Journal of Intelligent Computing and Networking
https://www.ffspub.com/index.php/jicn/index
ISSN 3079-9228 (print)
ISSN 3107-2798 (online)
E-mail: jicn.office@ffspub.com

Review

When One-Shot Federated Learning Meets Diffusion Models at the Edge: Technological Advances and Applications

Wanxiang Chen[1,‡], Dongshang Deng[1,‡], Chaocan Xiang[1,†], Zhi Liu[2], Bin Xiao[3]
[1]*College of Computer Science, Chongqing University, Chongqing 400044, China*
[2]*Department of Computer and Network Engineering, The University of Electro-Communications, Tokyo 1828585, Japan*
[3]*School of Artificial Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China*
[*]*E-mail: xiangchaocan@cqu.edu.cn*

Received: December 14, 2025 / Revised: March 18, 2026 / Accepted: March 22, 2026 / Published online: March 26, 2026

Abstract: The rapid growth of the Internet of Things (IoT) and edge devices has accelerated the adoption of edge computing, which processes data at the edge to reduce latency and enhance privacy. In this context, federated learning (FL) has emerged as a promising framework for distributed model training without sharing raw data. However, traditional FL methods are often impractical in edge scenarios due to their reliance on extensive, resource-intensive communication rounds. To tackle this issue, one-shot federated learning (OSFL) has been proposed, enabling model aggregation in a single communication round. Meanwhile, diffusion models have gained significant attention for their powerful generative capabilities, especially in image synthesis and data augmentation. Recently, researchers have begun exploring the implementation of OSFL with diffusion models. By leveraging diffusion-based data generation, these approaches efficiently combine knowledge from distributed sources. This synergy not only improves model performance under non-IID data but also addresses the challenges related to data scarcity and privacy in edge environments. In this review, we systematically analyze the intersection of these two advanced paradigms, highlighting their complementarity and discussing key design considerations. Furthermore, we outline the contributions of this work: (1) providing a comprehensive taxonomy of existing approaches that combine OSFL with diffusion models; (2) identifying open challenges and future research directions; and (3) offering practical insights for deploying such integrated systems in real-world edge computing applications.

Keywords: Edge Computing; One-Shot Federated Learning; Diffusion Model; Privacy-Preserving Learning
https://doi.org/10.64509/jicn.21.64

1 Introduction

The rapid growth of the Internet of Things (IoT) devices and mobile applications has made edge computing a crucial approach for processing data directly on resource-constrained devices [1, 2]. This method helps to reduce latency [3, 4] and preserve bandwidth [5]. However, training machine learning models using decentralized edge data raises significant privacy concerns and communication overheads. This has led to the emergence of federated learning (FL) [6], which allows for collaborative model training without sharing raw data. Despite its advantages, traditional FL requires multiple communication rounds, incurring high costs in unstable or bandwidth-limited edge environments [7, 8]. However, the increasing scale and mobility of edge devices further amplify these communication and reliability constraints, making multi-round coordination increasingly impractical in real deployments. To address this issue, one-shot federated learning (OSFL) [9] has been introduced. OSFL enables effective global model aggregation within a single communication round by exchanging either compact model updates [10] or synthetic data representations [11]. Nevertheless, achieving efficient OSFL faces numerous challenges. These include severe resource constraints on edge devices, a significant performance drop under non-IID data distributions, and increased privacy risks arising from the exchange of compact model

† Corresponding author: Chaocan Xiang
‡ These authors contributed equally to this work
[*]Academic Editor: Xiuli Bi

Table 1: A comparison on the existing surveys of FL and OSFL, highlighting their scope, limitations, and the unique perspective of this survey.

Survey	Summary	Tax.[1]	App.[2]	OSFL for DMs	DMs for OSFL	Limitation
Liu *et al.* [12]	Comprehensive taxonomy of OSFL, addressing data/model heterogeneity and practical deployment challenges in resource-constrained settings.	✓	✓	✗	✓	Lacks discussion on generative diffusion architectures for OSFL.
Amato *et al.* [13]	Systematic review of OSFL evolution, including communication-efficient aggregation and non-IID mitigation in real-world edge applications.	✓	✓	✗	✗	Reviews OSFL evolution but lacks discussion of generative paradigms.
Ayeelyan *et al.* [14]	Broad overview of FL aggregation models, including early OSFL variants like one-round knowledge sharing.	✓	✓	✗	✗	Neglects the bidirectional integration between OSFL and diffusion models.
Gargary *et al.* [15]	In-depth analysis of federated GANs, VAEs, and DMs for privacy-preserving generation; covers DM-OSFL hybrids but focuses on multi-round FL	✓	✗	✓	✗	Confined to multi-round FL; does not address OSFL scenarios.
Our paper	We provide a comprehensive survey of OSFL and develop a novel direction for it, namely OSFL for DMs and DMs for OSFL.	✓	✓	✓	✓	The first systematic survey of the bidirectional integration between OSFL and diffusion models.

[1]Taxonomy. [2]Application.

updates or synthetic proxies within a single communication round.

Diffusion models (DMs) [16] offer a promising solution by enabling high-fidelity data synthesis through iterative denoising in latent space. This enables clients to transmit only lightweight embeddings, such as noisy latents [16] and conditioning prompts [17], while the server reconstructs diverse, distribution-aligned samples. In edge computing, edge-adapted diffusion models leverage techniques like LoRA-based parameter-efficient fine-tuning(PEFT), latent diffusion architectures [17], and classifier-free guidance. These approaches facilitate lightweight client-side personalization [18] and efficient server-side generation [19, 20], significantly enhancing OSFL's robustness, privacy, and scalability in heterogeneous, communication-constrained environments. Therefore, DMs represent a promising candidate as a key vehicle for enabling efficient OSFL in the future.

The convergence of OSFL and DMs represents a transformative synergy in privacy-preserving generative learning. OSFL empowers DMs by enabling federated training of high-capacity generative models in a single communication round [21]. Clients upload only compact latent representations or conditional embeddings, allowing the server to aggregate a global denoising process without exchanging raw data or making iterative updates. Conversely, DMs supercharge OSFL by providing a robust mechanism for synthetic data generation and distribution alignment: clients transmit lightweight prompts [11] or statistics [22], while the server leverages pre-trained or collaboratively refined DMs to synthesize high-fidelity, domain-specific samples that mitigate non-IID challenges and enhance global model accuracy. This bidirectional enhancement enables efficient, scalable, and privacy-resilient generative modeling across heterogeneous edge networks.

Building on the synergy between OSFL and DMs, this paradigm holds transformative potential across diverse real-world domains. In healthcare, OSFL enables hospitals to collaboratively train personalized diagnostic models using DM-generated synthetic medical images (*e.g.,* MRIs or pathology slides) from latent embeddings shared in a single round. This approach preserves patient privacy while overcoming data silos and non-IID distributions across institutions [8, 23]. In battery swapping networks, electric vehicle stations employ OSFL with DMs to model spatiotemporal battery demand patterns: edge stations upload compact usage statistics, and the central server uses DMs to synthesize realistic demand forecasts [24]. This optimizes inventory allocation without exposing proprietary operational data [25]. In mobile crowdsensing, OSFL-DM frameworks facilitate collaborative environment modeling; drones share lightweight visual or sensor embeddings once [26], enabling a global DM to generate unified 3D terrain or obstacle maps for safe navigation, all while minimizing communication in bandwidth-constrained aerial networks [27, 28]. These applications underscore the scalability, privacy, and adaptability of OSFL-DM integration in mission-critical, data-heterogeneous systems.

As summarized in Table 1, while prior surveys have extensively reviewed either OSFL techniques [12, 13] or the application of generative models (including DMs) in multi-round FL [14, 15], none has systematically addressed the bidirectional integration of OSFL and DMs. Existing works either treat DMs solely as tools within conventional multi-round FL settings or analyze OSFL frameworks without exploring

generative enhancement through modern diffusion architectures. In contrast, our paper is the *first* comprehensive survey that explicitly focuses on the emerging paradigm of OSFL for DMs and DMs for OSFL. This work establishes a unified taxonomy, identifies unique challenges (*e.g.,* latent representation design, single-round distribution alignment, and edge-compatible diffusion), and presents novel insights for both directions. These contributions collectively fill a critical gap in the literature and outline a clear roadmap for achieving communication-efficient and privacy-preserving generative FL in heterogeneous edge environments.

In this article, we provide a comprehensive survey of OSFL and propose a novel direction for exploring the relationship between OSFL and DMs. The key contribution of this article can be summarized as follows.

- We conduct an in-depth analysis of the unique challenges and design principles when combining OSFL and DMs in edge environments, including latent representation design, single-round distribution alignment, privacy-utility trade-offs under differential privacy, and resource-efficient diffusion adaptation.
- To bridge this critical gap and systematically organize this rapidly evolving field, we present the first systematic and bidirectional survey on the integration of OSFL and DMs. This survey explicitly categorizes the field into two complementary yet underexplored research directions, *i.e.,* OSFL for DMs and DMs for OSFL.
- We investigate representative real-world applications in healthcare, battery-swapping networks, and mobile crowdsensing, showcasing the practical feasibility and superiority of the OSFL-DM paradigm in mission-critical, data-siloed scenarios. Moreover, we identify the existing challenges of OSFL, including data heterogeneity and privacy constraints, and propose forward-looking research directions to foster deep integration between OSFL and DMs.

The remainder of this paper is structured as follows. Section 2 introduces edge computing, federated learning, and one-shot federated learning. Section 3 reviews diffusion models in edge computing scenarios. Subsequently, Section 4 presents the interplay between diffusion models and one-shot federated learning. Section 5 discusses current applications. Finally, we summarize existing challenges and future research directions in Section 6.

2 OSFL in Edge Computing

The design philosophy of OSFL aligns well with edge computing scenarios, making the development of a practical OSFL framework for edge computing a promising research direction. This section begins by discussing the conceptual applications of edge computing and its existing challenges (Section 2.1), thereby introducing FL (Section 2.2) and OSFL (Section 2.3). Finally, we present the lessons learned (Section 2.4).

2.1 Edge Computing

Edge computing fundamentally shifts the cloud-centric paradigm to the network edge—devices and nodes near data generation sources [1]. This transition enables millisecond-level latency [3], efficient bandwidth usage [5, 29], and localized data processing [30]. Its adoption is propelled by demands from 5G, IoT, and real-time applications, such as distributed healthcare applications [31], real-time industrial energy systems [24, 32], and obstacle detection for autonomous vehicles [33]. The edge infrastructure, spanning from low-power sensors to base station-mounted edge computing servers, creates a distributed computing system. By processing data locally instead of routing it to distant clouds, this model slashes end-to-end latency from hundreds of milliseconds to under a few milliseconds, while offloading backbone network traffic [1]. It also inherently supports the data privacy requirements of regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) by minimizing the transmission of raw data. However, achieving an efficient edge computing framework remains challenging.

