Confounding-Aware Client Selection in Federated Learning via Causal Mediation Analysis

Xiaoyang Yi[1,2,3]* **Yuru Bao**[1,2,3]* **Rihao Chang**[4] **Binhan Yang**[5] **Jian Zhang**[1,2,3,5]†
[1]College of Cryptology and Cyber Science, Nankai University
[2]Tianjin Key Laboratory of Network and Data Security Technology
[3]Key Laboratory of Data and Intelligent System Security, Ministry of Education
[4]School of Microelectronics, Tianjin University
[5]College of Computer Science, Nankai University
zhang.jian@nankai.edu.cn

Abstract

Federated Learning (FL) enables collaborative model training across clients while preserving data privacy. However, FL typically relies on voluntary client participation with uniform rewards, which can lead to high-quality clients dropping out due to inadequate incentives and low-quality clients taking a free-ride, both of which degrade overall model performance. However, existing incentive mechanisms fail to capture the true contribution value of a client's data to the global model, and often skew client selection toward cost compression rather than quality enhancement, thereby exacerbating confounding bias. To address this, we propose CausalAFL, a causal-based framework that reduces information asymmetry and corrects for confounding factors. Specifically, CausalAFL builds a complete causal chain connecting client information, bid, and marginal contribution. By jointly optimizing an inference network and a generative network within a variational inference framework, it precisely adjusts for causal bias. Maximizing the evidence lower bound objective allows CausalAFL to separate true contribution from noise and highlight the value of high-quality clients. A social welfare objective that quantifies both natural direct and indirect effects guides the server to prioritize these clients, resulting in improved global model accuracy and social welfare. Experiments on multiple datasets demonstrate that CausalAFL can guide clients to adjust their bids, leading to winners with higher contributions, confirming its effectiveness.

1 Introduction

Federated Learning (FL) (McMahan et al., 2017) enables collaborative model training across distributed clients while preserving data privacy (Chen & Vikalo, 2024; Ghari & Shen, 2024; Xu et al., 2025), thereby facilitating knowledge sharing and providing critical support for the integration of mobile computing and edge intelligence (Lei et al., 2026; Chen et al., 2026; Hou et al., 2026). Client selection stands as a core module in FL, which directly determines the quality of model aggregation and system efficiency. However, traditional FL relies on clients voluntarily uploading model parameters without differentiated rewards (Weng et al., 2021; Murhekar et al., 2023; Lin et al., 2024), lacking an effective incentive mechanism to compensate for the computational and communication resource consumption incurred by clients (Huang et al., 2024). This risks driving away high-quality clients due to inadequate incentives while allowing low-quality clients to free-ride, undermining both

*These authors contributed equally to this work.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

fairness and overall performance. To mitigate this issue, Auction-based FL (AFL) has emerged as a promising paradigm (Tang et al., 2024a), using market-driven mechanisms to select participating clients and distribute payments, thereby improving resource utilization and model performance (Tang & Yu, 2023; Tang et al., 2024b; Lu et al., 2024; Ren et al., 2024).

As illustrated in Figure. 1, the buyer acts as the global model's demander in AFL and initiates an auction task to the server, which is the auctioneer. Meanwhile, the sellers, who own personalized models, submit bids to this auctioneer (Zhang et al., 2021; Yang et al., 2024b). The auctioneer selects sellers to participate in FL based on the buyer's budget constraints, favoring lower bids, and determines payments according to contribution to the global model using rules such as the Vickrey-Clarke-Groves (VCG) (Deng et al., 2021; Yang et al., 2024b). The goal is to efficiently select sellers and determine payment prices while satisfying the buyer's budget constraints, thereby achieving optimal resource allocation (Li et al., 2025). This hinges on a key assumption that a seller's bid is an unbiased estimate of its cost, and that cost is monotonically related to the quality of the contributed model. However, a seller's bid cannot convey the potential contribution value of its data to the model. The lowest-bid-wins competitive logic shifts sellers' focus toward cost reduction rather than quality improvement, failing to create a healthy competitive market.

This confusion stems from unobserved confounding factors within the cost-bid-contribution pipeline, which disrupt the intended causal logic. When the server observes a high bid, it cannot discern whether it is attributable to a high-quality seller requiring substantial processing costs or merely from an inefficient, low-quality seller. This causal identification confusion leads the auction into a lemons market trap, where high-cost, high-quality sellers are driven out while low-cost, low-quality sellers dominate by strategic underbidding (Wadleigh et al., 2015; Chade & Schlee, 2020; Mamada, 2022), ultimately harming the performance ceiling of the global model. To fill this causal gap, a feasible solution is to incorporate a representation of the intrinsic causal structure among cost, bid and contribution into the auction process. Causal Mediation Analysis (CMA) (Xu et al., 2022, 2023; Lim et al., 2024), as a decomposition method based on a counterfactual framework, stands out as a fitting choice to quantify the causal transmission mechanism among variables through path intervention, and correct for bias by controlling the interference of confounding factors on the dual pathways.

