# Adaptive Gradient Methods for Differentially Private TinyML in 6G

Chen Hou, Tao Huang, Qingyu Huang, Xu Yang, Xiaoding Wang, *Member, IEEE*, Jia Hu, Sunder Ali Khowaja, *Senior Member, IEEE*, and Kapal Dev, *Senior Member, IEEE*

## Abstract

The sixth-generation (6G) of wireless systems is poised to enable a hyper-connected world of intelligent devices, where tiny machine learning (TinyML) will drive pervasive, real-time applications. However, this paradigm, built on distributed data from billions of endpoints, introduces an unprecedented privacy attack surface. A fundamental challenge for deploying AI in 6G is ensuring robust data privacy on resource-constrained devices without sacrificing model utility. Differentially Private Stochastic Gradient Descent (DP-SGD), a cornerstone of private machine learning, critically depends on managing gradient sensitivity, a task traditionally hampered by the manual tuning of a static clipping threshold. This paper presents a comprehensive analysis of gradient control mechanisms for DP-SGD, evaluated from a 6G deployment perspective. We trace the evolution from static clipping to fully adaptive scaling methods that obviate the need for a fixed threshold. To unify these approaches, we propose a conceptual framework, culminating in a case study of the Differentially Private Per-sample Adaptive Scaling Clipping (DP-PSASC) algorithm. We argue that such adaptive methods are not merely algorithmic improvements but are essential, "6G-adaptive" solutions that can be integrated into next-generation network architectures, such as the Open RAN (O-RAN) framework, to deliver efficient, scalable, and trustworthy AI.

## I. Introduction

The convergence of 6G wireless systems and Tiny Machine Learning (TinyML) is not merely an incremental advance; it is fundamentally reshaping the landscape of edge intelligence. The vision of 6G, with its promise of massive Machine-Type Communications (mMTC), Ultra-Reliable Low-Latency Communication (URLLC), and integrated sensing capabilities, sets the stage for a world populated by billions of intelligent devices, from smart wearables to autonomous vehicles. This hyper-connected ecosystem is the natural habitat for TinyML, where lightweight models, running directly at the network's edge, become the engines of real-time, adaptive decision-making.

Consider the tangible applications this enables. In smart healthcare, a TinyML model on a wearable sensor, powered by a seamless 6G connection, could perform real-time analysis of vital signs to predict a cardiac event. In the Industrial Internet of Things (IIoT), an edge device could perform predictive maintenance on factory machinery by analyzing subtle vibrations, preventing costly failures. In both scenarios, the immense value is unlocked by training models on deeply personal or proprietary data. Yet, this very reliance on data opens a Pandora's box of privacy vulnerabilities, creating a significant barrier to public trust and adoption.

The threat is not merely theoretical. The massive data aggregation in 6G networks creates fertile ground for sophisticated attacks. A well-funded adversary could use a membership inference attack to determine if a specific person's data was part of a model's training set [1]. Imagine a commercial entity querying a public medical model to see if it confidently recognizes data patterns associated with a certain individual, thereby inferring a private health condition. An even more invasive threat is data reconstruction, where attackers have shown it is possible to reverse-engineer model updates to reconstruct recognizable faces or sensitive text from the training data [1]. Consequently, the central challenge has become clear: we must build powerful models without demanding a sacrifice of privacy.

Differential Privacy (DP) has emerged as the gold standard for providing rigorous, mathematical privacy guarantees. Its most prominent application in deep learning, DP-SGD [2], works by carefully managing the influence of any single data point during training. In each step, gradients from individual samples are first constrained in magnitude and then obscured with calibrated noise. While effective in theory, the practical deployment of DP-SGD has been persistently hindered by its reliance on a fixed gradient clipping threshold. This single hyperparameter presents an untenable dilemma. Set it too low, and you discard

*Chen Hou, Tao Huang, Qingyu Huang, and Xu Yang (corresponding author) are with the School of Computer and Data Science, Minjiang University, Fuzhou 350108, China; Xiaoding Wang (corresponding author) is with the College of Computer and Cyber Security, Fujian Normal University, Fuzhou 350117, China; Jia Hu is with the Department of Computer Science, University of Exeter, EX4 4QJ Exeter, U.K.; Sunder Ali Khowaja is with Dublin City University, 9 (D09) Dublin, Ireland; Kapal Dev is with Munster Technological University, T12 P928 Cork, Ireland.*
*Chen Hou and Tao Huang contributed equally to this work.*

Digital Object Identifier: 10.1109/MWC.2025.3640093

valuable information, introducing a clipping bias that can pull the model away from the optimal solution [3]. Set it too high, and you must add so much noise to maintain privacy that the learning signal is drowned out. This limitation is a critical bottleneck for deploying AI at scale in 6G, where the diversity of devices, data types, and network conditions makes a one-size-fits-all threshold impractical and inefficient.

To overcome this, future networks require solutions that are not just algorithmically sound but also architecturally aware. The 6G network is evolving towards a disaggregated, intelligent, and software-defined infrastructure. The Open Radio Access Network (O-RAN) alliance is at the forefront of this transformation, defining an architecture that decouples network components and introduces intelligence through the RAN Intelligent Controller (RIC). The RIC, with its Non-Real-Time (Non-RT) and Near-Real-Time (Near-RT) control loops, enables the deployment of third-party applications (xApps and rApps) that can monitor, manage, and optimize the network.