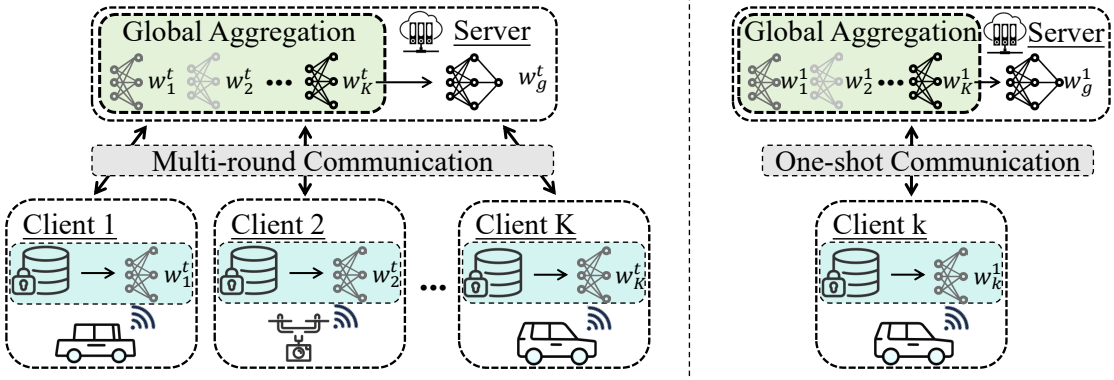
Edge computing environments are significantly more complex than homogeneous cloud clusters, constrained by three core challenges: resource heterogeneity [7, 8], network dynamics [34], and energy sensitivity [35]. Device capabilities span from milliwatt-level MCUs to hundred-watt-level GPUs, with memory ranging from kilobytes to gigabytes. Network conditions are volatile, with links switching between Wi-Fi, 4G/5G, and LoRa, and are prone to severe packet loss and bandwidth fluctuations [36]. Furthermore, the energy constraints of battery-powered or energy-harvesting nodes make communication—often consuming over 60% of total power—a critical bottleneck [35]. Frequent data exchange can rapidly deplete batteries, causing nodes to fail. This environment also exacerbates data silos, where device data is highly non-independent and identically distributed (non-IID) due to divergent user behaviors, local environments, and sensor accuracies [37, 38].

2.2 Federated Learning

Federated learning [6], introduced by Google in 2017, operates on the principle of "moving the model, not the data". Edge devices train locally on private data and transmit only model updates (*e.g.,* gradients or parameters) to a central server for aggregation, enabling collaborative learning without sharing raw data. The widely-used FedAvg [6] algorithm iteratively refines the global model through multiple rounds. As shown in Figure 1(a), the server distributes the current model, clients perform local SGD and submit updates, and the server aggregates them using a data-volume-weighted average before broadcasting an aggregated model.

Considering a system of K clients and one server collaborating to train a global model, where each client k possesses a local dataset $D_k \in \mathscr{D}_k$, where $\mathscr{D}_k$ is a local data distribution, the global optimization objective, without compromising client privacy, is formulated as follows:

$$\min_{w_g} \mathscr{L}(w_g) = \min_{w_g} p_k \mathscr{L}_k(w_g), \tag{1}$$



(a) Federated Learning (b) One-shot Federated Learning

Figure 1: Comparisons between the FL and OSFL frameworks. (a) Federated learning (FL): Multiple communication rounds are required between the server and clients. (b) One-shot Federated Learning : Only a single communication round is required.

where $\mathscr{L}(\cdot)$ and $\mathscr{L}_k(\cdot)$ are the global and local loss functions, respectively; p_k is the aggregation weight satisfying $\sum_{k=1}^{K} p_k = 1$. In FedAvg, p_k is set proportional to the local data volume. This method has proven highly effective in harnessing idle computational resources across millions of edge devices. Privacy protection is enhanced through techniques such as homomorphic encryption [39, 40] and differential privacy [41], which introduce noise into model updates to mitigate model inversion attacks.

While FL excels in privacy preservation, its multi-round communication protocol poses a significant performance bottleneck in edge environments [8]. Constrained by limited and unstable bandwidth, transmitting multi-megabyte model parameters per round can incur communication latency of seconds to minutes, resulting in a linear increase in overhead with rounds ($O(T \times |w_g|)$), where T is the number of communication rounds, and $|w_g|$ denotes model parameter size. To reduce communication overhead, existing methods can be categorized into model quantization, sparsification, and fewer communication rounds, as summarized in Table 2. These approaches aim to reduce the communication overhead, either in each round or in the total number of rounds.

- **Model quantization:** Model quantization reduces the data volume of each communication transmission by lowering the numerical precision of the updates [42–45]. Qu *et al.* [44] proposed integrating quantization with clipped SGD to mitigate the quantization error. To address the computational overhead and convergence issues introduced by quantization, Cui *et al.* [45] proposed an adaptive selection of quantization levels based on data quality and communication capability. However, by reducing the bit-precision of model updates, model quantization inevitably leads to a slowdown in the convergence of FL.
- **Model sparsification:** Model sparsification involves selecting only a subset of the most important parameters for transmission, while ignoring those deemed to have a minor impact on the model update [46–49]. adaMC [48] proposed a joint optimization strategy for communication and computation that effectively improves the energy efficiency and scalability of FL in resource-constrained environments. Wei *et al.* [49] proposed a gradient sparsification framework for FL over wireless channels to enhance training efficiency without sacrificing convergence performance. Similar to model quantization, model sparsification also discards a portion of the update information, leading to a delay in model convergence.
- **Faster convergence:** Model quantization and sparsification aim to reduce the communication overhead $|w_g|$ per round. Without compromising the model updates, these methods aim to reduce the number of communication rounds T required for model convergence [7, 8, 50, 51]. To address the degradation in convergence efficiency caused by aggregation loss, Deng *et al.* [8] proposed an adaptive personalized model calibration method. The key idea of AdaPC [51] is to adjust the aggregation period adaptively, enabling the federated learning framework to achieve fast convergence with minimal communication. However, a critical problem persists: is it possible to reduce the required communication rounds to just one?

While quantization and sparsification reduce per-round overhead, they often compromise model integrity; achieving convergence in a single communication round remains the theoretical upper bound for efficiency in unstable edge environments.

2.3 One-Shot Federated Learning

One-shot federated learning pushes the communication rounds of FL to the extreme—as shown in Figure 1(b), clients perform only one round of local training and upload the model parameters [10], soft labels [11], or feature representations [22]. This reduces the communication overhead from $O(T \times |w_g|)$ in conventional FL to just $O(|w_g|)$. The core assumption of this paradigm is that, through well-designed local training strategies and server-side fusion mechanisms, even a single round of information exchange is sufficient to capture the common knowledge pattern present across distributed data. Note that the OSFL setting typically operates under several practical assumptions. Primarily, due to strict communication constraints in edge environments, clients participate in only a single communication round without iterative feedback. Meanwhile, local data distributions across clients are typically heterogeneous, which motivates collaborative learning without centralized data sharing. Furthermore, privacy considerations prohibit the exchange of raw local data, and

Table 2: A summary of communication efficiency FL includes model quantization, sparsification, and faster convergence.

Category	Basic idea	Method	Weakness
Model quantization	By reducing the bit precision of the model in each communication round, it reduces the overall communication overhead.	QAFeL [42] FLoRA [43] FedQClip [44] LCO-AGQ [45]	Reducing the bit precision inevitably degrades local model performance, which in turn leads to a delay in convergence.
Model sparsification	Transmitting only the most critical parameters in each round lowers the total communication overhead.	adaMC [46] SparsiFL [47] ANC [48] [49]	The sparsification process can discard important information from model parameters, resulting in convergence delay.
Faster convergence	Reducing the number of rounds required for model convergence, thereby achieving faster convergence and lowering the total communication overhead.	FedASA [7] pFedCal [8] FedGau [50] AdaPC [51]	The theoretically optimal scenario is to achieve model convergence with only a single round of client-to-server communication.

only model parameters, feature representations, or distilled knowledge are exchanged.

According to the uploading medium, existing approaches can be broadly categorized into three main paths: data-based, model-based, and feature-based, as summarized in Table 3.

- **Data-based approaches:** These methods propose a one-shot upload of partial client data [9] or its distilled version [52] to the server for global training. The underlying assumption is that a minimal data subset can preserve the essential information for model aggregation. For instance, DOSFL [52] applies knowledge distillation to create a compact, noise-like dataset from each client's private data, which is designed to be functionally meaningful only to the specific global model being trained. However, uploading any form of client data to the server, even in a distilled format, not only introduces significant privacy risks but also fundamentally contradicts the privacy-by-design paradigm of FL. This category represents the most primitive form of OSFL implementation.
- **Model-based approaches:** Instead of uploading private data, model-based methods generate data by leveraging the model parameters sent by clients. It was theoretically demonstrated by Yang *et al.* [53] that multi-round client models can guide diffusion models for synthetic data generation. DENSE [10] and FedHydra [54] utilize the uploaded client model parameters to simultaneously optimize an auxiliary generator and the global model. However, using model parameters, which are learned from client data, to guide data generation inevitably introduces additional data-to-model errors.
- **Feature-based approaches:** Similar to data-based methods, feature-based methods similarly aim to build a public dataset on the server. Feature-based methods fully leverage the capabilities of pre-trained generative models to synthesize data based on the features extracted from client data. A representative example of this straightforward feature type is the label description, as employed in [11, 55]. Additionally, FedBiP [22] proposed the use of personalized features to bridge the distribution gap between real and synthetic data. Although these methods are highly promising, they encounter challenges including the real-to-synthetic distribution shift as well as data heterogeneity.

Overall, OSFL significantly reduces communication overhead by limiting training to a single communication round. However, its effectiveness largely depends on whether the adopted paradigm is well-suited to the communication constraints, data heterogeneity, and privacy requirements of edge environments.

2.4 Lessons

Edge devices deployed at the network edge are continuously generating vast amounts of data. Limited communication bandwidth and client privacy concerns hinder the transmission of this data to a central server. Although FL enables the training of a global model by collaborating with multiple clients without raising privacy concerns, its multi-round client-server communication is not well-suited for heterogeneous edge environments. Existing studies reduce communication overhead through model quantization, sparsification, and fewer communication rounds. However, both quantization and sparsification compromise local model knowledge, leading to convergence delay. While reducing the number of rounds is a promising direction, it prompts a fundamental question: is it possible, ideally, to achieve model convergence with just a single round of communication?

OSFL has emerged as a solution that not only inherits the privacy-preserving features of FL but also requires only a single round of client-server communication. The early OSFL method, being data-based, raised privacy concerns as it required uploading a portion of client data, even after distillation. Later studies introduced model-based and feature-based methods, which rely on a pre-trained model at the server for data generation to facilitate model training. While promising, the feature-based approach faces key challenges: obtaining a suitable pre-trained model, achieving personalization, and alleviating the real-to-synthetic distribution shift. Addressing these open challenges represents a key direction for future research in OSFL.

3 DMs in Edge Computing

Diffusion Models have emerged as a powerful new class of generative models, garnering significant attention from both academia and industry. This section begins by reviewing the evolution (Section 3.1) and fundamental mechanisms

Table 3: A summary of current OSFL methods includes data-based, feature-based, and model-based approaches.

Category	Basic idea	Method	Publication	Auxiliary dataset	Pre-trained model
Data-based	One-shot upload partial data or distilled data to the server for global model training.	[9]	Arxiv '19	Yes	No
		DOSFL [52]	Arxiv '21	Yes	No
Model-based	One-shot upload local model parameters as prompts guides the pre-trained model's data generation process.	DENSE [10]	NeurIPS '22	No	No
		FedLMG [53]	ICML '25	No	Yes
		FedHydra [54]	KDD '25	No	No
Feature-based	One-shot upload local extracted features to guide the pre-trained model's data generation process.	FGL [55]	Arxiv '23	No	Yes
		FedDEO [11]	MM '24	No	Yes
		FedBiP [22]	CVPR '25	No	Yes

(Section 3.2) of DMs. Then we analyze their emerging applications in edge computing (Section 3.3) and propose a critical taxonomy of lightweighting strategies (Section 3.4). Finally, we summarize the lessons learned (Section 3.5).