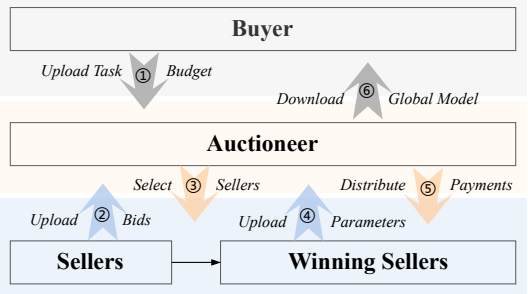


Figure 1: The framework of AFL. The buyer submits an auction request to a server acting as the auctioneer. Clients then participate as sellers in this auction, and the winners engage in FL to ultimately deliver the global model back to the buyer.

Building on these insights, we introduce **CausalAFL**, a **Causal**-based **AFL** framework, which extracts the client cost and converting it into information as the treatment variable, strategic bid as the mediator and marginal contribution as the outcome. Specifically, CausalAFL introduces dataset size and data usage cost as proxy confounders, then construct a generative causal model to estimate mediation pathways and correct bias. Under a variational inference framework, we jointly train an inference network and a generative network to maximize the Evidence Lower Bound Objective (ELBO), minimizing both reconstruction error and Kullback-Leibler (KL) divergence across all variables. By disentangling confounding effects, high-cost sellers can more accurately signal data quality and avoid being outbid by noise. CausalAFL further contains a social welfare objective driven by causal effects, quantifying Natural Direct Effects (NDE) and Natural Indirect Effects (NIE) to guide the server toward selecting high-quality clients and thus improve both global model performance. To summarize, our contributions are as follows:

- We propose CausalAFL, a novel causal-based framework for AFL that mitigates information asymmetry and confounding bias to enhance global model performance.
- We elaborately design a causal-guided bidding strategy that models the relationships among client information, bid, marginal contribution and incorporates latent confounders to infer key causal pathways.

- We develop a dual-network architecture combining inference and generative networks for modeling the cost-bid-contribution pathways, optimize it by maximizing the ELBO to correct multiple causal paths simultaneously.

- Experimental results demonstrate that CausalAFL can guide sellers to adjust their bids, leading to winners with higher contributions, showcasing its effectiveness.

2 Related Work

In recent years, AFL have been designed to balance multiple objectives, such as client participation, model performance and overall social welfare. Specifically, ACFL (Lu et al., 2023) applies mean-shift clustering to group clients by local data-distribution features, then employs an auction-based selection strategy driven by a composite cost function to dynamically filter participants. PrivAim (Wang et al., 2023) perturbs each client's local model gradients under local differential privacy, uses the exponential mechanism to obscure bid costs, and defines a marginal social welfare function to assess individual contributions. A_{FL} (Pang et al., 2023) decomposes social-cost minimization into a multi-stage winner-determination problem, selecting clients via a greedy strategy and critical-value payment rule. CSRA (Yang et al., 2024b) constructs a dual-layer dishonest-client detection and quality-maximization framework, integrating each client's past trustworthiness, local dataset size and privacy budget, and adjusts participation probabilities to exclude malicious nodes. The Long-Term Adaptive VCG Auction Mechanism (Wu et al., 2024a) employs Monte Carlo update strategies to optimize long-term cumulative reward, and proposes a multi-branch network architecture to enhance adaptability. MAGFL (Lu et al., 2024) normalizes client features for clustering into multiple groups, quantifies group training quality via trapezoidal membership functions, and defines a composite score to derive optimal bidding strategies for group agents. FPIN (Wang et al., 2025) dynamically evaluates client noise levels, selects clients with low noise and high contributions based on a combinatorial multi-armed bandit approach, designs a flexible pricing strategy to incentivize clients to proactively clean their noisy data. RAM (Li et al., 2025) employs a hybrid action reinforcement learning scheme to jointly learn user selection and payment outcomes, while each selected user maximizes its own utility by optimizing channel assignment, transmission power, and computation frequency.

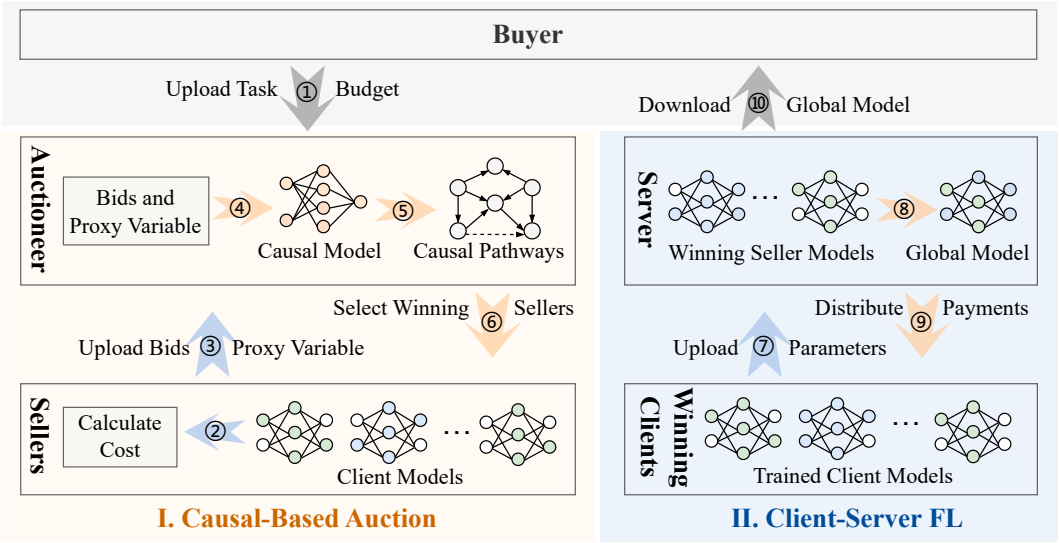


Figure 2: The overall framework of CausalAFL. Specifically, CausalAFL construct a generative causal graph model to estimate causal mediation pathways. By disentangling confounding effects, high-cost sellers can more accurately signal data quality and avoid being outbid by noise, thus improve both global model performance and social welfare.