This paper posits that advanced privacy-preserving mechanisms should be designed as integral components of this new architecture. We argue that the evolution of gradient control in DP-SGD, from static clipping to adaptive scaling, provides the foundation for developing a "privacy-as-a-service" xApp within the O-RAN framework. Such an application could reside in the Non-RT RIC to manage the long-term privacy policies and budgets for distributed learning tasks across thousands or millions of devices. This paper serves as a tutorial on this evolution, charting a course from static clipping to fully adaptive scaling. We will explore how each new idea attempts to solve the shortcomings of its predecessors, introduce a framework for understanding their core principles, and ground the discussion in a practical case study. Our goal is to pave the way for effective, privacy-preserving TinyML that is not just compatible with 6G but is a native, integrated component of a trustworthy 6G future.

# II. The Evolution of Gradient Control in DP-SGD

At the heart of DP-SGD is the need to control gradient sensitivity—to bound the influence of any single data sample. The story of research in this area is a journey from a simple, rigid rule to sophisticated, adaptive strategies. This evolution has been driven by the quest to resolve the fundamental tension between preserving data utility and ensuring privacy, a tension that is magnified in the complex 6G environment.

## A. The Foundational Paradigm: Static Gradient Clipping

The original approach, introduced alongside DP-SGD, is static gradient clipping [2]. The logic is straightforward: before training, a single, fixed threshold is chosen. During training, if any individual gradient's L2 norm exceeds this threshold, it is scaled down to match it, preserving its direction. Gradients with norms already below the threshold are left untouched.

While its simplicity is appealing, this static approach is a blunt instrument, particularly ill-suited for 6G. In a heterogeneous 6G network comprising a vast array of devices—from low-power IoT sensors in an mMTC scenario to high-computation ECUs in a vehicle leveraging URLLC—a single, static threshold is profoundly inefficient. The optimal threshold for one device type or data distribution will be suboptimal for another, leading to either excessive information loss or insufficient privacy. Furthermore, finding the "goldilocks" threshold that works well throughout the entire training process often requires extensive, computationally expensive tuning, which can ironically consume part of the very privacy budget it is meant to help manage. This core limitation made it clear that a one-size-fits-all approach was insufficient for the dynamic and diverse nature of 6G.

## B. Evolving Approaches: Dynamic Threshold Adjustments

A logical next step was to ask: what if the threshold could change during training? This led to dynamic adjustment methods. The core idea is that the ideal threshold should adapt as the model learns and gradients naturally shrink.

One popular strategy is to use a small part of the privacy budget at each step to privately estimate a good threshold from the current batch of gradients, for instance, by setting it to their median or 90th percentile norm [4]. Another approach is to follow a predefined schedule, such as decreasing the threshold over time [5]. While these methods are an improvement, they often just trade one problem for another, especially in a 6G context. They introduce new hyperparameters to tune (like the quantile or a decay rate), and in the case of quantile-based methods, the act of estimating the threshold itself consumes precious privacy budget and requires additional communication between devices and the edge server. This overhead can be prohibitive in large-scale mMTC scenarios where bandwidth and energy are scarce. They were a step in the right direction, but not the final destination for a truly scalable 6G solution.

## A. Paradigm Shift: Gradient Scaling

A more recent and powerful idea was to abandon the clipping threshold altogether. This led to adaptive scaling approaches, often called Automatic Clipping (Auto-S) or Normalized SGD (NSGD) [6], [7]. Here, instead of clipping, every per-sample gradient is simply divided by its own norm.

This elegant maneuver ensures every processed gradient has a uniform norm (typically 1), completely eliminating the hyperparameter search. However, it created a new, serious problem: the amplification of small gradients. In the late stages of training, gradients are naturally small as the model fine-tunes its parameters near a minimum. This new approach would take these small, often noisy, gradients and amplify their influence enormously. It is akin to using a magnifying glass on shaky hands—the amplified tremor makes precise work impossible and can destabilize learning. This instability is especially problematic for 6G's mission-critical URLLC services, such as autonomous driving or remote surgery, where model reliability is paramount. Furthermore, in 6G Integrated Sensing and Communication (ISAC) applications, where a model might be trained to detect faint environmental signals, amplifying noise could completely obscure the target information.

This deficiency sparked the latest wave of innovation: non-monotonic scaling functions. Methods like DP-PSAC [8] were designed to suppress the influence of very small gradients, preventing their harmful amplification. As we will see, however, even this can

