireless synthetic data exhibits prevalent limited affinity, with imbalanced diversity for cross-domain generation and good diversity for in-domain generation.

4.3 Investigation of Underlying Causes

We investigate the underlying causes of the prevalent low affinity and partially imbalanced diversity in wireless synthetic data and answer the following research question:

RQ2: What are the causes of low affinity and imbalanced diversity in the wireless synthetic data?

Analysis Method. We analyze the potential causes of synthetic data quality issues throughout the entire generation process, including pre-generation preparation and domain-specific post-generation processing. Based on guidelines above, we identify limitations of the representative generative models, which further inspire the development of our quality-aware synthetic data utilization scheme.

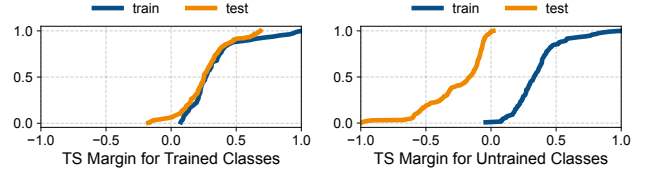


Figure 6: Diversity comparison between trained and untrained classes of generative models.

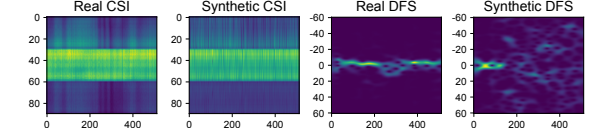


Figure 7: Data with Similar CSIs but Dissimilar DFSs.

Incapability of Condition Transfer. The pre-generation phase prepares training data for generative models, where models designed for cross-domain transfer often encounter incomplete class coverage in the training set [6, 51, 63]. This issue arises from the specific constraints of wireless applications, which aim to augment the target domain with limited available data. To further investigate the effect of class coverage, we compare the affinity and diversity between trained and untrained classes of generative models in Fig. 6. The TS margins for the train and test sets of the trained classes are well aligned, while the margins for the untrained classes are more divergent. This margin distribution difference shows a significant diversity advantage of trained classes over untrained ones. The diversity gap between trained and untrained classes underscores the challenges of effective cross-domain condition transfer.

Unawareness of Domain-specific Processing. The post-generation phase transforms synthetic data into application-specific formats. A typical example is WiFi signal processing, where generative models synthesize raw channel state information (CSI) [9, 64], and applications like human action recognition rely on higher-level features such as Doppler frequency shift (DFS) [38] and Body Velocity Profile (BVP) [59]. The discrepancy in signal formats arises from the differing objectives of generative models and applications: CSI synthesis captures diverse environmental factors and is effective for complementing real data in complex scenarios, whereas DFS and BVP are designed to be resilient to such variations and focus on target tasks. This divergence between the model’s raw output (CSI) and application’s input (DFS) is overlooked in prior work. Our analysis of RF-Diffusion highlights this issue and reveals that generative models trained on CSI fail to generate DFS similar to the real DFS, as shown in Fig. 7. In summary, generative models that do not account for domain-specific processing fail to align their outputs with the particular semantic properties required by target applications.

Insights for Better Utilization. The entire generation process of wireless synthetic data is inherently tied to domain-specific properties. While we have identified quality issues at both pre-generation and post-generation stages, the variety of generation techniques and applications makes this investigation far from exhaustive. Given the difficulty of comprehensive quality assurance throughout data generation, an effective utilization scheme for synthetic data should incorporate domain-specific knowledge, and overcome the quality limitations in a *task-oriented* and *generic* manner. To this end, both our quality metrics and utilization approach

leverage task models to analyze synthetic data. Since affinity deficiency directly impacts label correctness and compromises training data, and diversity imbalance results in uneven performance gains across classes, our first step in developing a better utilization scheme addresses the affinity limitation.

Answer to RQ2: Generative models agnostic to untrained condition distributions and domain-specific processing produce low-quality data. An effective utilization scheme for synthetic data should recognize the quality issues and underlying causes.

5 Quality-aware Utilization of Synthetic Data

Building on the quality quantification of affinity and diversity in synthetic data (RQ1), and exploring the causes of quality issues (RQ2), we conclude that the primary limitation of current wireless synthetic data is prevalent affinity deficiency, which leads to mislabeled synthetic data and degraded task performance. Keeping these observations in mind, we address the following research question:

RQ3: Given synthetic data with limited affinity, how can it be utilized to improve task performance in a generic way?

5.1 Approach Overview

We propose a quality-guided utilization scheme for synthetic data in a post-processing framework, adaptable to various generative models. This approach stems from the cause investigation that data quality is deeply intertwined with the entire generation process, making it challenging to design targeted solutions to cover every aspect of quality issues.

In contrast to previous methods that rely on condition-labeled synthetic data for fully supervised training [6, 9], our solution, SYNCHECK, eliminates the need for generation conditions as labels. Instead, it adopts a semi-supervised learning paradigm, leveraging labeled real data and unlabeled synthetic data. The model simultaneously performs the target task while assigning pseudo-labels to the synthetic samples. The pseudo-labels provide additional supervision with domain-specific processing to iteratively refine the model. Furthermore, since the actual label of a synthetic sample may not correspond to any real sample, we adopt an *open-set* setup [65, 66], as opposed to the *closed-set* setup. The closed-set setup assumes that the synthetic data classes are identical to those of the real data. In contrast, the open-set setup assumes that synthetic data may contain classes not present in real data. The samples falling outside the real data classes is identified as irrelevant and filtered out.

SYNCHECK Model Architecture. The SYNCHECK model comprises three main components: backbone, task classifier and inlier-outlier detector, as shown in Fig. 8. Specifically, the backbone model serves as a feature extractor for both classification and inlier-outlier detection. The classifier, denoted as f , is responsible for the target classification task and its output dimension is the number of classes C in real data. The detector decides whether the input sample is an inlier or outlier of a specific class. With the detector g^j for class j , the probability of a sample x being an inlier is $g^j(t = 0|x_b)$, while the probability of x being an outlier is $g^j(t = 1|x_b)$. This *one-vs-all* detection [67] has proven effective in identifying per-class inliers.