3 Method

The framework of CausalAFL is illustrated in Figure 2. Specifically, CausalAFL simulates a market-driven competition, where the buyer submits a training request and budget constraint, the server acts as the auctioneer coordinating the auction process, and clients participate as sellers competing to join the federated training to obtain rewards. In each communication round, the auctioneer determines the winning sellers based on the submitted bids, the buyer's budget constraint, and the causal model, after which the winning sellers are then involved in the federated training process. Unlike traditional mechanisms that rely solely on static bid rankings, CausalAFL systematically models the causal relationships among seller information, bid, and marginal contribution. By introducing a generative network to estimate mediation pathways and adjust seller bidding strategies, CausalAFL improves budget-constrained social welfare and provides rational bids, achieving optimal model performance.

3.1 Architecture of the Causal-based AFL

Conventional FL typically assumes that clients participate by default and receive uniform rewards, overlooking the non-negligible resource consumption incurred by clients in practical deployments. Within CausalAFL, we explicitly model sellers with two types of costs and redesign the auction pipeline to align participation with each client's actual contribution to federated training. This prioritizes clients that provide greater improvements to the global model, thereby improving both model performance and social welfare.

3.1.1 Task Initiation and Bid Preparation

In the communication round t, the buyer first releases a training task along with a budget constraint to the auctioneer, after which each potential seller evaluates its cost and determines a corresponding bid. The cost of seller i is regarded as its total resource consumption for local training, which is defined from two aspects:

Computational Energy Cost: Local training is often conducted on resource-constrained devices such as mobile or edge platforms. The energy consumption of these devices significantly affects overall system efficiency. To reasonably estimate the computational energy cost of a seller, we adopt a classical energy model (Burd & Brodersen, 1996). The energy consumption of seller i in the communication round t is modeled as:

$$\text{Comp}_{i,t}^{s} = \beta_i \, c \, d_i / \delta_i^2 \tag{1}$$

where β_i denotes the effective switching capacitance of the seller's chip, c represents the number of CPU cycles per sample, d_i denotes the size of the local dataset, and δ_i is the CPU frequency.

Data Usage Cost: Local data distributions typically differ from the global data distribution, which is considered a key factor affecting the performance of the global model. Specifically, the closer the local distribution T_i^s is to the global data distribution T^s, the better aligned the local model is with the convergence direction of the global model, resulting in more beneficial training results during aggregation. Conversely, when the local data distribution deviates severely from the global data distribution, the aggregation performance may deteriorate. Therefore, the divergence between the two distributions is used to evaluate the data usage cost:

$$\text{Usage}_i^s = e\left(T_i^s, T^s\right) = \sqrt{\sum_{n=1}^{N} \left(T_{i,n}^s - T_n^s\right)^2} \tag{2}$$

where $e(*)$ is a predefined distribution distance function. We adopt L_2 distance as the default metric for efficient computation and generalization. $T_{i,n}^s$ is the proportion of class-n samples in seller i, and $T_n^s = 1/N$ represents the ideal proportion for balanced global data distribution, with N being the total number of classes. The divergence metric is computed locally by each seller, and the server collects the resulting data usage cost from sellers, thereby avoiding direct exposure of precise data distribution and preserving privacy.

In summary, the overall cost of seller i in the communication round t, incorporating both device heterogeneity and data distribution differences, is formulated as:

$$\text{Cost}_{i,t}^{s} = \text{Comp}_{i,t}^{s} \times t + \text{Usage}_i^s \tag{3}$$

where the computational energy cost is linearly accumulated over communication rounds due to repeated local training and the communication overhead, while the data usage cost depends on the deviation between the local data distribution and the global data distribution, remaining constant throughout the training process. The data usage cost is computed only once during the seller's first participation in the auction, then reused as a fixed cost in bid adjustment. The seller cost is also treated as an information variable that reflects both the computational energy and data usage.

Due to individual rationality assumptions in AFL, sellers may intentionally perturb their cost estimates to improve their winning probability or utility. Moreover, real-world auction environments may introduce social noise, further affecting actual bidding behavior. Thus, the bid of seller i in the communication round t is defined as a strategically adjusted form of its cost:

$$\text{bid}_{i,t}^{s} = \text{Cost}_{i,t}^{s} + \varepsilon_{i,t}, \qquad \varepsilon_{i,t} = \epsilon_{i,t} + \xi_{i,t} \tag{4}$$

Here, $\epsilon_{i,t}$ denotes a controllable strategic deviation and $\xi_{i,t}$ denotes uncontrollable market noise. The Gaussian model $\varepsilon_{i,t} \sim \mathcal{N}(1, \sigma^2)$ specifies the synthetic total perturbation.