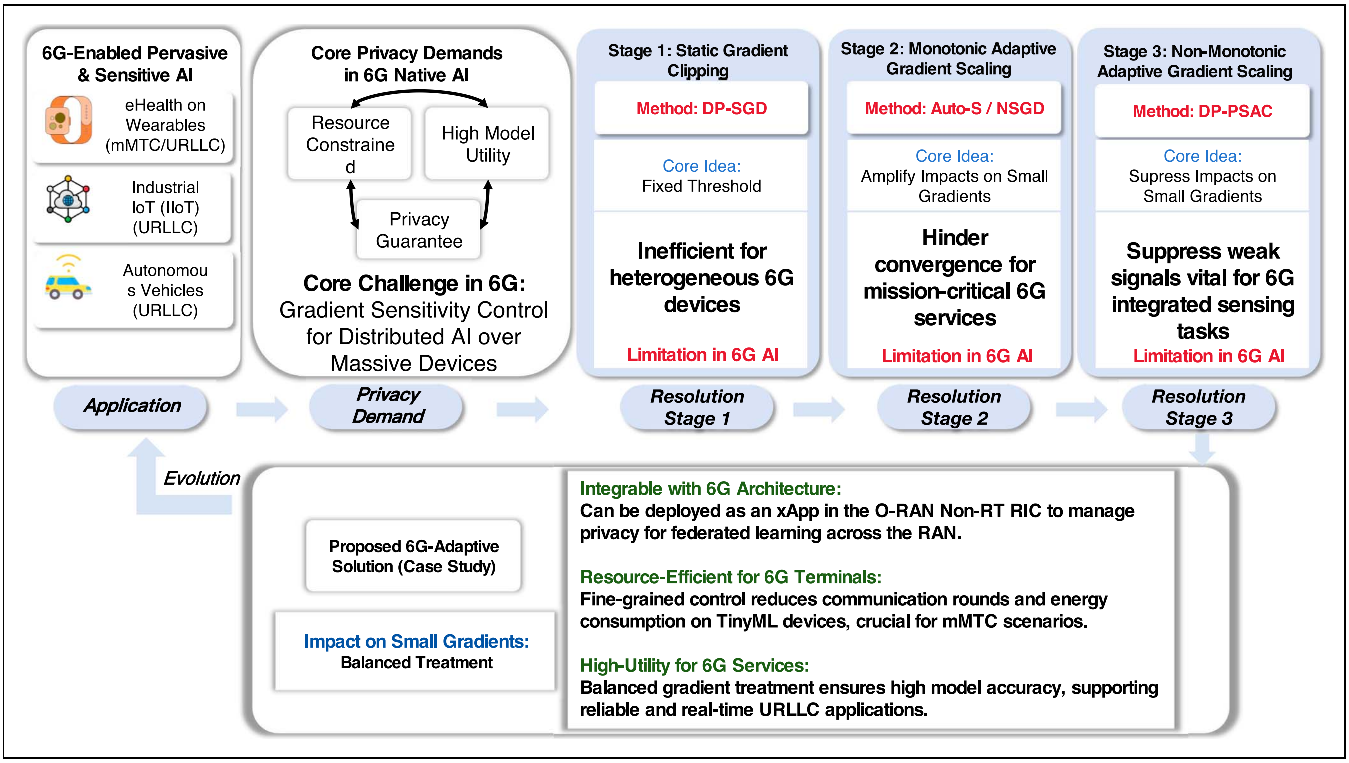

6G-Enabled Pervasive & Sensitive AI
eHealth on Wearables (mMTC/URLLC)
Industrial IoT (IIoT) (URLLC)
Autonomous Vehicles (URLLC)
Core Privacy Demands in 6G Native AI
Resource Constrained
High Model Utility
Privacy Guarantee
Core Challenge in 6G: Gradient Sensitivity Control for Distributed AI over Massive Devices
Stage 1: Static Gradient Clipping
Method: DP-SGD
Core Idea: Fixed Threshold
Inefficient for heterogeneous 6G devices
Limitation in 6G AI
Stage 2: Monotonic Adaptive Gradient Scaling
Method: Auto-S / NSGD
Core Idea: Amplify Impacts on Small Gradients
Hinder convergence for mission-critical 6G services
Limitation in 6G AI
Stage 3: Non-Monotonic Adaptive Gradient Scaling
Method: DP-PSAC
Core Idea: Supress Impacts on Small Gradients
Suppress weak signals vital for 6G integrated sensing tasks
Limitation in 6G AI
Application
Privacy Demand
Resolution Stage 1
Resolution Stage 2
Resolution Stage 3
Evolution
Proposed 6G-Adaptive Solution (Case Study)
Impact on Small Gradients: Balanced Treatment
Integrable with 6G Architecture:
Can be deployed as an xApp in the O-RAN Non-RT RIC to manage privacy for federated learning across the RAN.
Resource-Efficient for 6G Terminals:
Fine-grained control reduces communication rounds and energy consumption on TinyML devices, crucial for mMTC scenarios.
High-Utility for 6G Services:
Balanced gradient treatment ensures high model accuracy, supporting reliable and real-time URLLC applications.

FIG. 1. A conceptual framework illustrating the evolution of gradient control for privacy-preserving AI in 6G. The journey begins with 6G-enabled application demands (e.g., URLLC, mMTC), which create core privacy challenges for distributed AI over massive devices. This challenge is met by a series of increasingly sophisticated resolutions, from static clipping (Stage 1) to monotonic (Stage 2) and non-monotonic (Stage 3) adaptive scaling. The framework highlights how each stage's limitations are particularly acute in a 6G context, culminating in a proposed 6G-adaptive solution that is resource-efficient, high-utility, and architecturally integrable (e.g., as an O-RAN xApp).

be suboptimal, as it risks suppressing the very signals needed for the final, crucial steps of convergence. This ongoing evolution reveals a central theme: an ideal gradient control mechanism for 6G must intelligently and adaptively handle gradients of all sizes across a wide range of devices and applications. We therefore formalize the notion of "6G-adaptive" algorithms below and later connect it to O-RAN control loops and xApp/rApp realizations [9], [10].