SYNCHECK Workflow. The semi-supervised learning of SYNCHECK comprises two phases: model warm-up with labeled real data and unlabeled synthetic data, and model iteration with labeled real data and pseudo-labeled synthetic data. The training losses for each phase are shown in Fig. 8. In the first phase, the task classification and inlier-outlier detection models are trained in a supervised manner, followed by an unsupervised enhancement to further improve the robustness of both classification and detection. With a well-behaved model on classification and detection, the second phase introduces additional training losses by assigning pseudo-labels to the filtered synthetic data. These pseudo-labels, combined with application-specific processing, provide additional supervision that incorporates domain knowledge to guide model training.

5.2 Phase 1: Model Warm-up

The first phase focuses on training a model to achieve good performance in both task classification and inlier-outlier detection, preparing it for effective utilization of synthetic data in the subsequent phase. The model warm-up phase involves two key components: supervised training using labeled real data and self-supervised training with unlabeled synthetic data.

Supervision with Labeled Real Data. Using manually collected real data x with reliable labels y , we train the target task classifier f following common practice. Besides, we enforce the labeled real samples to be inliers of their label classes, and outliers of all other classes, as determined by the one-vs-all detectors g^j s. The supervised training losses with labeled real data are as follows:

$$\begin{aligned} L_s &= L_{\text{task}} + L_{\text{ova}} \\ L_{\text{task}} &= \text{cross entropy}(f(x), y) \\ L_{\text{ova}} &= -\log(g^y(t = 1|x) - \min_{j \neq y} \log(g^j(t = 0|x))) \end{aligned} \quad (2)$$

Self-supervision with Unlabeled Synthetic Data. With the low-affinity synthetic data u without usable labels, we employ consistency regularization to ensure robust model predictions and entropy minimization to enhance the distinction between inliers and outliers. The rationale behind consistency regularization is based on the idea that for an input sample with small perturbations, the model's prediction should remain similar to that of the original unperturbed sample [68], as long as the noise does not interfere with the task-specific semantics. For the wireless signals, we use Gaussian noise, a widely recognized model for propagation channel noise [69]. Furthermore, to encourage clearer discrimination between class inliers and outliers, we minimize the entropy of the outputs from the one-vs-all detectors g^j to avoid ambiguous inlier-outlier decision. The training losses for enhancing consistency and distinction with unlabeled synthetic data include:

$$\begin{aligned} L_u &= L_{\text{cons}} + L_{\text{ent}} \\ L_{\text{cons}} &= \sum_j \|g^j(u) - g^j(u + \text{noise})\|_2^2 + \|f(u) - f(u + \text{noise})\| \\ L_{\text{ent}} &= \sum_j \sum_{t=0}^1 -g^j(t|u) \log(g^j(t|u)) \end{aligned} \quad (3)$$

5.3 Phase 2: Model Iteration

After the warm-up phase, the model is now capable of predicting input classes and distinguishing between inliers and outliers. In this phase, we use the model to filter out synthetic data with low affinity

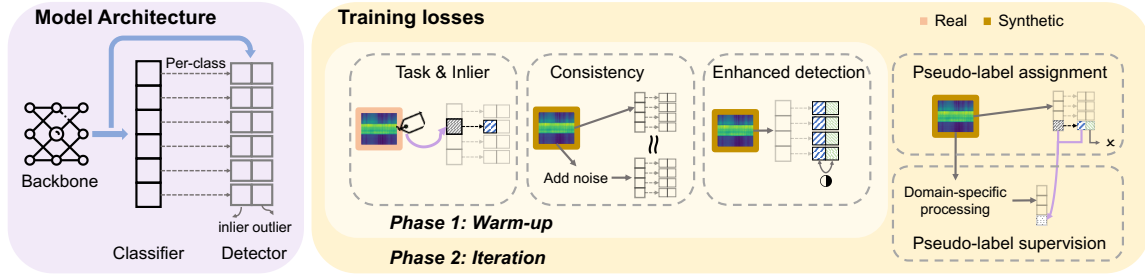


Figure 8: SYNCHECK Workflow.

(outliers) and assign pseudo-labels to the remaining high-affinity synthetic data (inliers). The pseudo-label supervision iteratively improves the model’s performance in a domain-dependent manner.

Pseudo-label Assignment. For each synthetic sample u , we first determine its most likely class \hat{y} using the task classifier f . We then assess its affinity by passing it through the inlier-outlier detector $g^{\hat{y}}$. If the sample is classified as an inlier of class \hat{y} , it is assigned the pseudo-label \hat{y} and can provide further supervision for the task.

Pseudo-label Supervision with Domain Knowledge. So far a subset of synthetic samples (denoted as S) has been identified as inliers and assigned pseudo-labels. To further improve the model, we employ a pseudo-label utilization strategy in which the model’s predictions for augmented samples should align with the assigned pseudo-labels [70]. What sets SYNCHECK apart is the incorporation of wireless-specific augmentations [38–40], such as varying time window lengths and random subcarrier distortion, to integrate domain-specific knowledge into the label prediction process. The training loss for inlier synthetic samples $s_b \in S$ with application-specific augmentation \mathcal{A}_s is

$$L_{\text{pseu}} = \text{cross entropy}(\mathcal{A}_s(s_b), \hat{y}) \quad (4)$$

Answer to RQ3: We propose a semi-supervised learning framework that treats synthetic data as unlabeled and selects high-affinity synthetic samples to aid model training.

6 Implementation

We reproduce existing generative models and implement our semi-supervised learning with PyTorch [71]. For CsiGAN, we attempt to align our results with the original paper, as the authors have publicly released their dataset processing scripts [72]. For RF-Diffusion, since the dataset processing code is not open-sourced, we use the public dataset partition from RFBoost [73] for the Widar dataset [59]. The model backbone is primarily ResNet34 [56], while we also include microbenchmarks to investigate the effect of backbones.

Model training and inference were performed on a workstation equipped with a Genuine Intel (R) CPU and six NVIDIA 3090 GPUs. For the evaluation of CsiGAN, the models were trained for 100 epochs with a learning rate of 2×10^{-4} , a batch size of 16, and a StepLR scheduler configured to decay the learning rate by a factor of 0.1 every 40 epochs. For the evaluation of RF-Diffusion, the models were trained for 10 epochs with a learning rate of 1×10^{-3} , a batch size of 64, and a StepLR scheduler set to decay the learning rate by a factor of 0.9 every 2 epochs. These configurations were based on the size and convergence characteristics of each dataset.

7 Evaluation

7.1 Experiment Setup

7.1.1 Dataset and Generative Models. Our empirical analysis focuses on two representative generative models: CsiGAN for cross-domain transfer and RF-Diffusion for in-domain amplification. Unless otherwise specified, the initial amount of synthetic data is the same as real data of the original dataset in our experiments.