3.1.2 Causal Selection and Model Training

In AFL, the cost of each seller directly influences its bidding behavior, which subsequently determines whether it is selected to participate in the global model update. Thus, a rational bidding mechanism is critical for enhancing the efficiency of AFL and ensuring fairness in seller incentives. However, previous work typically treats bid as an additive perturbation of cost, overlooking the potential complex causal dependencies among information, bid, and marginal contribution (Zhan et al., 2020; Deng et al., 2022). To address this gap, we transform the auctioneer from a passive bid-taker into an intelligent market designer, guiding sellers to generate more informative bids and addressing the causal blind spots in existing AFL methods when modeling client decision-making processes.

Based on causal estimation, the auctioneer refines each seller's submitted bid $\text{bid}_{i,t}^{s}$ to obtain the adjusted bid $\hat{\text{bid}}_{i,t}^{s}$, ranks the adjusted bids in ascending order, and selects the top K sellers as winners, forming the winner set W^t. After the winning sellers are determined, the server broadcasts the current global model to them for local training. Upon receiving the updated local models, the server computes the new global model through weighted aggregation. The global model ω_t^g in round t is updated as:

$$\omega_t^g = \sum_{i \in W^t} \frac{|D_i^s|}{\sum_{j \in W^t} |D_j^s|} \omega_{i,t}^s \tag{5}$$

where $\omega_{i,t}^s$ is the local model trained by the winning seller $i \in W^t$ and D_i^s represents the size of its local dataset. The aggregated global model is then delivered to the buyer as the training service result.

3.1.3 Utility Allocation and Welfare Realization

The server operates under the buyer's predefined budget constraint in each communication round t, which limits the total expenses available to reward sellers. Given this limitation, it is essential to optimize both the selection of winning sellers and the distribution of payments to maximize model performance and enhance social welfare. To this end, we define the utility of the buyer as:

$$\text{Utility}_t^b = \text{Acc}_t^b - \text{Budget} \tag{6}$$

where Acc_t^b is the accuracy of the global model. Budget is the predefined maximum expenses.

The utility of a winning seller i is defined as its actual payment minus training cost:

$$\text{Utility}_{i,t}^s = p_{i,t} - \text{Cost}_{i,t}^s \tag{7}$$

Here, $p_{i,t}$ denotes the actual payment from the server to the winning seller i in the communication round t. Social welfare in each communication round is defined as:

$$\text{Social-Welfare} = \text{Utility}_t^b + \sum_{i \in W^t} \text{Utility}_{i,t}^s = \text{Acc}_t^b - \sum_{i \in W^t} \text{Cost}_{i,t}^s \tag{8}$$

After training, the marginal contribution of each winning seller $i \in W^t$ is calculated as:

$$Con_{i,t} = \text{Social-Welfare}_{W^t \setminus \{i\}}(\psi^*) - \text{Social-Welfare}(\psi_{-i}^*) \tag{9}$$

where Social-Welfare$_{W^t \setminus \{i\}}(\psi^*)$ represents the social welfare that other sellers can achieve under the original optimal strategy after only removing the cost of seller i, and Social-Welfare(ψ^*_{-i}) represents the social welfare achieved by other sellers through re-optimization of the optimal strategy after removing seller i. The marginal contribution of the seller that does not win is 0.

To ensure budget efficiency, payments to winning sellers are distributed proportionally to their marginal contributions:

$$p_{i,t} = \frac{Con_{i,t}}{\sum\limits_{i \in \mathcal{W}^t} Con_{i,t}} \times \text{Budget} \tag{10}$$

In this way, we ensure that payments to winning sellers are allocated in a rational and contribution-aware manner. This incentivizes sellers to improve the quality of their personalized models to enhance competitiveness, leading to the selection of higher-contribution participants.

3.2 Causal Estimation and Bid Guidance

To overcome the mismatch between bid prices and true training value in AFL, we introduce CMA to quantify each seller's causal contribution to improving the global model. Specifically, CMA models the causal pathways linking seller information, bids, and marginal performance, and leverages proxy variables such as data size and usage cost to mitigate confounding effects. The resulting causal estimates are used to calibrate bids so that prices reflect both resource cost and expected performance gain. This transforms cost-driven bidding into causality-guided bidding, enabling the auctioneer to select sellers whose bids better reflect their contribution to the global model.

3.2.1 Confounder-Aware Variable Definition

To accurately capture the causal relationships behind AFL, we treat the involved factors as observable variables that are driven by underlying latent mechanisms. Specifically, the treatment variable is denoted by T, representing the resource consumption required to participate in training, as described in Eq. 3. The mediator variable is denoted by M, corresponding to the bid, which is obtained by adjusting the cost with a perturbation and reflects the seller's behavioral response, as in Eq. 4. The outcome variable is denoted by Y, referring to the marginal contribution that captures the improvement in global model accuracy brought by the seller after participating in the current communication round, according to Eq. 9.

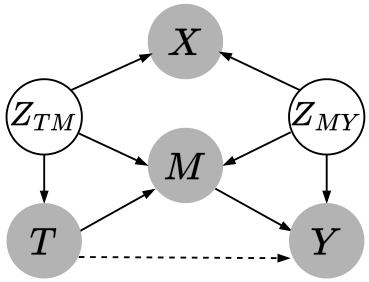


Figure 3: The partial mediation graph for CausalAFL, with both the direct path $T \to Y$ and the indirect path $T \to M \to Y$.