## III. A Framework for Analyzing Adaptive Scaling Methods

To better understand the design choices behind these advanced techniques, it is helpful to view them through a unifying framework. We can categorize and compare them based on two key principles: their treatment of small gradients and their adaptability across different training stages, as illustrated in Fig. 1.

We define a method as 6G-adaptive if it satisfies all of the following pillars: (i) resource efficiency on TinyML devices (compute, memory, and communication); (ii) high model utility under DP noise; (iii) scalability to massive device populations (mMTC) with non-IID data; (iv) resilience to dynamic conditions (mobility, varying channel quality, latency bounds for URLLC); and (v) architectural integrability with O-RAN RIC control loops (Non-RT analytics/policy and Near-RT enforcement) [9], [10].

The first, and perhaps most critical, dividing line among these methods is their treatment of small-magnitude gradients. These gradients are chameleons; early in training, they may be noise, but late in training, they carry the critical signal for fine-tuning. Handling them improperly has severe consequences. Amplifying them, as Auto-S does, is like trying to perform surgery with a sledgehammer—it introduces instability that could jeopardize a time-sensitive URLLC application. Conversely, excessively suppressing them, a risk in some non-monotonic methods, can cause training to stall, leaving the model stranded just short of its full potential. The sweet spot lies in a balanced approach that gives these small gradients just enough influence to guide the model home without shaking it apart.

The second dimension is how a method's behavior adapts over time. This perspective reveals a clear evolutionary story that aligns with the needs of persistent learning in 6G. Static clipping is "stage-agnostic," applying the same rigid rule from start to finish, making it brittle to the dynamic data streams from mobile 6G devices. Dynamic threshold methods are "stage-aware," adjusting over time, but their adjustments can be coarse and resource-intensive. The true promise lies in adaptive scaling methods, which offer "continuous stage adaptation." A well-designed scaling function should implicitly account for the natural shift in training dynamics—from the aggressive, large-gradient descents of the early stages to the delicate, small-gradient refinements of the final stages. This adaptability is key for long-running, federated learning tasks in 6G, where models must continuously evolve. This framework provides a lens through which we can now analyze a concrete implementation.

## IV. Case Study: Fine-Grained Gradient Control With DP-PSASC

To make these principles concrete, we now examine an advanced scaling method: Differentially Private

| Method | Core Idea | Handle Small Gradients | Key Limitation | Test Accuracy (%) | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | MNIST | F-MNIST | CIFAR10 | CelebA | Imagenette |
| **DP-SGD** | Fixed, predefined norm threshold. | Unchanged unless norm exceeds threshold. | Highly sensitive to the threshold; suboptimal for the entire training process. | 97.35 | 85.59 | 92.24 | 94.79 | 63.61 |
| **Auto-S** / **NSGD** | Normalizes every gradient to have a uniform norm. | Amplified significantly, giving them disproportionate weight. | Distorts learning by over-emphasizing small/noisy gradients, causing instability. | 97.95 | 86.15 | 92.65 | 95.12 | 63.87 |
| **DP-PSAC** | Non-monotonic scaling function. | Suppressed; assigned smaller weights to reduce influence. | May excessively dampen important signals, hindering final convergence. | 98.11 | 86.36 | 92.85 | 95.20 | 64.21 |
| **DP-PSASC** (Case Study) | Extends DP-PSAC with a tunable coefficient. | Given controlled, larger weights for effective fine-tuning. | Introduces a new, less sensitive hyperparameter that requires tuning. | 98.37 | 86.91 | 93.12 | 95.36 | 65.26 |
| **DP-PSASC*** (Case Study) | Integrates dual momentum with DP-PSASC. | Same as DP-PSASC, but direction is stabilized by momentum. | Increased complexity from the momentum mechanism. | **98.65** | **87.22** | **93.36** | **95.61** | **66.79** |

**TABLE I.** Comprehensive comparison of gradient control methods, detailing their core principles and test accuracy (%) across multiple datasets.

Per-sample Adaptive Scaling Clipping (DP-PSASC). This method and its momentum-enhanced variant, DP-PSASC*, exemplify a targeted, 6G-adaptive solution to the challenges we have discussed.

## A. Method Description

**DP-PSASC: A Tunable Scaling Function.** The core innovation of DP-PSASC is its unique, non-monotonic scaling function, which introduces a tunable coefficient. This coefficient acts as a knob, allowing for more nuanced control over how much weight is assigned to gradients of different magnitudes. By carefully setting this knob, the function can avoid the pitfalls of its predecessors: it can give small gradients more influence than methods that aggressively suppress them, without amplifying them to the dangerous extent of pure normalization. This design directly addresses the need for a balanced treatment of the small gradients that are so vital for the final phase of fine-tuning.

**DP-PSASC*: Integration with Advanced Momentum.** The enhanced version, DP-PSASC*, goes a step further by tackling the instability caused by noisy, single-sample gradients. It does this by integrating a sophisticated inner-outer momentum technique. This is particularly relevant for federated learning in 6G, which often involves non-IID (non-independently and identically distributed) data from clients, leading to noisy and divergent gradient updates. The momentum technique can be understood with an analogy:

- **Inner Momentum:** Before scaling, the algorithm computes a moving average of each individual sample's past gradients. This is like asking each person in a group to consider their recent opinions to form a more stable viewpoint before speaking. It stabilizes the direction of individual gradients, mitigating the effects of local data skew.
- **Outer Momentum:** After these stabilized gradients are scaled and aggregated, a second moving average is applied to the batch-level updates. This is like the group's final decision being a smoothed average of its recent collective decisions, preventing abrupt changes in the model's overall direction and improving convergence stability.