Cross-domain Transfer. CsiGAN is trained on the SignFi dataset [58], which contains WiFi CSI traces of 5 users performing 150 actions. Each sample consists of 200 packets captured with 3 receiving antennas and 30 subcarriers, resulting in a shape of $200 \times 30 \times 3$. Following the original experimental design, we treat each user as an independent domain due to their unique behavioral characteristics. For model training, we utilize a subset of 1,000 real samples, consistent with the original implementation.

In-domain Amplification. RF-Diffusion is trained on the Widar dataset [59], a widely used dataset for gesture recognition with WiFi. This dataset includes CSI traces from 17 users performing 22 gestures in 5 different locations and orientations. Each sample from a single receiver has the shape of $T \times 90$, where T is the packet amount. As its data partition is not open sourced, we employ the same dataset partition of RFBoost [38] with the 6 most frequent gestures from 15 users across three distinct scenarios, whose train set has 13,275 samples.

7.1.2 Evaluation Metrics. Given the goal of task performance enhancement with synthetic data, we use task performance as the primary evaluation criterion. We also re-assess the affinity and diversity of filtered synthetic data. The enhanced task performance and refined data quality will underscore the necessity of quality-guided utilization and the effectiveness of SYNCHECK in selecting high-quality synthetic data.

7.1.3 Baseline Methods for Comparison. We compare the performance of SYNCHECK against the following baselines: (1) *Real data only*, representing traditional training without synthetic data; (2) *Nonselective mixture* of synthetic and real data, a common practice in prior work (see Tab. 2); (3) *Synthetic data filtering with SSIM*, which selects synthetic data based on visual similarity to real data; and (4) *Synthetic data filtering with TRTS*, which trains a model on real data and excludes synthetic samples whose model predictions deviate from their labels (generation conditions). In contrast, SYNCHECK introduces iterative filtering and pseudo-label assignment for synthetic data utilization. Our evaluation aims to answer the following questions:

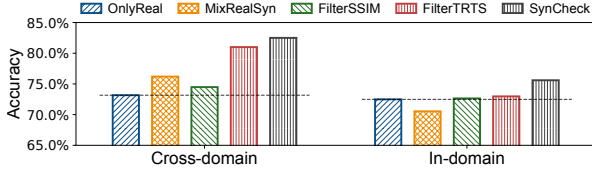


Figure 9: Overall Performance.

RQ4: How does SYNCHECK perform in task performance and data quality attributes of affinity and diversity?

RQ5: What is the impact of design choices, including loss components, backbone model, synthetic data volume, and label origin, in the SYNCHECK framework?

RQ6: What is the role of conditions in synthetic data generation and how do they affect data quality?

7.2 Overall Performance

End-to-end Performance on Target Task. The performance of SYNCHECK on the target task, compared to baseline methods, is shown in Fig. 9. The dashed line represents training with only real data. SYNCHECK outperforms all baselines, achieving 82.7% in cross-domain transfer and 75.6% in in-domain amplification, surpassing the nonselective mixture of synthetic and real data (76.2% and 70.5%) and the real-data-only baseline (73.2% and 72.5%). This represents an 8.6% and 7.2% improvement over nonselective utilization, highlighting SYNCHECK’s effectiveness in quality-guided synthetic data utilization. Performance degradation in the nonselective mixture stems from affinity limitations in cross-domain scenarios, while diversity gains partially mitigate this in in-domain setups. Synthetic data filtering with SSIM achieves 74.5% and 72.6%, showing minor gains over real data by addressing affinity through visual similarity. A stronger baseline, TRTS filtering, achieves 81.0% and 73.0%, demonstrating the advantage of task-oriented selection. However, TRTS filtering still falls short of SYNCHECK, as it relies on generation conditions as labels and causes mislabeled data.

Quality Re-evaluation. We re-evaluate the quality attributes of affinity and diversity for the filtered and relabeled synthetic data with SYNCHECK, as shown by the ‘test’ in Fig. 10. In contrast, the quality assessment of unfiltered synthetic data is denoted as ‘unfiltered test’. The filtered synthetic data demonstrates improved correlation between the TR margins of the train and test sets in both cross-domain and in-domain generation, indicating *improved affinity*. Moreover, the distribution of the TS margin before and after filtering remains similar, suggesting *preserved diversity*. To quantify this improvement, we calculate the JS Divergence between the margins of the training set and test set in Tab. 3. Our results show that SYNCHECK significantly improves the TR-margin, achieving a 57.3% and 29.5% reduction in JS Divergence for TR margins, thereby enhancing data affinity. In contrast, the TS margin divergence increases only slightly, preserving the diversity of synthetic data. This quality enhancement aligns with task performance, further validating the effectiveness of our quality assessment metrics.

Answer to RQ4: SYNCHECK achieves superior task performance and enhanced data affinity. These results underscore the importance of quality-guided utilization of synthetic data.

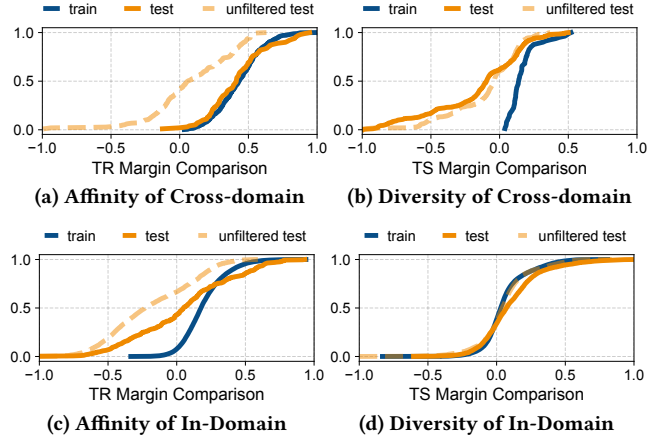


Figure 10: Quality re-quantification of the filtered synthetic data with SYNCHECK compared to original unfiltered data.

Table 3: Comparison of affinity and diversity between the original and SYNCHECK filtered synthetic data, where smaller scores indicates better affinity and diversity, respectively.