Additionally, the size of local dataset is directly related to both the bid and the cost, while the seller's data distribution directly affects the local model performance and thus influences the marginal contribution. That is, dataset size and data distribution act as confounders affecting the causal process, where dataset size influences both cost and bid, data distribution influences both bid and marginal contribution. However, directly accessing data distributions may lead to privacy concerns. To address this, we approximate the data distribution using the data usage cost derived from the divergence between the local and global data distributions.

According to (Angrist & Pischke, 2009), the observable proxy variable X is inherently related to the latent confounders. Therefore, we assume X_{TM} as the dataset size and X_{MY} as the data usage cost, respectively acting as proxies for the latent confounders Z_{TM} and Z_{MY}. These proxies provide information for path-specific latent confounding adjustment under the informativeness conditions. Figure 3 represents partial mediation: the direct path $T \to Y$ is retained alongside $T \to M \to Y$. Accordingly, the outcome model conditions on both T and M, and NDE captures effects that bypass the bid mediator.

3.2.2 ELBO-Driven Variational Inference

To model the above causal pathways, we construct a CMA model based on variational inference in the server, which consists of an inference network and a generative network. The inference network learns the distributions of two latent confounders from the proxy variables X_{TM} and X_{MY}, expressed as $q(Z_{TM}|X)$ and $q(Z_{MY}|X)$. Each is assumed to follow a Gaussian distribution, whose mean and variance parameters are generated by neural networks conditioned on the corresponding input.

$$q(Z_{TM}|X_{TM}) = \prod_{j=1}^{D_{Z_{TM}}} \mathcal{N}\left(\hat{\mu}_{TM_j},\ \hat{\sigma}^2_{TM_j}\right) \tag{11}$$

$$q(Z_{MY}|X_{MY}) = \prod_{j=1}^{D_{Z_{MY}}} \mathcal{N}\left(\hat{\mu}_{MY_j},\ \hat{\sigma}^2_{MY_j}\right) \tag{12}$$

where $\hat{\mu}_{TM_j}$ and $\hat{\mu}_{MY_j}$ represent the means, while $\hat{\sigma}^2_{TM_j}$ and $\hat{\sigma}^2_{MY_j}$ represent the variances of the Gaussian distributions parameterized by the inference network. $D_{Z_{TM}}$, $D_{Z_{MY}}$ denote the dimensions of Z_{TM} and Z_{MY}, respectively.

The generative network simulates variables generation pathways. Specifically, the latent variable Z_{TM} is first used to generate the seller information $\hat{T}$. Subsequently, conditioned on both $\hat{T}$ and Z_{TM}, the bid $\hat{M}$ is generated. Finally, the marginal contribution $\hat{Y}$, which reflects the impact of each seller on the global model performance, is generated based on $\hat{T}$, $\hat{M}$, and Z_{MY}. All variables are modeled as conditional Gaussian distributions. The generation processes of $\hat{T}$ and the proxy variable $\hat{X}$ are defined as follows:

$$p(\hat{T}|Z_{TM}) = \mathcal{N}(\mu_{TM}, \sigma^2_{TM}) \tag{13}$$

$$(\mu_{TM}, \sigma_{TM}) = \mathrm{NN}_T(Z_{TM}) \tag{14}$$

$$p(\hat{X}|Z_{TM}, Z_{MY}) = \prod_{j=1}^{D_{\hat{X}}} p(x_j|Z_{TM}, Z_{MY}) \tag{15}$$

where $\mathrm{NN}_T(*)$ denotes a neural network that outputs the mean and variance of the Gaussian distribution for $\hat{T}$, and $D_{\hat{X}}$ denotes the dimension of $\hat{X}$.

The generation processes of $\hat{M}$ and $\hat{Y}$ are calculated as follows:

$$p(\hat{M}|\hat{T}, Z_{TM}) = \mathcal{N}(\mu_{\hat{M}},\ \sigma_{\hat{M}^2}) \tag{16}$$

$$p(\hat{Y}|\hat{T}, \hat{M}, Z_{MY}) = \mathcal{N}(\mu_{\hat{Y}},\ \sigma_{\hat{Y}^2}) \tag{17}$$

Here, $\mu_{\hat{M}}$ and $\mu_{\hat{Y}}$ represent the means, while $\sigma_{\hat{M}^2}$ and $\sigma_{\hat{Y}^2}$ denote the variances of the Gaussian distributions parameterized by the generative network.

The prior distributions of the latent confounders Z_{TM} and Z_{MY} are modeled as standard normal distributions as follows:

$$p(Z_{TM}) = \prod_{j=1}^{D_{Z_{TM}}} \mathcal{N}(z_{TM_j}|0, 1) \tag{18}$$

$$p(Z_{MY}) = \prod_{j=1}^{D_{Z_{MY}}} \mathcal{N}(z_{MY_j}|0, 1) \tag{19}$$

This setting follows the classical framework of Variational Autoencoders (VAE) (Kingma et al., 2013), aiming to provide a unified and unbiased prior assumption for each latent factor, thereby guiding the learning of the posterior distribution to be more stable and controllable.