This dual-smoothing process leads to more stable and efficient training, a crucial benefit for resource-constrained TinyML applications running on 6G networks.

## B. Experimental Setup and Analysis

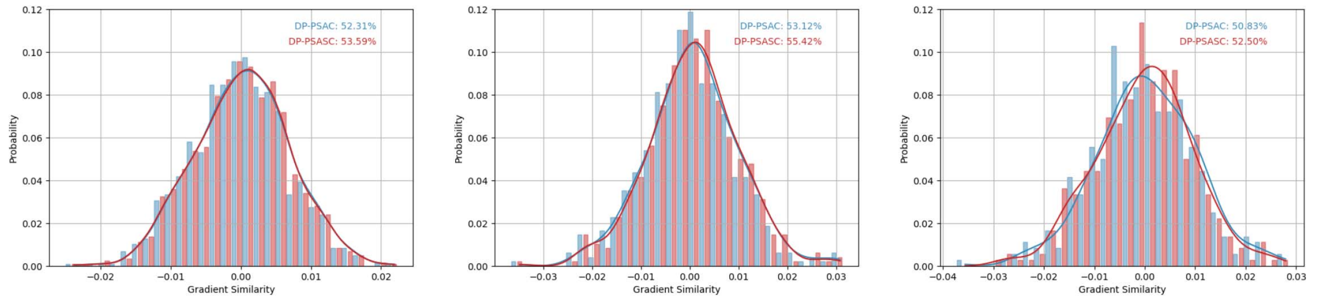
We evaluated performance on four datasets under a strict privacy budget ($\in = 8$, $\delta = 10^{-5}$). The models included a CNN for MNIST and FashionMNIST, a SimCLR-based model for CIFAR10, and a ResNet9 for Imagenette, a challenging subset of ImageNet.

Our analysis centers on gradient fidelity in the final training stages. We measure this using the cosine similarity between the private, noisy gradient and its non-private counterpart. In high-dimensional space, the direction of the update is often far more important than its exact magnitude. A high cosine similarity means the private update points in nearly the same direction as the "true" update would have, indicating that valuable learning information has been preserved despite the addition of privacy-preserving noise.

## D. Results and Discussion

The results in Table I tell a clear story: as the gradient control mechanism grows more sophisticated, model accuracy improves. The case study methods, DP-PSASC and DP-PSASC*, consistently outperform all baselines. The improvement is most striking on the challenging Imagenette dataset, where DP-PSASC* achieves an accuracy of 66.79%, a full 3 percentage points higher than standard DP-SGD. In the context of 6G, this is not just an incremental gain. For a perception model in an autonomous vehicle operating on a URLLC slice, a 3% accuracy improvement could be the difference between correctly identifying a pedestrian and causing a catastrophic failure. This is strong evidence that a principled approach to handling small gradients pays significant dividends in utility.

Fig. 2 reveals why. The plots show that the gradient similarities for DP-PSASC (red) are visibly shifted



(a) MNIST

(b) CIFAR10

(c) Imagenette

**FIG. 2.** Distribution of gradient similarities during the final training epochs. The similarity is measured between the private (noisy) and non-private gradients. A distribution shifted to the right indicates that private gradients better align with the true update direction. DP-PSASC (red) consistently maintains higher similarity, showing it better preserves crucial information for fine-tuning. This is vital for 6G applications, as higher gradient fidelity leads to faster convergence, which in turn means lower energy consumption and longer battery life for TinyML devices. It also enables higher model accuracy, increasing the reliability of critical on-device decisions.

to the right compared to its predecessor. This provides a quantitative explanation for the accuracy gains: by better preserving the directional information in small gradients, the model can make more precise adjustments during the critical fine-tuning phase. This higher “directional fidelity” helps it converge to a superior final solution. These findings have direct and important implications for the 6G and TinyML ecosystems. More stable and efficient convergence is not just an academic curiosity; it translates to tangible benefits on resource-constrained devices. For a battery-powered sensor in an mMTC network, converging in 20% fewer steps could significantly extend its operational life. For a critical application like medical diagnosis, higher accuracy under strong privacy guarantees means more reliable and trustworthy decisions.

**Computational/Energy Trade-off for TinyML.** Although the inner–outer momentum in DP-PSASC* introduces modest extra local computation and buffering, the improved gradient fidelity yields fewer communication rounds/epochs to reach a target accuracy. In resource-constrained deployments, this typically leads to lower end-to-end energy and reduced RAN signaling, consistent with recent communication-efficient FL [11].

From an architectural standpoint, these results demonstrate the viability of packaging an algorithm like DP-PSASC* as a high-performing xApp in the O-RAN Non-RT RIC. The RIC could use this xApp to orchestrate a federated learning task, setting privacy policies and managing the overall process. The efficiency of the algorithm means the xApp can achieve the target model accuracy with fewer communication rounds, reducing signaling overhead on the RAN and freeing up network resources. This case study, therefore, not only validates an algorithm but illuminates a viable path toward privacy-preserving AI that is practical, efficient, and architecturally aligned with the next generation of edge intelligence.