3.1 The Evolution of Diffusion Models

The development of DMs represents a significant milestone in generative modeling. Unlike traditional generative modeling paradigms—such as Variational Autoencoders (VAEs) [56] and Generative Adversarial Networks (GANs) [57]—DMs learn data distributions through a forward progress and a reverse denoising process, achieving a superior balance between training stability and generation quality. Specifically, DMs consist of two coupled stages: a forward process, which gradually perturbs data into an isotropic Gaussian prior, and a reverse process, which reconstructs structured samples from noise via neural approximation of ordinary or stochastic differential equations (ODE/SDE) dynamics [58, 59]. This framework provides a stable, interpretable objective and avoids the adversarial instability of GANs while outperforming VAEs and energy-based models in fidelity and diversity [60].

Early research, inspired by non-equilibrium thermodynamics [61], led to the Denoising Diffusion Probabilistic Model (DDPM) [16]. This model restructured the variational objective to predict noise directly, enabling stable training and achieving visual quality comparable to that of GANs. Building on this foundation, subsequent advances such as the Denoising Diffusion Implicit Model (DDIM) [62] and score-based SDE formulations [59], significantly reduced sampling steps while preserving high fidelity. Later, the Latent Diffusion Model (LDM) [17] further improved efficiency by performing diffusion in a compressed latent space, reducing computational cost by orders of magnitude and enabling large-scale systems such as Stable Diffusion. Nowadays, DMs have been widely applied to generation tasks, including image [63], audio [64], video [65], and molecular generation, demonstrating remarkable generalization and distribution alignment. These capabilities have further highlighted their potential in decentralized learning scenarios. Their robustness and adaptability make them particularly suitable for privacy-sensitive and heterogeneous edge environments, providing a solid foundation for integration with OSFL [66].

3.2 Fundamental Mechanism of Diffusion Models

Diffusion Models are a class of probabilistic generative models that learn to reverse a gradual noising process to recover the underlying data distribution and enable high-quality sample generation [16]. As illustrated in Figure 2, this mechanism consists of two opposing processes: a fixed forward diffusion process, in which data are progressively corrupted with Gaussian noise according to a predefined variance schedule, and a learnable reverse denoising process.

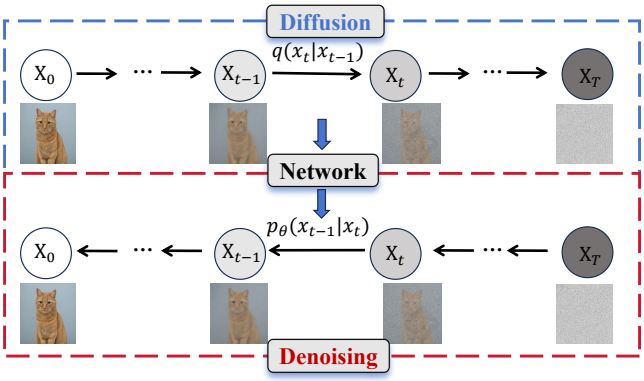


Figure 2: Illustration of the forward and reverse diffusion process in diffusion models. The forward diffusion process progressively adds Gaussian noise to the data until it becomes pure noise (top blue-dashed box), while the reverse process learns to reconstruct the original data distribution from noise (bottom red-dashed box).

3.2.1 Forward Process

The forward diffusion process q is a fixed (non-learnable) Markov process that progressively corrupts a real data sample x_0 with Gaussian noise over T timesteps, producing a sequence of latent variables $\{x_1, x_2, \ldots, x_T\}$. Formally, given the initial sample x_0, the joint distribution of this Markov chain factorizes as

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}), \tag{2}$$

where each transition $q(x_t \mid x_{t-1})$ is modeled as a Gaussian distribution that injects scaled noise at step t,

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(\sqrt{1-\beta_t}\, x_{t-1}, \beta_t \mathbf{I}\right), \tag{3}$$

where $\beta_t \in (0,1)$ denotes the predefined noise variance schedule, and $\mathbf{I}$ is the identity matrix. As the number of timesteps T becomes sufficiently large, the forward process drives the final latent variable toward an isotropic Gaussian distribution, *i.e.*,, $x_T \sim \mathcal{N}(0,\mathbf{I})$.

A critical property of this Markovian process is that we can sample x_t at an arbitrary timestep t directly from x_0 in a closed form. By defining $\alpha_t = 1-\beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, we derive a closed-form expression that avoids iterative sampling and thus enables efficient training:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, \quad \varepsilon \sim \mathcal{N}(0,I). \tag{4}$$

3.2.2 Reverse Process

In essence, DMs aim to learn a parameterized reverse denoising process, p_θ, that iteratively recovers the data distribution, starting from the prior noise $p(x_T) = \mathcal{N}(0,\mathbf{I})$. The reverse process is also a Markov chain:

$$p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t), \tag{5}$$

where the transition probability $p_\theta(x_{t-1}|x_t)$ is modeled as a Gaussian distribution, whose parameters are predicted by a neural network θ (typically a U-Net):

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\big(\mu_\theta(x_t,t), \Sigma_\theta(x_t,t)\big), \tag{6}$$

where $\mu_\theta(x_t,t)$ and $\Sigma_\theta(x_t,t)$ are the mean and variance of the Gaussian distribution, respectively.

3.2.3 Training Objective

Theoretically, training DMs requires maximizing the data log-likelihood, which often involves optimizing a complex evidence lower bound (ELBO). However, Ho *et al.* [16] achieved a revolutionary breakthrough by greatly simplifying the training objective.They found that by reparameterizing the mean μ_θ and fixing the variance Σ_θ (as a constant related to β_t), the neural network $\varepsilon_\theta(x_t,t)$ can be trained to directly predict the noise ε ($\varepsilon \sim \mathcal{N}(0,\mathbf{I})$) added at timestep t.

Consequently, the training objective is simplified to a Mean Squared Error (MSE) loss function, minimizing the difference between the predicted noise and the true noise:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,x_0,\varepsilon}\left[\|\varepsilon - \varepsilon_\theta(x_t,t)\|^2\right], \tag{7}$$

where t is uniformly sampled from $\{1,\ldots,T\}$ and $\varepsilon_\theta(x_t,t)$ denotes the model's approximation of the true noise ε. This formulation reframes the complex probabilistic modeling problem into an intuitive, efficient denoising auto-encoding task, making the training process of DMs exceptionally stable and straightforward to implement.

3.2.4 Extensions and Summary

Furthermore, DMs can be extended to conditional diffusion models [67, 68], where the denoise network is conditioned on an auxiliary variable y (e.g., class labels, textual prompts, or multimodal features), yielding $\varepsilon_\theta(x_t,y,t)$ and enabling semantically controllable generation [17]. During sampling, DMs start from Gaussian noise $x_T \sim \mathcal{N}(0,I)$ and progressively denoise it through the learned reverse diffusion process to obtain the final sample x_0. Although the original DDPM typically requires thousands of sampling steps and thus suffers from slow inference, subsequent variants such as DDIM and LDM substantially improve sampling efficiency [17, 62].

Overall, DMs explicitly model the noise distribution and reverse process, achieving high-quality generation with stable optimization. Their interpretable training objective and gradient stability make them well-suited for data-scarce, privacy-sensitive, and heterogeneous settings (*e.g.,* edge computing and federated learning), providing the theoretical foundation for subsequent discussions.

3.3 Diffusion Models in Edge Computing

Edge computing, as a distributed computing paradigm, deploys data processing and computational resources to the network edge near data generation sources (*e.g.,* IoT devices, mobile terminals, or intelligent sensors) [1, 69, 70]. This architecture aims to reduce latency, minimize bandwidth consumption, and bolster privacy protection [20, 70]. Nevertheless, the efficacy of this paradigm is often challenged by inherent limitations of edge nodes, including constrained computational capacity, unstable communication links, and data heterogeneity, which hinder the efficient operation of conventional deep generative models. Within this context, DMs, with their robust training mechanisms and superior data generation capabilities, have emerged as a supportive technology for edge computing scenarios [18–20].

- **Data augmentation and distribution alignment.** Edge nodes often suffer from limited data volume and imbalanced class distributions, leading to overfitting and poor generalization. DMs trained locally or in the cloud can generate high-fidelity samples to augment scarce data and alleviate Non-IID issues effectively. In medical imaging, DDPMs [16] have been synthesized underrepresented classes to improve diagnostic accuracy without exploring raw patient data [71]. Similarly, LDM [17] generates high-quality samples in a compressed latent space, significantly reducing computational and storage costs for edge deployment.
- **Privacy preservation and Secure Generation.** In privacy-sensitive domains such as healthcare, finance, and the Industrial Internet of Things (IIoT), DMs can synthesize semantically consistent data without revealing raw information [72]. This mechanism enables secure collaboration and model optimization under strict privacy constraints. For example, in smart home and industrial monitoring applications, conditional diffusion models have been deployed to reconstruct anomalous or missing sensor data from noise [73, 74]. Such generative privacy-preserving strategies establish a secure foundation for subsequent federated optimization and cross-node collaboration.
- **Multimodal and Collaborative Generation.** As edge–cloud collaborative intelligence advances, DMs are extending from single-modality generation to multimodal and cross-device collaboration. By jointly modeling visual, auditory, textual, and sensory data, they enable semantically consistent multimodal generation and interaction

Table 4: A summary of lightweight Diffusion Models categorized by their optimization strategies for edge deployment.

Category	Basic idea	Method	Main Weakness
Sampling Acceleration	Reducing inference latency on reducing the denoising steps by solver design or knowledge distillation.	DDIM [62] InstaFlow [75] ADD [76]	Aggressive step reduction often leads to fidelity loss or requires high training costs.
Model Compression	Reducing model size and memory footprint by quantization, pruning, or lightweight architecture design.	PTQ-Diffusion [77] LD-Pruner [78] BK-SDM [79]	These methods face quality degradation, which may require costly fine-tuning to recover.
Data Flow Optimization	Optimizing data and gradient flow during inference stage to reduce computation and communication cost.	LDM [17] LoRA [80] ControlNet [81]	Methods are constrained by auxiliary models or the frozen backbone's capacity.

at the network edge. Recent works integrate DMs with state space models (SSMs) to enhance temporal reasoning [82], while collaborative generation through shared latent representations achieves distribution-robust inference across devices [22]. These paradigms align closely with the architecture of federated and edge intelligence systems.

In summary, DMs in edge computing are evolving from cloud-side verification to on-device deployment, showing strong potential in data augmentation, privacy protection, and system optimization. With advances in lightweight design and distributed inference, they are expected to underpin privacy-preserving and efficient federated learning, laying the foundation for OSFL and next-generation edge intelligence.

3.4 Taxonomy of Diffusion Models for Edge Computing

The deployment of DMs in edge computing is fundamentally challenged by their resource-intensive nature, which contrasts sharply with the limited capacity of edge devices. To address this, recent studies propose various lightweighting strategies that systematically reduce the model's computational, memory, and latency overheads. This section presents a taxonomy of lightweighting strategies [83], broadly categorized into three main paths according to the bottleneck they address: sampling acceleration, model compression, and data flow optimization, as summarized in Table 4.