The server fits the generative model to the observed variables (X, T, M, Y) by minimizing reconstruction losses and KL penalties. Let $q = q(Z_{TM} \mid X_{TM})q(Z_{MY} \mid X_{MY})$. Writing the proxy-reconstruction part of the variational bound as ELBO_X, the objective is:

$$\begin{aligned} \mathrm{ELBO}_X = \mathbb{E}_q[\log p(X \mid Z_{TM}, Z_{MY})] - D_{\mathrm{KL}}(q(Z_{TM} \mid X_{TM}) \,\|\, p(Z_{TM})) \\ - D_{\mathrm{KL}}(q(Z_{MY} \mid X_{MY}) \,\|\, p(Z_{MY})) \end{aligned} \tag{20}$$

$$\begin{aligned} \mathcal{L}_{\mathrm{CMA}} = -\,\mathrm{ELBO}_X - \mathbb{E}_q\big[\log p(T \mid Z_{TM}) + \log p(M \mid T, Z_{TM}) \\ + \log p(Y \mid T, M, Z_{MY})\big] \end{aligned} \tag{21}$$

Minimizing $\mathcal{L}_{\mathrm{CMA}}$ is equivalent to maximizing the full variational lower bound for the observed tuple under this factorization. The likelihood terms are evaluated at the observations; the hatted quantities above describe generated predictions. The KL terms enter the minimized loss with positive signs and penalize posterior deviations from the priors.

3.2.3 ATE-Informed Bid Optimization

After training the CMA model, we can estimate the NDE, NIE_r, and Average Treatment Effect (ATE) by simulating the generative results under different variable pathways. NDE compares treatment levels while retaining the mediator distribution induced by the reference treatment; it includes pathways from T to Y that bypass M. NIE_r measures the indirect pathway using the reverse contrast defined below:

$$\begin{aligned}\text{NDE}(t_1, t_0) = \int \Big[\mathbb{E}(Y \mid T = t_1, M = m, Z_{MY}) - \mathbb{E}(Y \mid T = t_0, M = m, Z_{MY})\Big] \\ \cdot\ p(m \mid T = t_0, Z_{MY}, Z_{TM}) \cdot\ p(Z_{MY}, Z_{TM})\, dm\, dZ_{MY}\, dZ_{TM}\end{aligned} \tag{22}$$

$$\begin{aligned}\text{NIE}_r(t_1, t_0) = \int \mathbb{E}(Y \mid T = t_1, M = m, Z_{MY}) \cdot \Big[p(m \mid T = t_0, Z_{MY}, Z_{TM}) \\ - p(m \mid T = t_1, Z_{MY}, Z_{TM})\Big] \cdot p(Z_{MY}, Z_{TM})\, dm\, dZ_{MY}\, dZ_{TM}\end{aligned} \tag{23}$$

$$\text{ATE} = \text{NDE} - \text{NIE}_r \tag{24}$$

Here, t_1 and t_0 are two specific values of the processing variable T. $p(Z_{MY}, Z_{TM})$ denotes the joint distribution of the confounders.

Based on CMA, a causally driven bid adjustment mechanism is designed. In the communication round t, the bid of seller i is determined as the sum of its cost, a Gaussian perturbation term, and a causal guidance term:

$$\hat{\text{bid}}^s_{i,t} = \text{Cost}^s_{i,t} + \varepsilon_{i,t} + \text{ATE} \cdot t \tag{25}$$

This causally driven bidding mechanism enhances the marginal contribution level of the selected seller set W^t, thereby improving the global model performance Acc^b_t, and under a constraint budget, achieves social welfare maximization in each round of auction:

$$\max_{\{M_i\}} \left[\text{Acc}^b_t - \sum_{i \in W^t} \text{Cost}^s_{i,t}\right] \tag{26}$$

The server uses the causal estimates to calibrate bids for budget-constrained client selection.

4 Experiments

4.1 Experiment Setting

We evaluate CausalAFL on six benchmark datasets, including CIFAR-10 (Krizhevsky & Hinton, 2009), CIFAR-100 (Krizhevsky & Hinton, 2009), FashionMNIST (FMNIST) (Xiao et al., 2017), emnist-digits (EMNIST-D) (Cohen et al., 2017), emnist-letters (EMNIST-L) (Cohen et al., 2017), and Kuzushiji-MNIST (KMNIST) (Clanuwat et al., 2018). On CIFAR-10 and CIFAR-100 we use ResNet-18 (He et al., 2016) for image classification. On others we employ a convolutional neural network with two convolution layers and two fully connected layers. A variational autoencoder is trained to obtain disentangled representations of the confounding variables.

We simulate 20 clients as sellers in reverse auctions. In each communication round, 10 sellers win the auction and participate in FL under a total budget of 20 units. Each bid is affected by random noise ranging from the average bid of the clients to 50%. Seller data distributions are sampled from a Dirichlet distribution with concentration parameter α that controls heterogeneity. Smaller values of α produce greater heterogeneity. Each seller's computational cost is determined by 20 CPU cycles per bit of data, an effective switch capacitance chosen at random between 1×10^{-28} and 2×10^{-28}, a CPU frequency selected at random between 1 and 5 GHz. Training runs for 200 communication rounds with 5 local epochs each round. Each seller's causal network has 5 layers and is trained for 10 epochs. All client models use a learning rate of 0.1 and a batch size of 512. We compare CausalAFL against FedAvg (McMahan et al., 2017), Ditto (Li et al., 2021), FedAS (Yang et al., 2024a), FedDBE (Zhang et al., 2023), FedNTD (Lee et al., 2022) FedCAC (Wu et al., 2023), IAFL (Wu et al., 2024b), FuseFL (Tang et al., 2024c), K-FLA (Li et al., 2024), FedCIFL (Guo et al., 2025) and FedDif (Ahn et al., 2026). All experiments are conducted on a Tesla A100 GPU.