## V. Challenges and Future Directions

While adaptive scaling methods are a major step forward, the road to deploying robust, large-scale private learning in 6G is filled with exciting open problems. Answering them will require a multi-disciplinary approach spanning algorithms, network architecture, and hardware design.

**Integration as a Privacy xApp in O-RAN.** *Open Issue*: Although the proposed adaptive gradient methods can conceptually operate within O-RAN, the paper so far treats this only at a high level. The specific control interfaces, timing loops, and performance implications of deploying a privacy-preserving mechanism as an xApp remain insufficiently detailed. A clearer articulation is needed on what telemetry and KPIs the privacy xApp would monitor and how policy decisions would propagate through the RIC hierarchy. Without this, it is difficult to evaluate latency and communication overhead, particularly for URLLC services. Future Direction: A concrete integration pathway can be envisioned by aligning the DP-PSASC* workflow with the O-RAN control plane. In such a design, the Non-RT RIC would analyze network-wide KPIs—such as channel quality, cell load, energy consumption, and privacy budgets—and issue policy updates for gradient scaling and noise parameters through the A1 interface. The Near-RT RIC would then apply these configurations via the E2 interface to selected TinyML devices or federated clients. These control loops could operate at 10 ms–1 s intervals, maintaining feasibility for URLLC latency bounds while minimizing signaling overhead by leveraging existing A1/E2 procedures. Recent orchestration frameworks such as OrchestRAN and OREO demonstrate that such closed-loop intelligence is achievable at scale [9], [10], and GLOBECOM’24 reports a concrete privacy-preserving FL prototype for O-RAN that follows this architecture [11]. Establishing this integration experimentally would transform the proposed approach from a conceptual contribution into an implementable privacy service within the 6G ecosystem.

**The Quest for Full Automation and RIC Integration.** *Open Issue*: The tunable coefficient in DP-PSASC, while an improvement over a clipping threshold, is still a parameter that requires manual tuning. For a system managed by an O-RAN RIC, manual intervention is anathema. Future Direction: A significant leap would be to design a fully autonomous privacy xApp. This requires methods that automatically adapt the scaling function’s parameters during training without consuming additional privacy budget. A promising path lies in using reinforcement learning within the Non-RT RIC, where an agent learns an optimal schedule for the coefficient based on network KPIs (Key Performance