- **Lightweighting via Sampling Acceleration.** These methods directly tackle the prohibitive inference latency caused by the iterative denoising process. The core idea is to reduce the required sampling steps from hundreds or thousands to just a few without compromising generation quality. For instance, DDIM [62] reformulates the diffusion process as an ODE for deterministic and fast sampling, while InstaFlow [75] achieves high-quality one-step generation through rectified flows. Another major avenue uses knowledge distillation (KD), such as adversarial diffusion distillation (ADD) [76], which trains a compact student model to match the teacher's output in a single step. However, despite these advances, achieving ultra-fast inference under strict edge latency remains an open challenge.
- **Lightweighting via Model Compression.** This category focuses on reducing the memory footprint of DMs, typically dominated by U-Net backbones with hundreds of millions of parameters. Methods such as PTQ-Diffusion [77] apply post-training quantization to convert FP32 weights into low-precision formats, while LD-Pruner [78] performs structured pruning guided by parameter importance. More recently, BK-SDM [79] replace redundant residual and attention blocks to achieve drastic reductions in FLOPs and latency. Although effective, aggressive compression may still lead to quality degradation, calling for better trade-offs between model size and fidelity.
- **Lightweighting via Data Flow Optimization.** This paradigm shifts the focus from model structure to the data and gradient flow during inference and adaptation.LDM [17] operates in a compressed latent space, greatly reducing data dimensionality and computational cost. For model adaptation, PEFT [80, 84] freezes the large backbone and updates only a small set of adapters, while ControlNet [81] trains lightweight auxiliary branches for conditional control. These strategies significantly reduce on-device fine-tuning costs, making local adaptation feasible. Nonetheless, optimizing data flow non-IID edge data distributions remains a major bottleneck for large-scale deployment.

In summary, lightweight diffusion strategies aim to reduce the computational and memory overhead of generative models through model compression, efficient sampling, and architectural optimization, making diffusion models increasingly feasible for deployment in resource-constrained edge environments.

3.5 Lessons

Our review of DMs, particularly the taxonomy of lightweighting strategies (Section 3.4), highlights several critical lessons. Major obstacles to on-device inference—such as high latency and massive memory footprints—are being progressively alleviated through sampling acceleration and model compression. Meanwhile, PEFT techniques have markedly lowered local adaptation costs, enabling efficient deployment on edge devices.

However, despite these advances, lightweight DMs alone do not resolve the collaborative training bottleneck inherent in

distributed learning settings. As summarized in Section 2, traditional multi-round FL is ill-suited for the edge, while OSFL represents a promising paradigm to overcome this limitation.

These observations lead to a pivotal question: can DMs serve as high-quality generative engines compatible with OSFL, and can OSFL provide the minimal-communication framework necessary to coordinate DMs across distributed edge nodes? Exploring this mutual and highly complementary relationship defines the technical core of this survey and naturally leads to Section 4 where we investigate how diffusion models and OSFL can be integrated into a unified edge-intelligent paradigm.

4 Diffusion Models Meet OSFL

OSFL is an efficient FL variant that minimizes communication by aggregating knowledge from clients in a single round, often via knowledge distillation or synthetic data generation. DMs, which generate high-fidelity data through iterative denoising, have become a key tool in this context. This section discusses the synergy between DMs and OSFL: first defining a conceptual framework for positioning existing studies (Section 4.1), using OSFL to train DMs (Section 4.2), applying DMs to improve OSFL (Section 4.3), and summarizing insights from recent advances (Section 4.4).

4.1 Conceptual Framework and Method Positioning

While DMs have also been explored in conventional multi-round federated learning settings, this survey primarily focuses on their integration within the OSFL paradigm, which targets extreme communication and resource constraints in edge intelligence scenarios. By delineating this scope, we emphasize the unique challenges of single-round knowledge coordination and distinguish our discussion from approaches based on iterative multi-round training or fully centralized diffusion model optimization.

The integration of OSFL and DMs entails trade-offs among communication efficiency, computational overhead, and personalization requirements. Existing studies can be categorized into two primary integration paradigms: OSFL for DMs and DMs for OSFL. OSFL for DMs focuses on the collaborative training or adaptation of diffusion models themselves under strict single-round communication constraints [21]. In this context, the diffusion model serves as the primary optimization target [85], while OSFL provides the distributed training and aggregation framework. In contrast, the DMs for OSFL paradigm leverages the generative or representational capabilities of DMs to enhance the one-shot aggregation process [10, 86, 87]. Typical strategies include synthesizing proxy data to mitigate distribution shifts or employing generative priors to facilitate single-round knowledge distillation. While these two paradigms may share underlying technical components, they differ fundamentally in their optimization objectives and systemic roles. OSFL for DMs aims to achieve the convergence of high-dimensional generative models despite extreme communication limits, whereas DMs for OSFL treats the generative model as an ”auxiliary prior” to enhance the aggregation quality of traditional federated models. To clarify the synergy and technical boundaries between these two integration paradigms, we establish a conceptual framework based on three orthogonal dimensions:

- **Communication Object and Bottleneck:** In OSFL for DMs, the primary challenge lies in the transmission of massive model parameters (*e.g.,* U-Net backbones), necessitating PEFT [84] or model-based strategies [87] to fit single-round constraints. In contrast, in DMs for OSFL, the focus shifts to exchanging lightweight latent representations or prompts to guide server-side synthesis [10]. This approach aims to effectively mitigate the real-to-synthetic distribution shift with minimal bandwidth consumption, enabling the server to reconstruct data proxies that capture local statistical properties without raw data exchange.
- **Personalization Level:** Existing methods differ substantially in the extent of client-specific adaptation [88]. Some approaches focus on learning a unified global generative prior 4.3) [10], while others incorporate lightweight adapters (*e.g.,* LoRA) to strike a balance between global knowledge sharing and local domain specialization (Section 4.2) [89].
- **Edge Constraints:** The dependence on edge computational resources varies markedly across strategies. Paradigms involving on-device diffusion training (Section 4.2) impose substantial memory and energy burdens, making them suitable for high-end edge servers. In contrast, approaches leveraging pre-trained generative priors (Section 4.3) typically require only lightweight local processing, rendering them highly compatible with resource-constrained IoT environments.

This framework clarifies the conceptual boundaries between the two integration paradigms and provides a structured perspective for understanding existing approaches across communication characteristics, personalization levels, and edge resource constraints.

4.2 OSFL for Diffusion Models

In this subsection, we discuss how to leverage OSFL for training DMs, where the diffusion model is the primary optimization target, focusing on adapting its high-dimensional parameters to edge data in a single round. Prior to this, we first examine how DMs are trained within the FL framework. FL for DMs often requires multiple rounds of update aggregation in order to align with the models' inherent multi-step denoising procedure [90]. Building upon the FedAvg algorithm, Goede *et al.* [91] trained a denoising DM. By innovatively leveraging the underlying UNet backbone, they reduced the volume of parameters exchanged during training by up to 74% compared to the naive FedAvg approach, with comparable image generation quality to centralized training. To address the need for multi-modal data fusion in remote sensing and the limitations of existing DMs confined to single modalities and single clients, Li *et al.* [89] proposed FedDiff. The core of this framework is a dual-branch DM, which processes data from different modalities separately and establishes complementary connections through their intrinsic denoising steps. Furthermore, Yoon *et al.* [85] proposed the

VQ-FedDiff algorithm for training DMs in an FL setting. It is a personalized approach that can generate higher-quality images while reducing the risk of privacy leakage to a level comparable to that of a secure model trained only locally.

While there has been some research on training DMs using FL, employing OSFL for DM training remains a relatively unexplored research direction. We note that Peng *et al.* [21] have proposed a pioneering one-shot scheme. In [21], FedDDPM introduced server-side refinement using auxiliary datasets to correct local biases during aggregation. To adapt this to one-shot scenarios, enhanced variants such as FedDDPM+ employ a single corrective step on the server, detecting slow-converging models and applying one-shot adjustments with synthetic data reflecting global distributions. This enables convergence in non-convex settings typical of DMs' U-Net architectures, with theoretical guarantees under non-IID conditions. However, this study does not achieve the implementation of OSFL for DM training.

Based on a summary of existing research, it has been found that combining the inherent complexity of DMs with the stringent constraints of OSFL gives rise to a series of unique and difficult-to-overcome challenges.

- **Conflict between model parameters and single-round communication.** DMs (*e.g.,* those based on U-Net) contain a massive number of parameters. Even transmitting a compressed or partial subset of these parameters within a single OSFL communication round can cause significant network overhead. Transmitting all parameters once within the single communication round of OSFL places an unbearable burden on bandwidth [92].
- **Effective knowledge aggregation within a single round.** Client data is typically non-IID. This distributional heterogeneity complicates the alignment of knowledge learned from different clients within a single communication round. With only one communication round in OSFL, it is challenging to effectively fuse knowledge from diverse distributions through simple model averaging, which easily leads to a degradation in generation quality [93].
- **Privacy leakage risks in generative models.** From a threat model perspective, adversaries or an "honest-but-curious" server may attempt to infer sensitive information from shared model updates or generated outputs. DMs possess a strong capacity for memorization, which can inadvertently encode sensitive patterns from local training data into model parameters or latent features. Consequently, under vexplicit attacks such as model inversion [94] or attribute inference [95], shared model updates or generated features may implicitly contain sufficient information to reconstruct private data characteristics, creating potential privacy leakage hazards even in OSFL setting [96].
- **Client-side computational and energy demands.** DMs are characterized by computationally intensive training and inference procedures, which involve numerous iterative steps. This multi-step nature results in high computational and energy demands that are difficult to meet under typical edge resource constraints. Consequently, deploying DMs on edge devices poses significant challenges to both computational capacity and battery life [92].

In summary, applying OSFL to DMs enables collaborative generative model training under strict communication constraints. However, the large model size and complex generative dynamics of diffusion models make efficient knowledge aggregation and model synchronization particularly challenging in one-shot settings.

4.3 Diffusion Models for OSFL

Unlike OSFL for DMs, the application of pre-trained DMs in OSFL has been extensively demonstrated. In this subsection, we first introduce the contributions of DMs to FL, and then discuss their usefulness for OSFL. The data generation capability of DMs can be used to address challenges of F,L such as data heterogeneity, privacy attacks, and parameter aggregation, as summarized in Table 5.

- **Data heterogeneity.** DMs mitigate the data heterogeneity problem by virtue of their ability to generate high-quality, diverse, and balanced synthetic datasets [97–100]. Wang *et al.* [97] proposed FedDifRC, a framework that utilizes the rich semantic guidance from text-to-image diffusion models to alleviate data heterogeneity. Furthermore, CRFed [98] introduced a data harmonization mechanism that employs data augmentation, noise injection, and iterative denoising to reduce data distribution disparities among nodes and ensure consistent model updates. In [100], the authors proposed utilizing DMs to generate synthetic data to address data imbalance in industrial and medical domains.
- **Privacy attacks.** The powerful generative capability of DMs poses privacy risks such as generative reconstruction and membership inference attacks. These attacks may enable adversaries to memorize and expose sensitive training data details, potentially leading to inadvertent privacy leakage when model updates or synthetic samples are shared [101, 102]. Gu *et al.* [101] proposed GGDM, a novel training-free attack method that innovatively leverages pre-trained DMs. It enables the reconstructed images to closely mirror the private information of the original data. To mitigate these risks, researchers have explored both reactive and proactive defenses. For federated unlearning, Liu *et al.* [102] proposed FedDM, which uses a diffusion model to generate noisy data for precisely forgetting specific knowledge without compromising overall model performance. Furthermore, for practical deployment, we recommend incorporating differential privacy [72] during the noise prediction process and employing secure aggregation for one-shot updates to systematically enhance the privacy resilience of OSFL-DM systems [22].
- **Parameter aggregation.** Furthermore, DMs can also be applied to the aggregation of personalized models [103–105]. In [103], pFedGPA proposed a generative parameter aggregation framework. This method deploys a DM on the server side, which utilizes a parameter inversion technique to transform the parameters uploaded by each client into latent codes. These codes are then aggregated through the denoising sampling process of the diffusion model to generate a set of personalized parameters for each client. This non-linear aggregation approach effectively decouples the

Table 5: A summary of the roles of DMs in FL includes data heterogeneity, privacy attacks, and parameter aggregation.