Table 1: Main accuracy results of CausalAFL on CIFAR-10, EMNIST-L, and KMNIST. The best results are shown in bold.

	CIFAR-10			EMNIST-L			KMNIST		
	$\alpha = 1.0$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 1.0$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 1.0$	$\alpha = 0.1$	$\alpha = 0.01$
FedAvg	53.44	48.81	30.93	53.04	41.81	14.39	90.21	74.44	39.45
Ditto	74.69	55.94	39.77	80.36	79.29	37.71	90.79	89.33	85.15
FedAS	72.06	64.40	50.61	86.86	62.17	60.05	84.17	78.03	72.38
FedDBE	48.14	44.92	42.59	87.95	80.03	76.55	95.77	91.65	90.95
FedNTD	46.02	39.14	42.26	63.07	51.12	37.49	91.36	84.78	64.99
FedCAC	71.60	53.54	50.65	88.97	71.35	55.66	94.63	87.94	85.81
IAFL	48.74	46.17	32.45	60.28	48.77	31.56	92.22	81.23	77.75
FuseFL	62.23	53.65	52.50	68.59	61.53	46.77	89.66	83.09	78.57
K-FLA	72.78	64.76	43.74	67.57	59.14	42.19	91.61	88.92	87.74
FedCIFL	58.70	54.25	51.80	63.15	70.15	48.00	85.10	84.40	81.80
FedDif	54.02	47.65	38.53	69.65	69.77	65.03	83.55	80.10	79.17
Ours	**78.22**	**67.75**	**54.68**	**89.26**	**81.24**	**78.06**	**96.31**	**92.05**	**91.15**

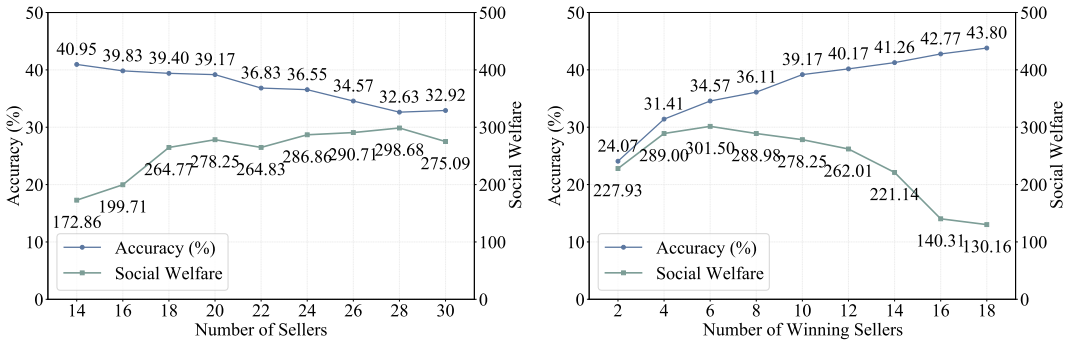
Table 2: Main accuracy results of CausalAFL on CIFAR-100, EMNIST-D, and FMNIST. The best results are shown in bold.

	CIFAR-100			EMNIST-D			FMNIST		
	$\alpha = 1.0$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 1.0$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 1.0$	$\alpha = 0.1$	$\alpha = 0.01$
FedAvg	29.30	16.42	12.68	84.24	69.03	63.78	71.01	48.83	42.13
Ditto	38.91	25.62	18.63	98.75	97.65	79.72	75.96	76.83	67.99
FedAS	48.05	37.79	19.15	98.00	81.73	73.38	88.03	85.82	75.21
FedDBE	25.24	21.64	19.18	98.93	96.78	71.41	88.47	85.67	85.85
FedNTD	21.18	16.77	18.86	95.49	90.01	75.55	74.96	69.21	46.56
FedCAC	53.61	38.03	16.67	99.21	97.41	80.95	89.03	85.30	79.59
IAFL	36.71	22.08	23.13	92.74	71.61	70.77	86.60	71.66	65.83
FuseFL	33.33	33.86	30.31	88.72	78.33	72.97	85.37	85.83	72.22
K-FLA	43.86	38.58	30.04	96.10	93.36	79.13	83.85	81.72	80.54
FedCIFL	24.65	21.15	22.60	89.65	76.51	73.50	85.75	84.70	79.85
FedDif	41.57	32.50	27.60	95.97	89.68	77.62	77.95	76.18	73.18
Ours	**54.46**	**42.12**	**39.17**	**99.30**	**98.06**	**82.22**	**89.05**	**87.11**	**86.64**

4.2 Main Results

We compare CausalAFL with baselines in Tables 1 and 2. As heterogeneity increases, the accuracy of FedAvg drops sharply, since it suffers from client drift when non-independent and identically distributed (non-IID) data cause local models to update in divergent directions. Ditto uses personalized fine-tuning to balance fairness and robustness, but it cannot guarantee that selected clients contribute most to the global model. FedAS is limited to internal model distributions and fails to stabilize performance at low α values, achieving only 19.15% accuracy on CIFAR-100 when $\alpha = 0.01$. FedDBE may select low-quality clients unreasonably and performs poorly on complex datasets such as CIFAR-10 and CIFAR-100. FedNTD cannot ensure that participating clients actually make contributions, resulting in only 46.56% and 37.49% accuracy on FMNIST and EMNIST-L when $\alpha = 0.01$. FedCAC cannot avoid choosing clients whose contributions do not match their bids, causing its accuracy on CIFAR-100 to fall from 53.61% to 16.67%.