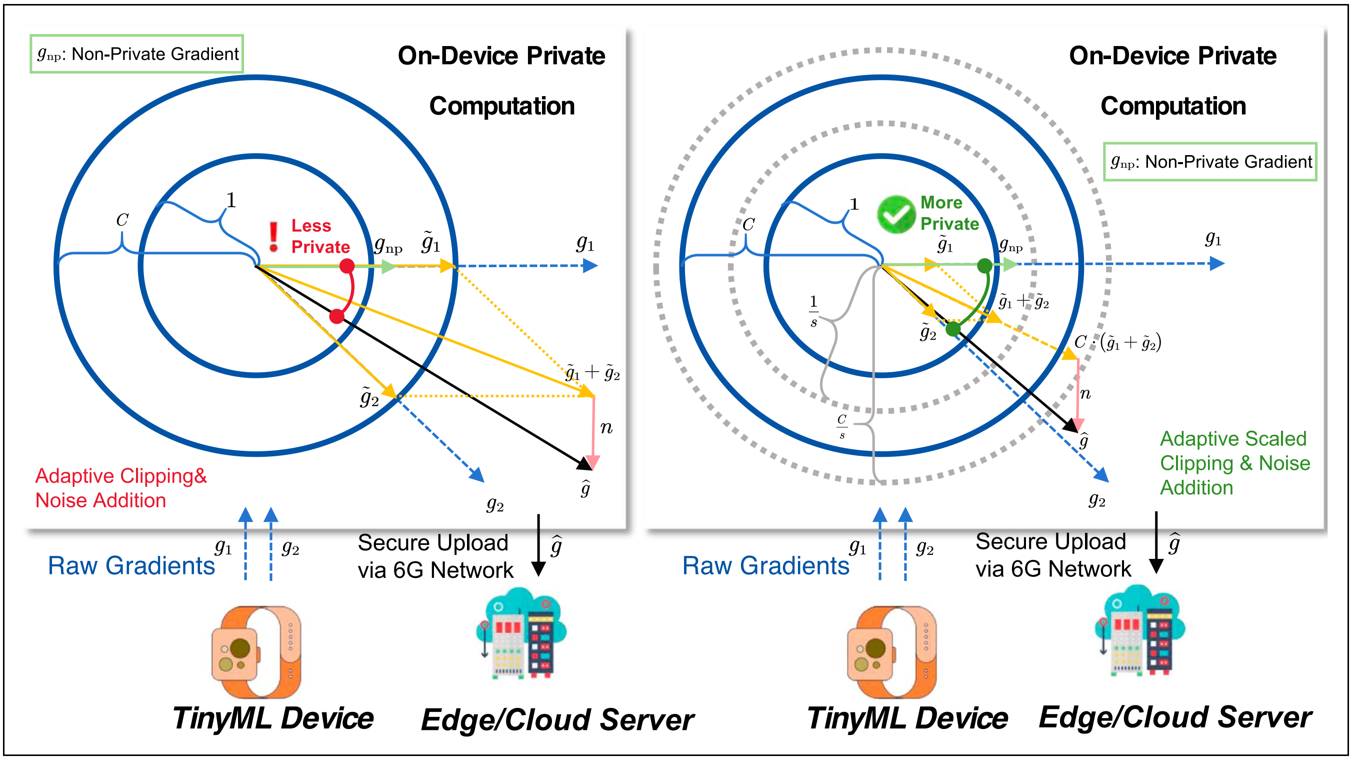

$g_{\mathrm{np}}$: Non-Private Gradient
On-Device Private Computation
$C$
1
Less Private
$g_{\mathrm{np}}$
$\tilde{g}_1$
$g_1$
$\tilde{g}_1 + \tilde{g}_2$
$\tilde{g}_2$
$n$
$\hat{g}$
$g_2$
Adaptive Clipping& Noise Addition
Raw Gradients
$g_1$
$g_2$
Secure Upload via 6G Network
$\hat{g}$
TinyML Device
Edge/Cloud Server
On-Device Private Computation
$g_{\mathrm{np}}$: Non-Private Gradient
More Private
$C$
1
$\frac{1}{s}$
$\frac{C}{s}$
$\tilde{g}_1$
$g_{\mathrm{np}}$
$g_1$
$\tilde{g}_1 + \tilde{g}_2$
$\tilde{g}_2$
$C \cdot (\tilde{g}_1 + \tilde{g}_2)$
$n$
$\hat{g}$
$g_2$
Adaptive Scaled Clipping & Noise Addition
Raw Gradients
$g_1$
$g_2$
Secure Upload via 6G Network
$\hat{g}$
TinyML Device
Edge/Cloud Server

FIG. 3. A conceptual illustration of the gradient update process within a 6G O-RAN architecture. Raw gradients are computed on-device by 6G-enabled UEs (e.g., TinyML devices). the on-device privacy-preserving computation (clipping/scaling and noise addition) is governed by policies from the network's control plane. The processed gradients are then securely uploaded via the 6G RAN to an edge server, which can be an O-Cloud node. The figure contrasts two policies that could be enforced by a privacy xApp in the RIC: a less effective method (left, e.g., DP-PSAC) and the more advanced scaling approach (right, e.g., DP-PSASC). for DP-PSASC, gradients are first scaled to a uniform norm, summed, and then stretched before noise addition. As shown, this can result in a final update that better preserves privacy (larger angle to the non-private gradient) while achieving higher utility, demonstrating a superior privacy-utility trade-off suitable for 6G.

Indicators) like convergence speed, device battery levels, and radio resource usage. This would enable a zero-touch, self-optimizing privacy service for the entire network.

**The Gap in Theoretical Understanding.** *Open Issue*: The empirical success of non-monotonic scaling functions is clear, but our theoretical understanding of their behavior, especially in the highly non-convex world of deep learning, lags behind. This gap becomes a critical liability when deploying AI for mission-critical 6G services. Future Direction: We need a unified convergence theory that explains precisely how different gradient manipulations affect the optimization path under the influence of network dynamics (e.g., latency, packet loss). A more robust theory, perhaps one grounded in stochastic differential equations [12], could provide formal guarantees on the reliability and fairness of AI models trained over 6G, a prerequisite for their use in applications like autonomous transportation and public safety.

**The Challenge of Scale**. *Open Issue*: The elephant in the room for all the methods discussed is their computational cost. The need to compute per-sample gradients is prohibitively expensive for today's massive models and for the sheer scale of 6G, which envisions trillions of connected devices. Future Direction: Research is urgently needed into more memory- and compute-efficient variations that can scale. Within the O-RAN context, this could involve exploring hierarchical or group-wise scaling, where an xApp directs different privacy policies to different slices of the network or groups of devices. Another avenue is developing new low-rank approximation techniques [13] that can estimate gradient norms without ever materializing the full vectors, drastically reducing on-device computation. This computational barrier is recognized as a primary obstacle, with recent theoretical analyses highlighting the fundamental trade-offs between privacy, utility, and scalability for massive models [14].

**Broader Applications and Composition in O-RAN**. *Open Issue*: The principles of adaptive scaling are general, but their application in other privacy-preserving paradigms and their interaction with other network functions are still nascent. Future Direction: A natural and exciting frontier is co-designing privacy xApps with other RIC applications. For example, a privacy xApp in the Non-RT RIC could provide long-term budget information to a traffic scheduling xApp in the Near-RT RIC. The scheduling xApp could then prioritize radio resources for users who are close to exhausting their privacy budget, creating a truly "privacy-aware" network fabric. This synergy is a crucial step for practical deployment in 6G-enabled Federated Learning (FL) systems [15]. Further analysis is also needed on how privacy budgets compose over many rounds and interact with advanced privacy accountants [16], a task perfectly suited for a centralized rApp in the Non-RT RIC. We also attempt to situate DP with complementary paradigms (e.g., secure aggregation, unlearning) to explore how on-device adaptive DP can be composed with other paradigms [17].

**Hardware Co-Design for TinyML**. *Open Issue*: Ultimately, for TinyML, the algorithm must be efficient on the silicon itself. Software optimizations alone are insufficient to meet the stringent energy constraints of many 6G devices. Future Direction: This opens the door for algorithm-hardware co-design.