Methods	JS Div of TR Margin (\downarrow)		JS Div of TS Margin (\downarrow)	
	Unfiltered	SYNCHECK	Unfiltered	SYNCHECK
CsiGAN [6]	0.503	0.215	0.640	0.657
RF-Diffusion [9]	0.694	0.489	0.135	0.176

7.3 Ablation Studies

We conduct several ablation studies to validate the effectiveness and necessity of the loss function components in the semi-supervised learning framework. We categorize these losses into two groups based on their functions: (1) The one-vs-all loss L_{ova} for per-class inlier-outlier detectors g^j , which are trained on real labeled data to facilitate outlier detection on unlabeled synthetic data. (2) The consistency loss L_{cons} and entropy loss L_{ent} , applied to unlabeled synthetic data. These losses are designed to enforce the prediction robustness and discrimination of synthetic data for the target task and inlier-outlier detection.

Table 4: Ablation Studies for Loss Function Design.

Design Choices	Cross-domain Acc.	In-domain Acc.
Full SYNCHECK	0.827	0.756
w/o L_{ova}	0.760	0.746
w/o L_{cons} and L_{ent}	0.788	0.751
Nonselective mixture	0.762	0.705

Tab. 4 presents the task performance comparison for ablated models. The original SYNCHECK achieves the highest performance in both in-domain and cross-domain setups, underscoring the importance of all loss components. In contrast, removing the one-vs-all loss L_{ova} leads to a 8.10% and 1.3% drop in task accuracy for cross-domain and in-domain setups, respectively. Similarly, removing the consistency and entropy constraints (L_{cons} and L_{ent}) on unlabeled synthetic data results in a 4.72% and 0.67% decline in performance.

7.4 Microbenchmark

We analyze the impact of the backbone models, original synthetic data volume and generation conditions on task performance with

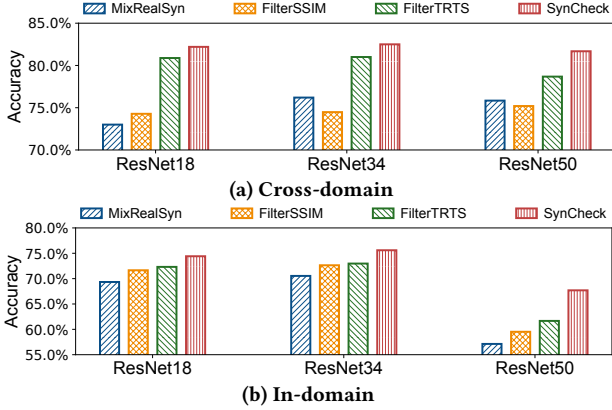


Figure 11: Performance v.s. Backbone Models.

SYNCHECK. In our semi-supervised learning framework, the backbone model functions as a feature extractor, providing outputs that are used by task classifier f and per-class inlier-outlier detector g^j .

Impact of Backbone Model. The results for different backbone models are presented in Fig. 11. SYNCHECK consistently achieves superior task performance across all architectures, demonstrating its robustness and adaptability. Furthermore, the performance variations with RF-Diffusion highlight that increasing model parameters does not guarantee improved performance and may instead increase the risk of overfitting.

Impact of Synthetic Data Amount. We investigate the impact of synthetic data volume on task performance, as illustrated in Fig. 12, by varying the initial synthetic data amounts from 0% to 500% of the real data in increments of 50%. We compare SYNCHECK with the three synthetic data utilization methods employed in the overall performance evaluation. SYNCHECK consistently outperforms all three baseline methods across varying volumes of synthetic data. For instance, with 50% synthetic data, SYNCHECK achieves performance metrics of 81.8% and 75.5% for in-domain and cross-domain setups, respectively, compared to 76.8% and 70.5% achieved by the nonselective mixture approach. This corresponds to a significant performance improvement of 6.5% and 7.1%, respectively. In the in-domain setup with 500% synthetic data, the nonselective mixture approach results in a 13.4% performance degradation compared to using real data alone, whereas SYNCHECK demonstrates a 4.3% performance improvement. Overall, filtering synthetic data with TRTS using the task model yields superior results compared to filtering based on SSIM, highlighting the advantage of task-oriented filtering over methods relying on visual similarity metrics.

As the volume of synthetic data increases, the performance of the nonselective mixture approach exhibits distinct trends in cross-domain and in-domain setups. In the cross-domain scenario, the performance of the nonselective mixture generally improves, primarily due to the increased diversity of the trained classes. However, this performance gain saturates once the synthetic data exceeds 150% of the real data volume, indicating that the diversity benefits of synthetic data reach a point of diminishing returns. On the contrary, in the in-domain setup, the nonselective mixture approach results in consistent performance degradation. This suggests that the affinity limitation cannot be mitigated by increased diversity, leading to train set distribution shifts from over-reliance on synthetic data.

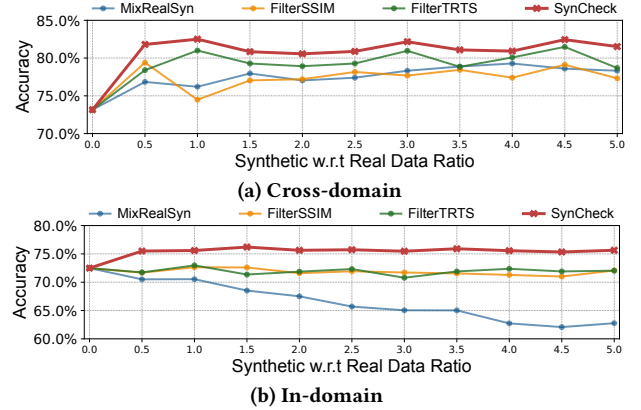


Figure 12: Performance v.s. Synthetic Data Amount.

In contrast, methods incorporating synthetic data filtering do not experience performance degradation with increased synthetic data volume, as they ensure the selection of data with better affinity. The gradual increase in synthetic data enhances diversity, contributing to further improvements in task performance. However, the rate of performance improvement begins to slow once the synthetic data volume exceeds 100% of the real data.

Impact of Pseudo-label Origin. We assess the task performance of filtering synthetic data while utilizing generation conditions as labels. Specifically, we maintain the data filtering mechanism of SynCheck but replace the assigned pseudo-labels with generation conditions as labels during the training of the task model, referred to as ‘Filter+CondLabel’ in Fig. 13. Our evaluation reveals that replacing inlier labels with generation conditions leads to a decline in performance compared to the original SYNCHECK. This performance degradation can be attributed to a dual loss of affinity and diversity: the loss of affinity arises from the use of generation conditions as labels, while the loss of diversity results from the reduced volume of synthetic data after filtering. The observed decline underscores the critical importance of preserving both affinity and diversity in synthetic data utilization for optimal task performance.