Category	Basic idea	Method	Publication	Intermediary
Data heterogeneity	DMs mitigate the data heterogeneity problem by virtue of their ability to generate high-quality, diverse, and balanced synthetic datasets.	FedDifRC [97]	ICCV '25	Data
		CRFed [98]	NeurIPS '24	Data
		FL-DiG [99]	TCSS '25	Data
		Gen-FedSD [100]	BigData '24	Data
Privacy attacks	Leverage client-side updates to guide DMs in generating private data.	GGDM [101]	WWW '24	Gradient
		FedDM [102]	Arxiv '24	Gradient
Parameter aggregation	By mitigating client dift among all nodes through DMs, so as to guide client selection and data sampling.	pFedGPA [103]	AAAI '25	Parameter
		DDGR [104]	Arxiv '24	Parameter
		FedDiffRec [105]	ICASSP '25	Parameter

complexity of the global parameter distribution from that of individual clients.

Regarding the application of DMs to OSFL, their principal application lies in enabling the OSFL training workflow. As summarized in Table 3, it can be observed that both model-based and feature-based approaches can leverage pre-trained DMs to achieve one-shot communication. Pre-trained DMs serve as a key module in OSFL to enable efficient one-shot communication [11, 22, 53, 55, 86, 87]. FedDISC [87] proposesd that the server utilizes a small amount of stylistic and semantic features extracted from the clients to guide a pre-trained DM. This model then generates a high-quality synthetic dataset that conforms to the client data distribution, which is subsequently used to train a global model. Additionally, Zaland *et al.* [86] proposed OSCAR, aiming to address the high computational and communication overhead caused by the requirement for pre-trained DMs to train additional classifiers on the client side. However, several challenges remain to be addressed, *e.g.,* how to obtain pre-trained DMs, the real-to-synthetic distribution shift, and privacy concerns regarding pre-trained DMs.

- **How to obtain pre-trained DMs.** In edge computing scenarios, pre-trained DMs are typically large, posing a significant challenge for resource-constrained edge devices *e.g.,* smartphones and IoT devices. The acquisition process involves downloading the model from a central server [22]. However, limited edge network bandwidth and high latency can lead to transmission failures or excessively long download times. Furthermore, insufficient device storage and computational power make it difficult to store or load these models efficiently. Additionally, pre-trained models suitable for specific tasks are not readily available in all edge scenarios, and customized training on the edge side is highly challenging.
- **The real-to-synthetic distribution shift.** Pre-trained DMs excel at generating synthetic data to augment local datasets in OSFL. However, a significant distribution shift exists between the synthetic and real-world data. In edge computing, real data typically originates from specific device environments (*e.g.,* sensor data, user behavior) and is characterized by noise, diversity, and temporal dynamics. In contrast, synthetic data generated by pre-trained DMs tends to align more with a generic distribution, leading to poor model generalization. To protect privacy and conserve bandwidth, clients usually upload only highly compressed data representations, *e.g.,* the model parameters [10], soft labels [11], or feature representations [22]. Yet, this condensed information may be insufficient to fully reconstruct the complex distribution of the original real data.
- **Privacy concerns regarding pre-trained DMs.** The application of pre-trained DMs in OSFL involves privacy risks, as these models may retain sensitive information from their training data (such as biases learned from public datasets or reversible inferences of original samples) [101]. In edge computing, when DMs are used locally on devices to generate data, sharing or aggregating model parameters with the server could expose user-private data through model inversion attacks or membership inference attacks. Furthermore, if synthetic data closely resembles real data, it may indirectly reveal personal information (*e.g.,* generated user portraits or location data).

In short, DMs enhance OSFL by generating high-fidelity synthetic data to mitigate non-IID challenges and data scarcity. The core objective shifts from traditional weight aggregation to aligning latent representations, which effectively bridges the real-to-synthetic distribution gap while maintaining strong privacy guarantees.

4.4 Lessons

The integration of DMs and OSFL represents an emerging paradigm that, while not yet fully explored, holds significant promise. This approach can effectively address data heterogeneity and privacy concerns within a single communication round, while also improving the training efficiency of generative models under a federated learning framework. In light of recent research developments, this paper first establishes the conceptual framework and method positioning of two integration paradigms, followed by summarizing key insights and lessons from two primary perspectives: (1) methods for training DMs based on OSFL, and (2) approaches to enhance OSFL performance using DMs.

OSFL for DMs focuses on the federated training of the DMs themselves within the OSFL framework. It achieves collaborative and privacy-preserving generative model training by having clients upload their local representations in a single round. However, existing research still faces several challenges, such as the conflict between model parameters and single-round communication, the difficulty of effective knowledge aggregation within a single round, privacy leakage risks in generative models, and high client-side computational and energy demands.

In FL, DMs can address its inherent data heterogeneity issues, enable privacy attacks, and facilitate parameter aggregation. In the OSFL framework, DMs serve as a generative tool to synthesize client data proxies within a single communication round, thereby avoiding multiple iterations and reducing data leakage risks. The core concept involves leveraging server-side pre-trained DMs to generate synthetic datasets for efficient global model training. However, in the context of edge computing, several challenges remain to be addressed, such as how to obtain pre-trained DMs, the real-to-synthetic data distribution shift, and privacy concerns associated with the pre-trained models.

5 Applications

The theoretical integration between DMs and OSFL, as discussed in Section 4, provides a powerful framework for deploying privacy-preserving generative AI in real-world settings. The inherent characteristics of edge computing environments—such as data sensitivity, constrained communication, and the need for low-latency responses—make this framework particularly suitable for several critical application domains. This section examines how this unified paradigm can be instantiated in three representative application scenarios: healthcare (Section 5.1), battery swapping networks (Section 5.2), and Mobile Crowdsensing networks (Section 5.3).

5.1 Healthcare

The healthcare domain presents a clear and compelling use case. Medical data, such as Magnetic Resonance Imaging (MRI) or Computed Tomography (CT) scans, is protected by stringent privacy regulations (*e.g.,* HIPAA or GDPR) and is consequently siloed within individual hospitals. Aggregating this data to train a robust global diagnostic model is a primary goal. While traditional FL offers a privacy-preserving baseline, its multi-round communication cost is a significant barrier.

The DMs for OSFL paradigm offers a promising solution [106]. In this model, edge hospitals (clients) train models on their private local data. These clients can employ lightweight, personalized FL strategies to efficiently capture local data characteristics, such as those explored in pFedCal [8]. After local training, the clients perform a single communication round, uploading only abstracted parameters or distilled features. The central server aggregates this one-shot information to guide a robust, central diffusion model. This DM then generates a large-scale, high-fidelity, and privacy-safe synthetic medical dataset. Frameworks like FedDrip [107] further demonstrate this approach, using diffusion-generated images at a 'pseudo-site' to enhance model generalization and overcome data scarcity in non-IID medical datasets. This synthetic dataset, which captures the statistical properties of the global data distribution without containing any real patient data, is subsequently used to train a high-performance global diagnostic model, effectively mitigating the real-to-synthetic distribution shift.

5.2 Battery Swapping

Battery swapping networks (BSNs) for electric vehicles (EVs) and electric bicycles represent a critical IoT-edge application, characterized by highly dynamic and heterogeneous demand patterns. Due to their large-scale urban deployment, each swapping station continuously generates real-time spatiotemporal data regarding local battery usage, charging status, and user mobility flows. Modeling and balancing such networks remain significant challenges. Prior studies have highlighted this complexity; for example, Zhou *et al.* [25] demonstrated the need for spatio-temporal recommendation techniques to balance electric scooter swapping networks. However, city-wide logistics optimization and demand prediction require aggregating knowledge from all stations, which is hindered by commercial sensitivity and the high communication overhead typically inherent in traditional multi-round FL.

This scenario is ideally suited for the DMs for OSFL paradigm. In this framework, battery swapping stations perform local training to capture their unique demand distributions and upload their model updates to the server only once. The server then leverages this aggregated knowledge to guide a powerful generative model. A prime example of this synergy is presented by Cheng *et al.* [24], who introduced a Knowledge-guided Diffusion model (KGDM) for the prediction of cross-city battery swap demand. By incorporating domain knowledge into the diffusion process, this approach enables significantly more accurate spatio-temporal prediction than traditional forecasting methods. This OSFL+DM framework therefore preserves sensitive station-level data while enabling accurate city-scale demand prediction and logistics planning.

5.3 Mobile Crowdsensing

Mobile crowdsensing (MCS) is a practical paradigm of edge intelligence, in which numerous mobile users and devices collaboratively collect and upload sensing data to support large-scale urban analytics [108]. Each participant typically operates under stringent constraints on computation, energy, and wireless bandwidth, while producing heterogeneous data streams from cameras, wearable sensors, or smartphones [109]. Real-world deployments reveal that MCS environments are inherently dynamic: participating devices exhibit diverse hardware capabilities, mobility patterns, and intermittent connectivity, requiring adaptive scheduling and robust communication [28, 110, 111]. Many tasks—such as crowd monitoring, traffic flow estimation, and anomaly detection—require a careful balance among latency, sensing accuracy, and energy consumption, as demonstrated in edge-assisted mobile sensing systems [27]. These characteristics render traditional multi-round FL impractical, since frequent synchronization is incompatible with fluctuating link quality and opportunistic participation patterns in MCS networks.

This setting aligns well with the OSFL for DMs paradigm. Each participant performs local adaptation on a lightweight diffusion model—often through parameter-efficient tuning such as LoRA [80]—to learn task-specific latent representations, such as scene-conditioned sensing or mobility-aware

prediction. When connectivity is available (*e.g.,* upon encountering a WiFi hotspot or returning to a charging station), devices upload their local updates in a single communication round. The server aggregates these one-shot updates to maintain a global diffusion model, which is then redistributed to participants to guide future sensing tasks. This approach minimizes communication overhead while enabling robust collaborative learning in highly dynamic and resource-limited mobile crowdsensing systems.

6 Conclusion and Future Directions

This section synthesizes the key insights derived from our examination of the OSFL–DM paradigm and highlights their broader implications for edge intelligence. Building on the analyses and applications presented earlier, we summarize the key lessons learned (Section 6.1)and outline the remaining challenges that motivate future research directions(Section 6.2) in communication-efficient and generative edge learning.

6.1 Insights and conclusion

This survey provides a comprehensive investigation into the integration between OSFL and DMs, two paradigms that naturally complement each other in heterogeneous, communication-constrained edge environments. Building upon the foundations established in the introduction, our analysis reaffirmed that OSFL addresses a fundamental bottleneck of edge intelligence—namely the prohibitive communication cost of traditional multi-round FL—by enabling global aggregation through a single exchange of compact representations. However, its effectiveness critically hinges on the availability of expressive generative mechanisms capable of mitigating non-IID data divergence, representation sparsity, and real-to-synthetic distribution mismatch.

Diffusion models naturally fill this critical void by generating high-fidelity data, performing latent-space reconstruction, and aligning distributions. To enable client-side deployment, recent advances in latent diffusion architectures, sampling acceleration, parameter-efficient adaptation (*e.g.,* LoRA), and model compression techniques have significantly lowered inference and adaptation barriers. However, lightweight inference alone does not solve the collaborative learning bottleneck. This necessitates a synergistic approach in which OSFL provides a principled, minimal-communication pathway to federate diffusion models. At the same time, DMs enhance OSFL by providing robust generative priors that compensate for heterogeneity and data scarcity.