Future work could explore simplified, fixed-point arithmetic versions of scaling functions that run efficiently on microcontrollers. Taking this further, one can envision future "privacy co-processors" or dedicated hardware blocks within 6G-native chipsets. These accelerators would have dedicated instructions for privacy-preserving operations, making on-device private learning a widespread, energy-efficient reality. This vision is beginning to take shape, with emerging research exploring specific architectural designs for privacy-preserving hardware accelerators [18].

## VI. Conclusion

In this paper, we have charted the evolution of gradient control mechanisms in differentially private learning, framing this journey within the context of the emerging 6G landscape. We have argued that the fundamental trade-off between model utility and data privacy must be resolved not in an algorithmic vacuum, but within the architectural realities of next-generation wireless systems. We followed the path from the rigid, foundational paradigm of static clipping to the flexible and powerful approach of adaptive scaling, highlighting how each step moves us closer to a solution viable for the scale, diversity, and dynamism of 6G.

Our central argument is that the key to this challenge lies in intelligent, adaptive management of gradient sensitivity. Through our framework and case study of DP-PSASC, we showed that a principled, non-monotonic scaling function can lead to state-of-the-art performance. The results confirm that by carefully weighting gradients of all sizes, it is possible to preserve crucial directional information for optimization while upholding rigorous privacy guarantees. More importantly, we have positioned these advanced algorithms as foundational technologies for new services within future network architectures. By envisioning their deployment as xApps within an O-RAN framework, we provide a concrete roadmap for integrating privacy as a native, manageable, and optimizable function of the 6G network itself. This work sets the stage for future innovations in creating the autonomous, efficient, and trustworthy intelligence that the 6G and TinyML era demands.

## References

[1] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 14774–14784.

[2] M. Abadi et al., "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 308–318.

[3] H. Xiao, Z. Xiang, D. Wang, and S. Devadas, "A theory to instruct differentially-private learning via clipping bias reduction," in *Proc. IEEE Symp. Secur. Privacy (SP).*, Piscataway, NJ, USA: IEEE Press, 2023, pp. 2170–2189.

[4] G. Andrew, O. Thakkar, B. McMahan, and S. Ramaswamy, "Differentially private learning with adaptive clipping," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 17455–17466.

[5] C. Wei, W. Li, C. Gong, and W. Chen, DC-SGD: Differentially private SGD with dynamic clipping through gradient norm distribution estimation. *IEEE Trans. Inf. Forensics Secur.*, pp. 4498–4511, 2025.

[6] Z. Bu, Y.-X. Wang, S. Zha, and G. Karypis, "Automatic clipping: Differentially private deep learning made easier and stronger," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 41727–41764.

[7] F. Hübler, I. Fatkhullin and N. He, "From gradient clipping to normalization for heavy tailed SGD," in *Proc. 28th Int. Conf. Artif. Intell. and Stat., PMLR*, 2025, vol. 258, pp. 2413–2421.

[8] T. Xia et al., "Differentially private learning with per-sample adaptive clipping," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 10444–10452.

[9] S. D'Oro, L. Bonati, M. Polese, and T. Melodia, "OrchestRAN: Orchestrating network intelligence in the open RAN," *IEEE Trans. Mobile Comput.*, vol. 23, no. 7, pp. 7952–7968, Jul. 2024. [Online]. Available: https://ece.northeastern.edu/wineslab/papers/doro2024orchestran.pdf

[10] F. Mungari, C. Puligheddu, A. Garcia-Saavedra, and C. F. Chiasserini, "OREO: O-RAN intelligence orchestration of xApp-based network services," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Piscataway, NJ, USA: IEEE Press, 2024, pp. 71–80.

[11] S. Wijethilaka and A. K. Yadav, "Privacy-preserving federated learning framework for open radio access networks (ORAN)," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Cape Town, South Africa, 2024, pp. 4515–4520.

[12] R. Zhang, M. Lei, M. Ding, Z. Xiang, J. Xu, and D. Wang, "Improved rates of differentially private nonconvex-strongly-concave minimax optimization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 21, 2025, pp. 22524–22532.

[13] D. Yu, H. Zhang, W. Chen, J. Yin, and T.-Y. Liu, "Large scale private learning via low-rank reparametrization," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 12208–12218.

[14] S., Siddharth et al., "Democratizing AI: Opensource scalable LLM training on GPU-based supercomputers," in *Proc. Int. Conf. High Perfor. Comp., Netw., Stor. and Anal.*, 2024, pp. 1–14.

[15] L. Chen, J. Wang, and W. Zhang, "Communication-efficient and adaptively private federated learning for 6G IoT networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Piscataway, NJ, USA: IEEE Press, 2024, pp. 1–6.

[16] J. Li, Y.-W. Wu, and M. Chen, "FL-PR: A novel privacy-preserving scheme for federated learning in 6G-enabled IoT," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Piscataway, NJ, USA: IEEE Press, 2024, pp. 1–6.

[17] F. Wang, B. Li, and B. Li, "Federated unlearning and its privacy threats," *IEEE Netw.*, vol. 38, no. 2, pp. 294–300, Mar. 2024.

[18] S. S. Saha, S. S. Sandha and M. Srivastava, "Machine learning for microcontroller-class hardware: A review," *IEEE Sensors J.*, vol. 22, no. 22, pp. 21362–21390.