Answer to RQ5: Each design component in SYNCHECK contributes to task performance enhancement in the semi-supervised learning framework. While other factors, including backbone model and volume of initial synthetic data, affect the absolute performance, SYNCHECK consistently outperforms the nonselective mixture and other data selection methods.

Necessity of Generation Conditions. To underscore the significance of generation conditions, we replace conditional generative models with unconditional ones. The task performance using synthetic data generated from unconditional models, denoted as ‘UncondSYNCHECK’ in Fig. 13, exhibits a decline compared to its conditional counterpart. This performance degradation emphasizes the necessity of generation conditions in synthetic data quality. We attribute this decline to the absence of generation control [74]. The condition-based control ensures that generative models produce structured and meaningful samples. In contrast, unconditional models, lacking such constraints, inefficiently explore the data space [75]

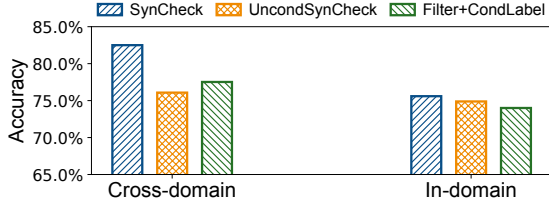


Figure 13: Condition assessment: pseudo-label origin and generation control.

Table 5: Per-batch Training Overhead

Methods	Parameters (M)	FLOPs (G)	GPU memory (MB)
Standard Model	7.58	15.2491	337.96
SYNCHECK	7.59	15.2495	606.42

during both training and inference, leading to the generation of lower-quality synthetic data.

The aforementioned experiments underscore the significance of conditions in both the data synthesis and data utilization processes. However, the cross-domain setup exhibits more pronounced performance degradation compared to the in-domain setup. We attribute this variation in performance to the data synthesis methodology employed by the selected in-domain generative model, RF-Diffusion. Specifically, RF-Diffusion operates by purifying noisy real-world data, where the data itself, apart from generation conditions, provides more informative cues than the cross-domain synthesis method utilizing a GAN model.

Answer to RQ6: While wireless synthetic data from conditional generative models exhibits limited affinity, incorporating conditions during generative models’ training is essential to ensure controllable and beneficial diversity.

7.5 Computational Overhead Analysis

We evaluate the computational overhead of SYNCHECK in comparison to baseline methods. The additional costs incurred by SYNCHECK arise from two main factors: (1) the increased parameter amount due to per-class inlier-outlier detectors, and (2) the additional forward passes required for synthetic data with domain-specific processing to enable pseudo-label supervision. These factors only impact the training phase, as inference in SYNCHECK relies solely on the task classifier and does not involve inlier-outlier detection or pseudo-label assignment. In contrast, all baseline methods employ standard task training procedures. When filtering is applied (*e.g.*, SSIM or TRTS), synthetic data is selected before training. As a result, the per-batch computational cost remains consistent across all baselines. To quantify the overhead, we compare SYNCHECK and the baselines in three aspects: parameter count, FLOPs (floating-point operations), and GPU memory consumption. The results are presented in Tab. 5. It turns out that SYNCHECK introduces a 0.13% growth in parameters and a negligible 0.0026% rise in FLOPs. However, GPU memory usage during training nearly doubles compared to standard training. This overhead is incurred only during training and can be amortized during inference, making the additional resource requirement acceptable given the accuracy improvement.

8 Discussion

Data Quality Metrics. We adopt the principles of using task performance as the ultimate standard for data quality [76] and assessing quality on two key attributes: affinity and diversity. Our quality metrics have demonstrated their effectiveness through correlations with target task performance. While prior work identifies another attribute of *generalization* [19], it stresses no direct copy of real data for privacy issues [77] and beyond the scope of our paper.

Trade-offs between Affinity and Diversity. Affinity and diversity are interdependent attributes of synthetic data, where improving one may compromise the other. In this paper, we focus on enhancing affinity while preserving diversity to improve task performance compared to the quality-oblivious utilization. The optimal balance between the two quality metrics is left for future work.

Empirical Analysis Scale. Our approach to selecting representative generative models balances two criteria: (1) generation techniques, which encompass the widely adopted frameworks of GAN (implicit likelihood estimation [78]), and diffusion models (explicit likelihood estimation [79]); and (2) generation domains, which include both cross-domain and in-domain configurations to ensure coverage of diverse data generation scenarios. Additionally, to maintain evaluation fidelity and reproducibility, we prioritize models with open-source implementations, enabling fair and consistent comparison of our data utilization methods.

Better Conditional Generative Models. Given the deficiencies of existing wireless generative models, an alternative solution is to develop improved generative models tailored to the wireless domain. This approach is orthogonal to our utilization scheme and can be seamlessly integrated to further enhance task performance.

Availability of High-quality Synthetic Data. The efficacy of our data selection framework relies on generative models’ ability to produce synthetic data with high affinity. Our experiments show that modern generative models can reliably produce high-affinity samples with enough sampling attempts.

Applicability to Other Tasks and Modalities. While SYNCHECK primarily focuses on classification tasks within the WiFi domain, there have been efforts to integrate synthetic data with real data for non-classification tasks and other modalities, such as channel prediction [80] and mmWave applications [10]. For regression tasks, techniques from anomaly detection in time-series data [81] could be adapted for inlier-outlier detection. The detection aligns with existing anomaly detection methods that map data to a unified space and compute anomaly scores [82]. For other modalities, since our framework’s design is not inherently tied to specific signal properties, SYNCHECK can be extended by adapting the domain-specific processing in the label assignment phase. We leave the exploration of these directions to future work.

9 Related Work

Data-driven Generative Models. We focus on evaluating the quality of wireless synthetic data generated by data-driven generative models and propose a taxonomy that categorizes these models into three groups based on the source and target data domains,

as detailed in §4.1. In addition to the representative models considered in our assessment, other relevant work in cross-domain transfer includes DFAR [51], which aligns cross-domain samples using decision boundaries and centers, and SCTRFL [52] as well as AutoFi [53], which calibrates samples distorted by temporal environmental changes using autoencoders. In-domain generative models include RF-EATS [43] amplifying liquid-related measurements using variational autoencoders, and AF-DCGAN [55] extending localization fingerprints. Although these works use closed-source models and datasets, which complicate the fair evaluation of original systems, their design rationales align with representative models in our study. Therefore, we argue that they can also benefit from quality-guided utilization for better performance.