Furthermore, this OSFL–DM synergy extends beyond theoretical frameworks, demonstrating practical viability in mission-critical domains, including privacy-preserving healthcare collaboration, spatiotemporal battery-demand prediction in swapping networks, and environmental modeling for mobile crowdsensing systems. Collectively, these insights highlight OSFL–DM integration as a promising foundation for next-generation generative, collaborative, and privacy-aware edge AI.

6.2 Existing Challenges and Future Directions

Building on the foundations laid in this survey, several critical avenues for future research emerge to address unresolved challenges in resource constraints, temporal dynamics, trust, and heterogeneity.

- **Hardware-Aware Generative Edge AI.** The severe resource constraints on edge devices require moving beyond theoretical lightweighting. Future research should explore hardware–algorithm co-design for DMs, including ultra-low-bit quantization ($<$ 4-bit) and accelerators tailored to iterative denoising, ensuring that computational and energy costs remain compatible with highly constrained platforms such as micro-UAVs [27].
- **Lifelong and Continual OSFL-DM Adaptation.** Most OSFL methods assume a static data snapshot, whereas real-world edge environments undergo continuous non-stationary drift [112]. A key direction is extending OSFL-DM toward lifelong adaptation frameworks that integrate continual learning principles with diffusion-based updates, enabling global models to evolve over time without catastrophic forgetting.
- **Heterogeneity-Robust One-Shot Aggregation.** Non-IID data and heterogeneous device capabilities remain major bottlenecks in single-round aggregation. Future frameworks may incorporate heterogeneity-resilient mechanisms such as knowledge trimming [23] and adaptive calibration to align diverse local latent distributions during the one-shot aggregation process [8].
- **Trustworthy, Verifiable, and Auditable Systems.** Beyond privacy considerations, deploying generative models at the edge introduces broader challenges in trust and accountability. Promising directions include differentially private diffusion models [72] for one-shot settings, verifiable generation mechanisms to prevent synthetic data from encoding sensitive attributes, and auditable pipelines for detecting backdoors or biases in aggregated global models—particularly in safety-critical applications such as healthcare.
- **Multimodal and Controllable Synthesis.** Edge applications are inherently multimodal. Future OSFL-DM systems should support multimodal synthesis that integrates visual, textual, LiDAR, and RF modalities, while enabling fine-grained controllable generation via lightweight conditioning mechanisms that translate high-level objectives into precise domain-specific synthetic samples [113].

In summary, advancing OSFL-DM research requires addressing efficiency, robustness, trust, and multimodality at their intersection. These future directions outline a pathway toward scalable, privacy-preserving, and generative intelligence for next-generation edge ecosystems.

Funding

This work was supported in part by NSF China under Project 62572082 and in part by NSF of Chongqing under Project CSTB2024NSCQ-JQX0020.

Author Contribution

Conceptualization: W. Chen and D. Deng; methodology: W. Chen, D. Deng and C. Xiang; formal analysis: W. Chen and D. Deng; investigation: W. Chen and D. Deng; data curation: W. Chen and D. Deng; writing—original draft preparation: W. Chen and D. Deng; writing—review and editing: W. Chen, D. Deng, C. Xiang, Z. Liu and B. Xiao; visualization: W. Chen and D. Deng; supervision: C. Xiang, Z. Liu and B. Xiao; project administration: C. Xiang. All authors have read and agreed to the published version of the manuscript.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Zhang, T., Kong, F., Deng, D., Tang, X., Wu, X., Xu, C., Zhu, L., Liu, J., Ai, B., Han, Z., *et al.*: Moving Target Defense Meets Artificial Intelligence-Driven Network: A Comprehensive Survey. IEEE Internet of Things Journal **12**(10), 13384–13397 (2025). https://doi.org/10.1109/JIOT.2025.3533016

[2] Wu, T., Fan, X., Wei, H., Qu, Y., Xiang, C., Yang, P., Wu, F.: Predictive Service Provisioning with Online Learning in Wireless Edge Networks. IEEE Transactions on Mobile Computing **23**(5), 4076–4091 (2024). https://doi.org/10.1109/TMC.2023.3286847

[3] Deng, D., Xiang, C., Gu, C., Yu, B., Chen, H., Wu, X., Gao, R.: Harmonizing Time-of-use Pricing and Rigid Driver Schedule in Battery Swapping Service. IEEE Transactions on Intelligent Transportation Systems **27**(3), 3605–3618 (2026). https://doi.org/10.1109/TITS.2025.3639263

[4] Zhang, W., Deng, D., Wu, X., Zhang, T., Zheng, X., Niyato, D., Kim, D.I.: Mitigating Catastrophic Forgetting in Personalized Federated Learning for Edge Devices using State-Space Models. IEEE Transactions on Mobile Computing **25**(3), 3568–3582 (2026). https://doi.org/10.1109/TMC.2025.3618886

[5] Deng, D., Zhao, W., Wu, X., Zhang, T., Zheng, J., Kang, J., Niyato, D.: DecFFD: A Personalized Federated Learning Framework for Cross-Location Fault Diagnosis. IEEE Transactions on Industrial Informatics **20**(5), 7082–7091 (2024). https://doi.org/10.1109/TII.2024.3353920

[6] McMahan, B., Moore, E., Ramage, D., Hampson, S., yArcas, B.A.: Communication-Efficient Learning of Deep Networks from Decentralized Data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, pp. 1273–1282 (2017)

[7] Deng, D., Wu, X., Zhang, T., Tang, X., Du, H., Kang, J., Liu, J., Niyato, D.: FedASA: A Personalized Federated Learning with Adaptive Model Aggregation for Heterogeneous Mobile Edge Computing. IEEE Transactions on Mobile Computing **23**(12), 14787–14802 (2024). https://doi.org/10.1109/TMC.2024.3446271

[8] Deng, D., Wu, X., Zhang, T., Xiang, C., Zhao, W., Xu, M., Kang, J., Han, Z., Niyato, D.: pFedCal: Lightweight Personalized Federated Learning with Adaptive Calibration Strategy. IEEE Transactions on Services Computing **18**(3), 1627–1640 (2025). https://doi.org/10.1109/TSC.2025.3553707

[9] Guha, N., Talwalkar, A., Smith, V.: One-shot federated learning. arXiv preprint arXiv:1902.11175 (2019). https://doi.org/10.48550/arXiv.1902.11175

[10] Zhang, J., Chen, C., Li, B., Lyu, L., Wu, S., Ding, S., Shen, C., Wu, C.: Dense: Data-free one-shot federated learning. In Proceedings of the 36th International Conference on Neural Information Processing Systems, pp. 21414–21428 (2022)

[11] Yang, M., Su, S., Li, B., Xue, X.: FedDEO: Description-Enhanced One-Shot Federated Learning with Diffusion Models. In Proceedings of the 32nd ACM International Conference on Multimedia, pp. 6666–6675 (2024). https://doi.org/10.1145/3664647.3681490

[12] Liu, X., Tang, Z., Li, X., Song, Y., Ji, S., Liu, Z., Han, B., Jiang, L., Li, J.: One-shot federated learning methods: A practical guide. In Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, pp. 10573–10581 (2025). https://doi.org/10.24963/ijcai.2025/1174

[13] Amato, F., Qiu, L., Tanveer, M., Cuomo, S., Giampaolo, F., Piccialli, F.: Towards One-shot Federated Learning: Advances, Challenges, and Future Directions. Neurocomputing **664**, 132088 (2026). https://doi.org/10.1016/j.neucom.2025.132088

[14] Ayeelyan, J., Utomo, S., Rouniyar, A., Hsu, H.-C., Hsiung, P.-A.: Federated learning design and functional models: Survey. Artificial Intelligence Review **58**, 21 (2024). https://doi.org/10.1007/s10462-024-10969-y

[15] Gargary, A.V., De Cristofaro, E.: A systematic review of federated generative models. arXiv preprint arXiv:2405.16682 (2024). https://doi.org/10.48550/arXiv.2405.16682

[16] Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In Proceedings of the 34th International Conference on Neural Information Processing Systems, pp. 6840–6851 (2020)

[17] Rombach, R., Blattmann, A., Lorenz, D., Esser, P.,

Ommer, B.: High-resolution image synthesis with latent diffusion models. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695 (2022). https://doi.org/10.1109/CVPR52688.2022.01042

[18] Ye, Y., Xu, K., Huang, Y., Yi, R., Cai, Z.: DiffusionEdge: Diffusion Probabilistic Model for Crisp Edge Detection. In Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, pp. 6675–6683 (2024). https://doi.org/10.1609/aaai.v38i7.28490

[19] Zheng, D.: Diffusion Models on the Edge: Challenges, Optimizations, and Applications. arXiv preprint arXiv:2504.15298 (2025). https://doi.org/10.48550/arXiv.2504.15298

[20] Chai, Z., Lin, Y., Gao, Z., Yu, X., Xie, Z.: Diffusion Model Empowered Efficient Data Distillation Method for Cloud-Edge Collaboration. IEEE Transactions on Cognitive Communications and Networking **11**(2), 902–913 (2025). https://doi.org/10.1109/TCCN.2025.3527647

[21] Peng, Z., Wang, X., Chen, S., Rao, H., Shen, C.: Federated Learning for Diffusion Models. IEEE Transactions on Cognitive Communications and Networking **11**(6), 4093–4109 (2025). https://doi.org/10.1109/TCCN.2025.3550359

[22] Chen, H., Li, H., Zhang, Y., Bi, J., Zhang, G., Zhang, Y., Torr, P., Gu, J., Krompass, D., Tresp, V.: FedBiP: Heterogeneous One-Shot Federated Learning with Personalized Latent Diffusion Models. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 30440–30450 (2025). https://doi.org/10.1109/CVPR52734.2025.02834

[23] Deng, D., Zhang, T., Gu, C., Xiang, C., Wu, X.: Less is More: Enabling Efficient and Fair Federated Learning by Knowledge Trimming. In 2025 IEEE/ACM 33rd International Symposium on Quality of Service (IWQoS), pp. 1–10 (2025). https://doi.org/10.1109/IWQoS65803.2025.11143456

[24] Cheng, W., Lu, H., Xiang, C., Liu, D., Xiang, T.: Breaking 'Chicken-Egg': Cross-city Battery Swap Demand Prediction via Knowledge-guided Diffusion. In IEEE Conference on Computer Communications, pp. 1–10 (2025). https://doi.org/10.1109/INFOCOM55648.2025.11044523

[25] Zhou, E., Li, Z., Liu, D., Xiang, C., Chen, J., Cheng, W.: Balancing Electric Scooter Battery Swapping Network by Spatio-Temporal Recommendation. IEEE Transactions on Intelligent Transportation Systems **25**(12), 21315–21326 (2024). https://doi.org/10.1109/TITS.2024.3457786

[26] Jiang, K., Wang, Y., Wang, H., Liu, Z., Han, Q., Zhou, A., Xiang, C., Cai, Z.: A Reinforcement Learning-Based Incentive Mechanism for Task Allocation Under Spatiotemporal Crowdsensing. IEEE Transactions on Computational Social Systems **11**(2), 2179–2189 (2024). https://doi.org/10.1109/TCSS.2023.3263821

[27] Xiang, C., Li, Z., Zhang, Q., Wu, X., Zheng, X., Guo, Y.: EagleEye: Balancing Latency, Accuracy, and Power on Edge-Assisted UAVs for Urban Crowd Surveillance. IEEE Transactions on Mobile Computing **24**(12), 13062–13077 (2025). https://doi.org/10.1109/TMC.2025.3586860

[28] Gu, C., Deng, D., Xiang, C.: Risk-aware model predictive control framework for collision avoidance in unmanned surface vehicles. Ocean Engineering **342**, 122954 (2025). https://doi.org/10.1016/j.oceaneng.2025.122954