Physics-Informed Generative Models. In addition to data-driven generative models, another category includes physics-informed models, which integrate physical laws during data generation. The differences in their mechanisms mean our current framework cannot be directly applied to purely physics-based models. However, existing works on physical models tend to combine with data-driven models [10, 43] for synthetic data refinement. This partial reliance on data-driven elements suggests potential applicability of our method, and we leave this exploration for future work.

Semi-supervised Learning. Semi-supervised learning leverages both labeled and unlabeled data for model training. Key techniques include decision consistency regularization across different augmented versions of the same sample [68] and pseudo-label assignment for unlabeled data [83]. While our semi-supervised framework builds on these established methods, it uniquely treats synthetic data as unlabeled and selects a useful subset based on its quality. This approach offers a plug-and-play solution to address synthetic data quality issues, distinguishing it from most prior semi-supervised learning that focuses solely on real data.

10 Conclusion

While synthetic data complements the data quantity for the wireless community, its quality has often been overlooked, limiting its ability to enhance task performance. In this paper, we introduce tractable and generalizable quantification metrics for two key data quality attributes—affinity and diversity—to systematically evaluate the quality of existing wireless synthetic data. Based on insights derived from our quality assessment, we propose a universal utilization scheme that enhances data affinity while preserving diversity, leading to improved task performance. We anticipate that the quantification of affinity and diversity will become a standard practice in future evaluations of wireless synthetic data. Additionally, our quality-aware utilization framework can be seamlessly integrated with various wireless generative models to boost their performance.

A Theorem Proof

Lemma 1. Given the task of predicting label y from sample x , the performance optimization is achieved through training a classifier $\hat{y} = f(x)$ to minimize the loss [34]

$$\mathbb{E}_{p(x,y)}[\mathcal{L}(y, \hat{y})] = \mathbb{E}_{p(x)}[\mathbb{E}_{p(y|x)}[\mathcal{L}(y, \hat{y} | X = x)]] \quad (5)$$

During training with concrete samples, we empirically optimize the loss $\frac{1}{N}\mathcal{L}(y, \hat{y})$ and the optimal prediction is:

$$\hat{y} = \arg \min \mathbb{E}_{p(y|x)}[\mathcal{L}(y, \hat{y} | X = x)] \quad (6)$$

Definition 4 (Total Variation). Consider p, q probabilities on E , the total variation (TV) distance [84] between p and q is $\text{TV}(p, q) = \sup_{A \subseteq E} |p(A) - q(A)| = \frac{1}{2} \int |p(x) - q(x)| dx$.

PROOF OF THEOREM 1. Based on Eqn. 6, given the training data distribution $p(x, y)$, the test loss under the data distribution $p_\theta(x, y)$ is expressed as:

$$\mathbb{E}_{TRTS}[\mathcal{L}(y, \hat{y})] = \mathbb{E}_{p_\theta(x)}[\mathbb{E}_{p(y|x)}[\mathcal{L}(y, \hat{y} | X = x)]] .$$

The marginal distributions of synthetic and real samples are:

$$p_\theta(x) = \int_y p_\theta(x, y) dy, \quad p(x) = \int_y p(x, y) dy.$$

For any measurable set A , the probability difference satisfies:

$$|p_\theta(x \in A) - p(x \in A)| \leq \int_y |p_\theta(x|y) - p(x|y)| p(y) dy,$$

which implies proximity in the total variation (TV) distance:

$$\text{TV}(p_\theta, p) \leq \mathbb{E}_{p(y)}[\text{TV}(p_\theta(x|y), p(x|y))] .$$

Thus, similarity in the conditional distributions ensures similarity in the marginal distributions. Let $h(x) = \mathbb{E}_{p(y|x)}[\mathcal{L}(y, \hat{y} | X = x)]$. The difference between the training and testing losses is given by:

$$|\mathcal{L}_{\text{test}} - \mathcal{L}_{\text{train}}| = |\mathbb{E}_{p_\theta(x)}[h(x)] - \mathbb{E}_{p(x)}[h(x)]| .$$

If $h(x)$ is bounded (e.g., $\mathcal{L} \in [0, 1]$), it follows that:

$$\begin{aligned} |\mathbb{E}_{p_\theta}[h] - \mathbb{E}_p[h]| &\leq 2 \cdot \text{TV}(p_\theta, p) \cdot \sup |h(x)| \\ &\leq 2 \cdot \mathbb{E}_{p(y)}[\text{TV}(p_\theta(x|y), p(x|y))] \cdot \sup |h(x)|. \end{aligned}$$

This implies that the test loss $\mathcal{L}_{\text{test}}$ approximates the training loss $\mathcal{L}_{\text{train}}$. In summary, under the sufficient condition that the synthetic data distribution $p_\theta(x|y)$ is close to the real data distribution $p(x|y)$, a model achieving a low training loss will also achieve a low testing loss on synthetic data. \square

PROOF OF THEOREM 2. With training distribution $p_\theta(x, y)$, the test loss under the data distribution $p(x, y)$ is given by:

$$\mathbb{E}_{TSTR}[\mathcal{L}(y, \hat{y})] = \mathbb{E}_{p(x)}[\mathbb{E}_{p_\theta(y|x)}[\mathcal{L}(y, \hat{y} | X = x)]]$$

Under the sufficient condition that $p_\theta(y|x) \approx p(y|x)$, the test loss on real data becomes:

$$\mathbb{E}_{p(x)}[\mathbb{E}_{p_\theta(y|x)}[\mathcal{L}(y, \hat{y} | X = x)]] .$$

This aligns with training loss on real data, resulting in low test loss. Therefore, consistency between $p_\theta(y|x)$ and $p(y|x)$ ensures good test performance in train-synthetic setup. \square

Acknowledgments

We sincerely thank the anonymous shepherd and reviewers for their insightful critique and constructive feedback, which have significantly improved the quality of this paper. This work is supported by National Key R&D Program of China (Grant No.2023YFF0725004) and National Natural Science Foundation of China (Grant No.62272010 and 62061146001). Chenren Xu (chenren@pku.edu.cn) is the corresponding author.

References

- [1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.
- [2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 2014.
- [3] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *NeurIPS*, 2019.
- [4] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *ArXiv:1411.1784*, 2014.
- [5] Lei Kang, Pau Riba, Yaxing Wang, Marçal Rusiñol, Alicia Fornés, and Mauricio Villegas. Ganwriting: Content-conditioned generation of styled handwritten word images. *Springer ECCV*, 2020.
- [6] Chunjing Xiao, Daojun Han, Yongsan Ma, and Zhiguang Qin. Csgan: Robust channel state information-based activity recognition with gans. *IEEE Internet of Things Journal*, 2019.
- [7] Sevgi Zubeyde Gurbuz, Mohammad Muntasir Rahman, Emre Kurtoglu, Trevor Macks, and Francesco Fioranelli. Cross-frequency training with adversarial learning for radar micro-doppler signature classification. In *Commercial Sensing*, 2020.
- [8] Xi Chen, Hang Li, Chenyi Zhou, Xue Liu, Di Wu, and Gregory Dudek. Fido: Ubiquitous fine-grained wifi-based localization for unlabelled users via domain adaptation. In *ACM WWW*, 2020.
- [9] Guoxuan Chi, Zheng Yang, Chenshu Wu, Jingao Xu, Yuchong Gao, Yunhao Liu, and Tony Xiao Han. RF-diffusion: Radio signal generation via time-frequency diffusion. In *ACM MobiCom*, 2024.
- [10] Xingyu Chen and Xinyu Zhang. Rf genesis: Zero-shot generalization of mmwave sensing through simulation-based data synthesis and generative diffusion models. In *ACM SenSys*, 2023.
- [11] Xinyi Li, Liqiong Chang, Fangfang Song, Ju Wang, Xiaojian Chen, Zhanyong Tang, and Zheng Wang. Crossgr: Accurate and low-cost cross-target gesture recognition using wi-fi. *ACM IMWUT*, 2021.
- [12] Xi Chen, Hang Li, Chenyi Zhou, Xue Liu, Di Wu, and Gregory Dudek. Fidora: Robust wifi-based indoor localization via unsupervised domain adaptation. *IEEE Internet of Things Journal*, 2022.
- [13] Nripendra Kumar Singh and Khalid Raza. Medical image generation using generative adversarial networks: A review. *Health informatics: A computational perspective in healthcare*, 2021.
- [14] Chen Chen, Gang Zhou, and Youfang Lin. Cross-domain wifi sensing with channel state information: A survey. *ACM Computing Surveys*, 2023.
- [15] Jie Wang, Qinhua Gao, Xiaorui Ma, Yunong Zhao, and Yuguang Fang. Learning to sense: Deep learning for wireless sensing with less training efforts. *IEEE Wireless Communications*, 2020.
- [16] Lisa C Adams, Felix Busch, Daniel Truhn, Marcus R Makowski, Hugo JW Aerts, and Keno K Bressen. What does dall-e 2 know about radiology? *Journal of Medical Internet Research*, 2023.
- [17] Pierre Chambon, Christian Bluthgen, Curtis P Langlotz, and Akshay Chaudhari. Adapting pretrained vision-language foundational models to medical imaging domains. *arXiv:2210.04133*, 2022.
- [18] Shenghuan Sun, Greg Goldf, Atul Butte, and Ahmed M Alaa. Aligning synthetic medical images with clinical knowledge using human feedback. *NeurIPS*, 2024.
- [19] Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *ICML*, 2022.
- [20] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 2021.
- [21] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. *NeurIPS*, 2018.
- [22] George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villegas, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *NeurIPS*, 2024.
- [23] Xin Li, Hongbo Wang, Zhe Chen, Zhiping Jiang, and Jun Luo. Uwb-fi: Pushing wi-fi towards ultra-wideband for fine-granularity sensing. In *MobiSys*, 2024.
- [24] Xiaomin Ouyang, Jason Wu, Tomoyoshi Kimura, Yihan Lin, Gunjan Verma, Tarek Abdelzaher, and Mani Srivastava. Mmbind: Unleashing the potential of distributed and heterogeneous data for multimodal learning in iot. *arXiv:2411.12126*, 2024.
- [25] Xiaomin Ouyang and Mani Srivastava. Lmsense: Harnessing llms for high-level reasoning over spatiotemporal sensor traces. *arXiv:2403.19857*, 2024.
- [26] Shenghong Dai, Shiqi Jiang, Yifan Yang, Ting Cao, Mo Li, Suman Banerjee, and Lili Qiu. Advancing multi-modal sensing through expandable modality alignment. *arXiv:2407.17777*, 2024.
- [27] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. How good is my gan? In *Springer ECCV*, 2018.
- [28] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *NeurIPS*, 2017.
- [29] Gamaleldin Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. *NeurIPS*, 2018.
- [30] Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. Identifying mislabeled data using the area under the margin ranking. *NeurIPS*, 2020.
- [31] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 2021.
- [32] Hongfei Xue, Qiming Cao, Chenglin Miao, Yan Ju, Haochen Hu, Aidong Zhang, and Lu Su. Towards generalized mmwave-based human pose estimation through signal augmentation. In *ACM MobiCom*, 2023.
- [33] Kancharla Parimala and Sumohana Channappayya. Quality aware generative adversarial networks. *NeurIPS*, 2019.
- [34] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *NeurIPS*, 2019.
- [35] Hong Cai, Belal Korany, Chitra R Karanam, and Yasamin Mostofi. Teaching rf to sense without rf training measurements. *ACM IMWUT*, 2020.
- [36] Yongsan Ma, Gang Zhou, and Shuangquan Wang. Wifi sensing with channel state information: A survey. *ACM Computing Surveys*, 2019.
- [37] Haochen Hu, Zihan Sun, Chao Li, Yong Zhang, and Chunxiao Xing. Survey of time series data generation in iot. *Sensors*, 2023.
- [38] Weiyang Hou and Chenshu Wu. Rfbost: Understanding and boosting deep wifi sensing via physical data augmentation. *ACM IMWUT*, 2024.
- [39] Omer Gokalp Serbetci, Ju-Hyung Lee, Daoud Burghal, and Andreas F Molisch. Simple and effective augmentation methods for csi based indoor localization. In *IEEE GLOBECOM*, 2023.
- [40] Julian Strohmer and Martin Koppel. Data augmentation techniques for cross-domain wifi csi-based human activity recognition. In *IFIP AIAI*, 2024.
- [41] John W McKown and R Lee Hamilton. Ray tracing as a design tool for radio networks. *IEEE Network*, 1991.
- [42] Xiaopeng Zhao, Zhenlin An, Qingrui Pan, and Lei Yang. Nerf2: Neural radio-frequency radiance fields. In *ACM MobiCom*, 2023.
- [43] Unsoo Ha, Junshan Leng, Alaa Khaddaj, and Fadel Adib. Food and liquid sensing in practical environments using {RFIDs}. In *NSDI*, 2020.
- [44] Alvaro Figueira and Bruno Vaz. Survey on synthetic data generation, evaluation methods and gans. *MDPI Mathematics*, 2022.
- [45] Ali Borji. Pros and cons of gan evaluation measures. *Elsevier Computer vision and image understanding*, 2019.
- [46] Kancharla Parimala and Sumohana Channappayya. Quality aware generative adversarial networks. *NeurIPS*, 2019.
- [47] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017.
- [48] Florence Regol and Mark Coates. Categorical generative model evaluation via synthetic distribution coarsening. In *AISTATS*, 2024.
- [49] Wikipedia contributors. Bayes' theorem — Wikipedia, the free encyclopedia, 2025. [Online; accessed 29-March-2025].
- [50] Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *Springer MICCAI*, 2017.
- [51] Jie Wang, Yunong Zhao, Xiaorui Ma, Qinghua Gao, Miao Pan, and Hongyu Wang. Cross-scenario device-free activity recognition based on deep adversarial networks. *IEEE Transactions on Vehicular Technology*, 2020.
- [52] Lili Zheng, Bin-Jie Hu, Jinguang Qiu, and Manman Cui. A deep-learning-based self-calibration time-reversal fingerprinting localization approach on wi-fi platform. *IEEE Internet of Things Journal*, 2020.
- [53] Xi Chen, Chen Ma, Michel Allegue, and Xue Liu. Taming the inconsistency of wi-fi fingerprints for device-free passive indoor localization. In *IEEE INFOCOM*, 2017.
- [54] Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. In *ICLR*, 2016.
- [55] Qiyue Li, Heng Qu, Zhi Liu, Nana Zhou, Wei Sun, Stephan Sigg, and Jie Li. Af-dcgan: Amplitude feature deep convolutional gan for fingerprint construction in indoor localization systems. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2021.
- [56] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF CVPR*, 2016.
- [57] Jianfei Yang, Xinyan Chen, Dazhuo Wang, Han Zou, Chris Xiaoxuan Lu, Sumei Sun, and Lihua Xie. Sensefi: A library and benchmark on deep-learning-empowered wifi human sensing. *Elsevier Patterns*, 2023.
- [58] Yongsan Ma, Gang Zhou, Shuangquan Wang, Hongyang Zhao, and Woosub Jung. Signfi: Sign language recognition using wifi. *ACM IMWUT*, 2018.
- [59] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. Zero-effort cross-domain gesture recognition with wi-fi. In *MobiSys*, 2019.
- [60] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv:1609.04836*, 2016.