[29] Jiao, X., Ou, H., Chen, S., Guo, S., Qu, Y., Xiang, C., Shang, J.: Deep reinforcement learning for time-energy tradeoff online offloading in MEC-enabled industrial internet of things. IEEE Transactions on Network Science and Engineering **10**(6), 3465–3479 (2023). https://doi.org/10.1109/TNSE.2023.3263169

[30] Sharma, M., Tomar, A., Hazra, A.: Edge computing for industry 5.0: Fundamental, applications, and research challenges. IEEE Internet of Things Journal **11**(11), 19070–19093 (2024). https://doi.org/10.1109/JIOT.2024.3359297

[31] Geng, H., Deng, D., Zhang, W., Ji, P., Wu, X.: Personalized federated learning based on bidirectional knowledge distillation for wifi gesture recognition. Electronics **12**(24), 5016 (2023). https://doi.org/10.3390/electronics12245016

[32] Zhang, W., Deng, D., Wang, L.: FedScrap: Layer-Wise Personalized Federated Learning for Scrap Detection. Electronics **13**(3), 527 (2024). https://doi.org/10.3390/electronics13030527

[33] Hasan, M.K., Jahan, N., Nazri, M.Z.A., Islam, S., Khan, M.A., Alzahrani, A.I., Alalwan, N., Nam, Y.: Federated learning for computational offloading and resource management of vehicular edge computing in 6G-V2X network. IEEE Transactions on Consumer Electronics **70**(1), 3827–3847 (2024). https://doi.org/10.1109/TCE.2024.3357530

[34] Zhang, W., Deng, D., Wu, X., Zhao, W., Liu, Z., Zhang, T., Kang, J., Niyato, D.: An adaptive asynchronous federated learning framework for heterogeneous Internet of things. Information Sciences **689**, 121458 (2025). https://doi.org/10.1016/j.ins.2024.121458

[35] Aslanpour, M.S., Toosi, A.N., Cheema, M.A., Chhetri, M.B.: FaasHouse: sustainable serverless edge computing through energy-aware resource scheduling. IEEE Transactions on Services Computing **17**(4), 1533–1547 (2024). https://doi.org/10.1109/TSC.2024.3354296

[36] Zhang, X., Tian, J., Zhang, J., Xiang, C.: Fine-grained caching and resource scheduling for adaptive bitrate videos in edge networks. ACM Transactions on Sensor Networks **19**(4), 1–30 (2023). https://doi.org/10.1145/3604555

[37] Xiang, C., Zhang, Z., Qu, Y., Lu, D., Fan, X., Yang, P., Wu, F.: Edge Computing-Empowered Large-Scale Traffic Data Recovery Leveraging Low-Rank Theory. IEEE Transactions on Network Science and Engineering **7**(4), 2205–2218 (2020). https://doi.org/10.1109/TNSE.2020.2984658

[38] Fan, L., He, L., Wu, Y., Zhang, S., Wang, Z., Li, J., Yang, J., Xiang, C., Ma, X.: AutoIoT: Automatically Updated IoT Device Identification With Semi-Supervised Learning. IEEE Transactions on Mobile Computing **22**(10), 5769–5786 (2022). https://doi.org/10.1109/TMC.2022.3183118

[39] Hu, C., Li, B.: Maskcrypt: Federated learning with selective homomorphic encryption. IEEE Transactions on Dependable and Secure Computing **22**(1), 221–233 (2024). https://doi.org/10.1109/TDSC.2024.3392424

[40] Kumar, K.N., Mitra, R., Mohan, C.K.: Revamping Federated Learning Security from a Defender's Perspective: A Unified Defense with Homomorphic Encrypted Data Space. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 24387–24397 (2024). https://doi.org/10.1109/CVPR52733.2024.02302

[41] Hu, J., Du, J., Wang, Z., Pang, X., Zhou, Y., Sun, P., Ren, K.: Does Differential Privacy Really Protect Federated Learning From Gradient Leakage Attacks? IEEE Transactions on Mobile Computing **23**(12), 12635–12649 (2024). https://doi.org/10.1109/TMC.2024.3417930

[42] Ortega, T., Jafarkhani, H.: Quantized and Asynchronous Federated Learning. IEEE Transactions on Communications **73**(4), 2361–2374 (2025). https://doi.org/10.1109/TCOMM.2024.3471996

[43] Gao, Z., Zhang, Z., Guo, Y., Gong, Y.: Federated Adaptive Fine-Tuning of Large Language Models with Heterogeneous Quantization and LoRA. In IEEE Conference on Computer Communications, pp. 1–10 (2025). https://doi.org/10.1109/INFOCOM55648.2025.11044641

[44] Qu, Z., Jia, N., Ye, B., Hu, S., Guo, S.: FedQClip: Accelerating Federated Learning via Quantized Clipped SGD. IEEE Transactions on Computers **74**(2), 717–730 (2025). https://doi.org/10.1109/TC.2024.3477972

[45] Cui, H., Qu, Z., Wang, X., Tang, B., Ye, B.: LCO-AGQ: A Lightweight Client-Oriented Adaptive Gradient Quantization Algorithm for Federated Learning. In IEEE Conference on Computer Communications, pp. 1–10 (2025). https://doi.org/10.1109/INFOCOM55648.2025.11044636

[46] Zhu, X., Wang, J., Sato, K., Zheng, Z.: Adaptive Model Compression for Efficient Federated Learning in IoT Systems. IEEE Internet of Things Journal **12**(14), 26155–26168 (2025). https://doi.org/10.1109/JIOT.2025.3557861

[47] Zhang, J., Li, X., Vijayakumar, P., Liang, W., Chang, V., Gupta, B.B.: Graph Sparsification-based Secure Federated Learning for Consumer-Driven Internet of Things. IEEE Transactions on Consumer Electronics **70**(3), 5188–5200 (2024). https://doi.org/10.1109/TCE.2024.3411551

[48] Yang, Y., Dang, S., Zhang, Z.: An Adaptive Compression and Communication Framework for Wireless Federated Learning. IEEE Transactions on Mobile Computing **23**(12), 10835–10854 (2024). https://doi.org/10.1109/TMC.2024.3382776

[49] Wei, K., Li, J., Ma, C., Ding, M., Shu, F., Zhao, H., Chen, W., Zhu, H.: Gradient sparsification for efficient wireless federated learning with differential privacy. Science China Information Sciences **67**, 142303 (2024). https://doi.org/10.1007/s11432-023-3918-9

[50] Kou, W.-B., Lin, Q., Tang, M., Ye, R., Wang, S., Zhu, G., Wu, Y.-C.: Fast-Convergent and Communication-Alleviated Heterogeneous Hierarchical Federated Learning in Autonomous Driving. IEEE Transactions on Intelligent Transportation Systems **26**(7), 10496–10511 (2025). https://doi.org/10.1109/TITS.2025.3543235

[51] Tchaye-Kondi, J., Zhai, Y., Shen, J., Telikani, A., Zhu, L.: Adaptive Period Control for Communication Efficient and Fast Convergent Federated Learning. IEEE Transactions on Mobile Computing **23**(12), 12572–12586 (2024). https://doi.org/10.1109/TMC.2024.3416312

[52] Zhou, Y., Pu, G., Ma, X., Li, X., Wu, D.: Distilled one-shot federated learning. arXiv preprint arXiv:2009.07999 (2021). https://doi.org/10.48550/arXiv.2009.07999

[53] Yang, M., Su, S., Li, B., Xue, X.: One-Shot Heterogeneous Federated Learning with Local Model-Guided Diffusion Models. In Proceedings of the 42nd International Conference on Machine Learnin, pp. 71157–71176 (2025)

[54] Bai, J., Song, Y., Wu, D., Sajjanhar, A., Xiang, Y., Zhou, W., Tao, X., Li, Y., Li, Y.: A Unified Solution to Diverse Heterogeneities in One-Shot Federated Learning. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 71–82 (2025). https://doi.org/10.1145/3711896.3736825

[55] Zhang, J., Qi, X., Zhao, B.: Federated Generative Learning with Foundation Models. arXiv preprint arXiv:2306.16064 (2023). https://doi.org/10.48550/arXiv.2306.16064

[56] Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. arXiv preprint arXiv:1312.6114 (2013). https://doi.org/10.48550/arXiv.1312.6114

[57] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In Proceedings of the 28th International Conference on Neural Information Processing Systems, pp. 2672–2680 (2014)

[58] Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the Design Space of Diffusion-Based Generative Models. In Proceedings of the 36th International Conference on Neural Information Processing System, pp. 26565–26577 (2022)

[59] Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-Based Generative Modeling through Stochastic Differential Equations. In 9th International Conference on Learning Representations, pp. 1–36 (2021)

[60] Dhariwal, P., Nichol, A.: Diffusion Models Beat GANs on Image Synthesis. In Proceedings of the 35th International Conference on Neural Information Processing Systems, pp. 8780–8794 (2021)

[61] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the 32nd International Conference on Machine Learning, pp. 2256–2265 (2015)

[62] Song, J., Meng, C., Ermon, S.: Denoising Diffusion Implicit Models. In International Conference on Learning Representations, ICLR 2021, pp. 1–20 (2021)

[63] Wang, T., Zhang, K., Zhang, Y., Luo, W., Stenger, B., Lu, T., Kim, T.-K., Liu, W.: LLDiffusion: Learning degradation representations in diffusion models for low-light image enhancement. Pattern Recognition **166**(C), 111628 (2025). https://doi.org/10.1016/j.patcog.2025.111628

[64] Huang, R., Huang, J., Yang, D., Ren, Y., Liu, L., Li, M., Ye, Z., Liu, J., Yin, X., Zhao, Z.: Make-an-Audio: Text-to-Audio Generation with Prompt-Enhanced Diffusion Models. In Proceedings of the 40th International Conference on Machine Learning, pp. 13916–13932 (2023)

[65] Mei, K., Patel, V.M.: VIDM: Video Implicit Diffusion Models. In Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence, vol. 37, pp. 9117–9125 (2023). https://doi.org/10.1609/aaai.v37i8.26094

[66] Mendieta, M., Sun, G., Chen, C.: Exploring the Effectiveness of Diffusion Models in One-Shot Federated Learning. In The Twelfth International Conference on Learning Representations, pp. 22340–22348 (2024)

[67] Nichol, A.Q., Dhariwal, P.: Improved Denoising Diffusion Probabilistic Models. In Proceedings of the 38th International Conference on Machine Learning, pp. 8162–8171 (2021)

[68] Ho, J., Salimans, T.: Classifier-Free Diffusion Guidance. arXiv preprint arXiv:2207.12598 (2022). https://doi.org/10.48550/arXiv.2207.12598

[69] Wu, J., Dong, F., Leung, H., Zhu, Z., Zhou, J., Drew, S.: Topology-aware federated learning in edge computing: A comprehensive survey. ACM Computing Surveys **56**(10), 1–41 (2024). https://doi.org/10.1145/3659205

[70] Hua, H., Li, Y., Wang, T., Dong, N., Li, W., Cao, J.: Edge Computing with Artificial Intelligence: A Machine Learning Perspective. ACM Computing Surveys **55**(9), 1–35 (2023). https://doi.org/10.1145/3555802

[71] Dorjsembe, Z., Pao, H.-K., Odonchimed, S., Xiao, F.: Conditional diffusion models for semantic 3D brain MRI synthesis. IEEE Journal of Biomedical and Health Informatics **28**(7), 4084–4093 (2024). https://doi.org/10.1109/JBHI.2024.3385504

[72] Dockhorn, T., Cao, T., Vahdat, A., Kreis, K.: Differentially Private Diffusion Models. arXiv preprint arXiv:2210.09929 (2022). https://doi.org/10.48550/arXiv.2210.09929

[73] Liu, M., Huang, H., Feng, H., Sun, L., Du, B., Fu, Y.: PriSTI: A Conditional Diffusion Framework for Spatiotemporal Imputation. In 2023 IEEE 39th International Conference on Data Engineering (ICDE), pp. 1927–1939 (2023). https://doi.org/10.1109/ICDE55515.2023.00150

[74] Zhang, P., Lin, Y., Cui, H., Gu, J.: Channel Attention-Based Conditional Diffusion Model Applied to Fault Diagnosis Under Imbalanced Data. Electronics **13**(23), 4807 (2024). https://doi.org/10.3390/

electronics13234807

[75] Liu, X., Zhang, X., Ma, J., Peng, J., Liu, Q.: InstaFlow: One Step Is Enough for High-Quality Diffusion-Based Text-to-Image Generation. In The Twelfth International Conference on Learning Representations, ICLR 2024, pp. 1–30 (2024)

[76] Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial Diffusion Distillation. In Computer Vision – ECCV 2024: 18th European Conference, pp. 87–103 (2024). https://doi.org/10.1007/978-3-031-73016-0_6

[77] Shang, Y., Yuan, Z., Xie, B., Wu, B., Yan, Y.: Post-Training Quantization on Diffusion Models. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1972–1981 (2023). https://doi.org/10.1109/CVPR52729.2023.00196

[78] Castells, T., Song, H.-K., Kim, B.-K., Choi, S.: LD-Pruner: Efficient Pruning of Latent Diffusion Models Using Task-Agnostic Insights. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 821–830 (2024). https://doi.org/10.1109/CVPRW63382.2024.00087

[79] Kim, B.-K., Song, H.-K., Castells, T., Choi, S.: BK-SDM: A Lightweight, Fast, and Cheap Version of Stable Diffusion. In Computer Vision – ECCV 2024: 18th European Conference, pp. 381–399 (2024). https://doi.org/10.1007/978-3-031-72949-2_22

[80] Hu, E.J., shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models. In The Tenth International Conference on Learning Representations, ICLR 2022, pp. 1–13 (2022)

[81] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3836–3847 (2023). https://doi.org/10.1109/ICCV51070.2023.00355

[82] Hong, Y., Mai, L., Yao, Y., Liu, F.: Pushing the Boundaries of State Space Models for Image and Video Generation. arXiv preprint arXiv:2502.00972 (2025). https://doi.org/10.48550/arXiv.2502.00972

[83] Song, W., Ma, W., Zhang, M., Zhang, Y., Zhao, X.: Lightweight Diffusion Models: A Survey. Artificial Intelligence Review **57**, 161 (2024). https://doi.org/10.1007/s10462-024-10800-8

[84] Xu, L., Xie, H., Qin, S.-Z.J., Tao, X., Wang, F.L.: Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment. arXiv preprint arXiv:2312.12148 (2023). https://doi.org/10.48550/arXiv.2312.12148

[85] Yoon, T., Hwang, M., Yang, E.: VQ-FedDiff: Federated Learning Algorithm of Diffusion Models With Client-Specific Vector-Quantized Conditioning. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(12), 11863–11873 (2025). https://doi.org/10.1109/TPAMI.2025.3602282

[86] Zaland, O., Jin, S., Pokorny, F.T., Bhuyan, M.: One-Shot Federated Learning with Classifier-Free Diffusion Models. In IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2025). https://doi.org/10.1109/ICME59968.2025.11209111

[87] Yang, M., Su, S., Li, B., Xue, X.: Exploring one-shot semi-supervised federated learning with pre-trained diffusion models. In The Thirty-Eighth AAAI Conference on Artificial Intelligence, pp. 16325–16333 (2024)

[88] Tan, A.Z., Yu, H., Cui, L., Yang, Q.: Towards Personalized Federated Learning. IEEE Transactions on Neural Networks and Learning Systems **34**(12), 9587–9603 (2023). https://doi.org/10.1109/TNNLS.2022.3160699

[89] Li, D., Xie, W., Wang, Z., Lu, Y., Li, Y., Fang, L.: Feddiff: Diffusion model driven federated learning for multi-modal and multi-clients. IEEE Transactions on Circuits and Systems for Video Technology **34**(10), 10353–10367 (2024). https://doi.org/10.1109/TCSVT.2024.3407131

[90] Gao, R., Kang, J., Lai, B., Xu, M., Sun, G., Zhang, T., Zhang, W., Yang, D.: High-quality Trajectory Generation for Autonomous Driving: A Lightweight Federated Learning-based Diffusion Model. In GLOBECOM 2024-2024 IEEE Global Communications Conference, pp. 1641–1646 (2024). https://doi.org/10.1109/GLOBECOM52923.2024.10901719

[91] Goede, M., Cox, B., Decouchant, J.: Training diffusion models with federated learning. arXiv preprint arXiv:2406.12575 (2024). https://doi.org/10.48550/arXiv.2406.12575

[92] Lai, B., He, J., Kang, J., Li, G., Xu, M., Xie, S.: On-demand quantization for green federated generative diffusion in mobile edge networks. In ICC 2024 - IEEE International Conference on Communications, pp. 2883–2888 (2024). https://doi.org/10.1109/ICC51166.2024.10622695

[93] Balan, K.G., Gilbert, A., Collomosse, J.: Pdfed: Privacy-preserving and decentralized asynchronous federated learning for diffusion models. In Proceedings of 21st ACM SIGGRAPH Conference on Visual Media Production, pp. 1–9 (2024). https://doi.org/10.1145/3697294.3697306

[94] Li, O., Hao, Y., Wang, Z., Zhu, B., Wang, S., Zhang, Z., Feng, F.: Model inversion attacks through

target-specific conditional diffusion models. arXiv preprint arXiv:2407.11424 (2024). https://doi.org/10.48550/arXiv.2407.11424

[95] Kandpal, N., Pillutla, K., Oprea, A., Kairouz, P., Choquette-Choo, C.A., Xu, Z.: User Inference Attacks on Large Language Models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 18238–18265 (2024). https://doi.org/10.18653/v1/2024.emnlp-main.1014

[96] Gan, Y., Miao, J., Yang, Y.: DataStealing: Steal Data from Diffusion Models in Federated Learning with Multiple Trojans. In Proceedings of the 38th International Conference on Neural Information Processing Systems, pp. 132614–132646 (2024)

[97] Wang, H., Li, H., Chen, H., Yan, J., Shi, J., Shen, J.: FedDifRC: Unlocking the Potential of Text-to-Image Diffusion Models in Heterogeneous Federated Learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3726–3736 (2025)

[98] Xiao, C., Chen, X., Liu, Y.: Confusion-resistant federated learning via diffusion-based data harmonization on non-IID data. Proceedings of the 38th International Conference on Neural Information Processing Systems, 137495–137520 (2024)

[99] Sun, H., Zhao, H., Xu, L., Zhang, R., Zhang, W., Jiang, W., Guan, H., Zhang, B.: Diffusion Generation-Based Federated Learning for Non-IID Defect Recognition. IEEE Transactions on Computational Social Systems, 1–14 (2025). https://doi.org/10.1109/TCSS.2025.3564325

[100] Hoefler, M.A., Mazouka, T., Mueller, K., Samek, W.: Boosting Federated Learning with Diffusion Models for Non-IID and Imbalanced Data. In 2024 IEEE International Conference on Big Data (BigData), pp. 7790–7799 (2024). https://doi.org/10.1109/BigData62323.2024.10825355

[101] Gu, H., Zhang, X., Li, J., Wei, H., Li, B., Huang, X.: Federated learning vulnerabilities: Privacy attacks with denoising diffusion probabilistic models. In Proceedings of the ACM Web Conference 2024, pp. 1149–1157 (2024). https://doi.org/10.1145/3589334.3645514

[102] Liu, B., Fang, Y.: Federated knowledge graph unlearning via diffusion model. arXiv preprint arXiv:2403.08554 (2024). https://doi.org/10.48550/arXiv.2403.08554

[103] Lai, J., Li, J., Xu, J., Wu, Y., Tang, B., Chen, S., Huang, Y., Ding, W., Li, Y.: pFedGPA: Diffusion-based Generative Parameter Aggregation for Personalized Federated Learning. In Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence, pp. 17999–18007 (2025). https://doi.org/10.1609/aaai.v39i17.33980

[104] Mei, Y., Yuan, L., Han, D.-J., Chan, K.S., Brinton, C.G., Lan, T.: Using Diffusion Models as Generative Replay in Continual Federated Learning–What will Happen? arXiv preprint arXiv:2411.06618 (2024). https://doi.org/10.48550/arXiv.2411.06618

[105] Li, G., Zhang, L., Rong, Q., Ding, X., Yuan, L.: FedDiffRec: A Module-wise Training Approach for Diffusion-Based Recommendation in Federated Learning. In 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2025). https://doi.org/10.1109/ICASSP49660.2025.10889647

[106] Zhou, Z., Luo, G., Chen, M., Weng, Z., Zhu, Y.: Federated Learning for Medical Image Classification: A Comprehensive Benchmark. arXiv preprint arXiv:2504.05238 (2025). https://doi.org/10.48550/arXiv.2504.05238

[107] Huangsuwan, K., Liu, T., See, S., Beng Ng, A., Vateekul, P.: FedDrip: Federated Learning With Diffusion-Generated Synthetic Image. IEEE Access **13**, 10111–10125 (2025). https://doi.org/10.1109/ACCESS.2025.3525806

[108] Wang, E., Liu, W., Liu, W., Xiang, C., Yang, B., Yang, Y.: Spatiotemporal Transformer for Data Inference and Long Prediction in Sparse Mobile Crowd-Sensing. In IEEE Conference on Computer Communications, pp. 1–10 (2023). https://doi.org/10.1109/INFOCOM53939.2023.10228982

[109] Huang, R., Cheng, L., Li, Z., Xiang, C., Guo, Y.: iPatrol: Illegal Roadside Parking Detection Leveraging On-road Vehicle Crowdsensing. In 2024 IEEE/ACM 32nd International Symposium on Quality of Service (IWQoS), pp. 1–10 (2024). https://doi.org/10.1109/IWQoS61813.2024.10682887

[110] Li, S., Xiang, C., Xu, W., Peng, J., Xu, Z., Li, J., Liang, W., Jia, X.: Coverage Maximization of Heterogeneous UAV Networks. In 2023 IEEE 43rd International Conference on Distributed Computing Systems (ICDCS), pp. 120–130 (2023). https://doi.org/10.1109/ICDCS57875.2023.00026

[111] Zhan, Z., Wang, Y., Duan, P., Sai, A.M.V.V., Liu, Z., Xiang, C., Tong, X., Wang, W., Cai, Z.: Enhancing worker recruitment in collaborative mobile crowdsourcing: A graph neural network trust evaluation approach. IEEE Transactions on Mobile Computing **23**(10), 10093–10110 (2024). https://doi.org/10.1109/TMC.2024.3373469

[112] Verbraeken, J., Wolting, M., Katzy, J., Kloppenburg, J., Verbelen, T., Rellermeyer, J.S.: A Survey on Distributed Machine Learning. ACM Computing Surveys **53**(2), 1–33 (2020). https://doi.org/10.1145/3377454

[113] Niu, R., Wu, W., Chen, J., Ma, L., Wu, Z.: A Multi-Stage Framework for Multimodal Controllable Speech Synthesis. arXiv preprint arXiv:2506.20945 (2025). https://doi.org/10.48550/arXiv.2506.20945