- [61] Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. *arXiv:1802.08760*, 2018.
- [62] María Luisa Menéndez, JA Pardo, L Pardo, and MC Pardo. The jensen-shannon divergence. *Journal of the Franklin Institute*, 1997.
- [63] Aditya Virmani and Muhammad Shahzad. Position and orientation agnostic gesture recognition using wifi. In *MobiSys*, 2017.
- [64] Shuokang Huang, Po-Yu Chen, and Julie Ann McCann. Diffar: Adaptive conditional diffusion model for temporal-augmented human activity recognition. In *IJCAI*, 2023.
- [65] Jongjin Park, Sukmin Yun, Jongheon Jeong, and Jinwoo Shin. Opencos: Contrastive semi-supervised learning for handling open-set unlabeled data. In *Springer ECCV*, 2022.
- [66] Zekun Li, Lei Qi, Yinghuan Shi, and Yang Gao. Iomatch: Simplifying open-set semi-supervised learning with joint inliers and outliers utilization. In *IEEE/CVF ICCV*, 2023.
- [67] Kuniaki Saito and Kate Saenko. Ovanet: One-vs-all network for universal domain adaptation. In *IEEE/CVF ICCV*, 2021.
- [68] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv:1610.02242*, 2016.
- [69] J.G. Proakis and M. Salehi. *Digital Communications*. McGraw-Hill, 2008.
- [70] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NeurIPS*, 2020.
- [71] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019.
- [72] Chunjing Xiao, Daojun Han, Yongsun Ma, and Zhiguang Qin. Csigan: Robust channel state information-based activity recognition with gans. <https://github.com/ChunjingXiao/CsiGAN>, 2019.
- [73] Weiying Hou and Chenshu Wu. Rfboost: Understanding and boosting deep wifi sensing via physical data augmentation. <https://github.com/aiot-lab/RfBoost>, 2024.
- [74] Alexandros Graikos, Srikar Yellapragada, and Dimitris Samaras. Conditional generation from unconditional diffusion models using denoiser representations. *arXiv:2306.01900*, 2023.
- [75] Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. Vaes meet diffusion models: Efficient and high-fidelity generation. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021.
- [76] Raphael Gontijo-Lopes, Sylvia J Smullin, Ekin D Cubuk, and Ethan Dyer. Affinity and diversity: Quantifying mechanisms of data augmentation. *arXiv:2002.08973*, 2020.
- [77] James Jordon, Daniel Jarrett, Evgeny Saveliev, Jinsung Yoon, Paul Elbers, Patrick Thoral, Ari Ercole, Cheng Zhang, Danielle Belgrave, and Mihaela van der Schaar. Hide-and-seek privacy challenge: Synthetic data generation vs. patient re-identification. In *NeurIPS Competition and Demonstration Track*, 2021.
- [78] Aditya Grover, Manik Dhar, and Stefano Ermon. Flow-gan: Combining maximum likelihood and adversarial learning in generative models. In *AAAI*, 2018.
- [79] Kushagra Pandey and Stephan Mandt. A complete recipe for diffusion generative models. In *IEEE/CVF ICCV*, 2023.
- [80] Haofan Lu, Christopher Vratheuer, Baharan Mirzasoleiman, and Omid Abari. Newrf: A deep learning framework for wireless radiation field reconstruction and channel prediction. In *ICML*, 2024.
- [81] Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A Lozano. A review on outlier/anomaly detection in time series data. *ACM computing surveys*, 2021.
- [82] Ruoying Wang, Kexin Nie, Tie Wang, Yang Yang, and Bo Long. Deep learning for anomaly detection. In *WSDM*, 2020.
- [83] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, 2013.
- [84] Wikipedia contributors. Total variation distance of probability measures — Wikipedia, the free encyclopedia, 2025. [Online; accessed 18-March-2025].