In addition, compared to AFL methods such as FedDif and K-FLA, CausalAFL introduces a causally guided bidding adjustment mechanism, avoiding the dominance of low-quality clients, thereby selecting participants with higher contributions. Compared to IAFL, which includes incentive mechanisms, CausalAFL not only considers economic incentives but also quantifies the true impact of clients on model performance through causal effects, achieving precise alignment between incentives and contributions. Furthermore, compared to causal-based FL methods like FedCIFL and FuseFL, CausalAFL constructs a complete causal graph model linking cost, bid, and contribution, jointly optimizes inference and generative networks within a variational inference framework. Therefore, CausalAFL can guide the server to prioritize clients that contribute the most to global performance,



(a) Seller Numbers

(b) Winning Seller Numbers

Figure 4: Seller study results of CausalAFL with social welfare and test accuracy on CIFAR-10 and CIFAR-100 when $\alpha = 0.01$.

enabling it to consistently select high-quality clients under data heterogeneity, resulting in expressive outcomes across all datasets compared to baselines.

4.3 Seller Study

Seller Numbers. We explore the impact of the number of sellers on the results on CIFAR-100 with $\alpha = 0.01$, as shown in Figure 4a. When the total number of sellers increases from 14 to 30, the global model's test accuracy smoothly declines from approximately 40.95% to 32.92% while the social welfare value climbs from 172.86 to 275.09 before gently receding. As the pool of candidate sellers grows and the number of winners remains fixed, the server gains a broader selection set and can more easily identify high-contribution clients, which directly drives social welfare upward. At the same time, expanding the seller pool introduces more low-quality or noisy participants and, when these nodes are not filtered out early, their erroneous gradients gradually dilute the signals of high-quality clients. As a result, although social welfare generally rises, model accuracy inevitably suffers because part of the budget is allocated to sellers whose actual contributions are limited.

Winning Seller Numbers. We also examine how the number of winning sellers affects the results as shown in Figure 4b. Keeping the candidate pool fixed at 20, we vary the number of winners from 2 to 18 and observe that test accuracy climbs from 24.07% to 43.80% while social welfare first falls from 227.93 to 130.16 before surging to 301.50 at 14 winners and then sliding back to 288.98. When too few sellers win, each client receives a larger reward but the server cannot fully leverage the diversity of data contributions, resulting in low social welfare and accuracy. As the number of winners increases, more high-quality clients join and the global model benefits from greater data diversity and aggregate contributions, causing accuracy and social welfare to rise sharply. Beyond the optimal point. However, as the fixed budget becomes too diluted across many clients, individual incentives weaken, and noise-prone clients begin to degrade performance, thereby causing both social welfare and accuracy to decline. All in all, these experiments demonstrate that CausalAFL can finely tune the number of winners to balance social welfare and model accuracy under budget constraints.

5 Conclusion

This paper proposes a causal-based AFL framework, CausalAFL, which is designed to reduce information asymmetry and confounding bias. Specifically, CausalAFL models the causal pathways linking client information, bid, marginal performance, and it employs proxy variables such as data size and usage cost to correct for confounding effects. The proxy-informed model supports bid calibration under the stated causal identification assumptions. Simultaneously, CausalAFL further implements a dual architecture of inference and generative networks, along with a causality-guided bid adjustment strategy. Experimental results demonstrate that CausalAFL delivers consistently expressive outcomes across multiple datasets, confirming its effectiveness and robustness.

References

Ahn, S., Kim, S., Kwon, Y., Youn, J., Park, J., and Cho, S. Auction-guided model diffusion for communication-efficient federated learning on non-iid data. *Neural Networks*, 193:108066, 2026. doi: 10.1016/J.NEUNET.2025.108066. URL https://doi.org/10.1016/j.neunet.2025.108066.

Angrist, J. D. and Pischke, J.-S. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

Burd, T. D. and Brodersen, R. W. Processor design for portable systems. *Journal of VLSI signal processing systems for signal, image and video technology*, 13(2):203–221, 1996.

Chade, H. and Schlee, E. Insurance as a lemons market: Coverage denials and pooling. *J. Econ. Theory*, 189:105085, 2020. doi: 10.1016/J.JET.2020.105085. URL https://doi.org/10.1016/j.jet.2020.105085.

Chen, H. and Vikalo, H. Heterogeneity-guided client sampling: Towards fast and efficient non-iid federated learning. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/7886b9bafe76c52fd568db10ff9772df-Abstract-Conference.html.

Chen, Y., Wang, Z., Cai, H., Qin, Z., and Deng, S. Federated knowledge distillation using hierarchical reinforcement learning in resource-constrained iot edge-cloud computing environments. *IEEE Trans. Mob. Comput.*, 25(1):326–340, 2026. doi: 10.1109/TMC.2025.3591610. URL https://doi.org/10.1109/TMC.2025.3591610.

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. Deep learning for classical japanese literature. *CoRR*, abs/1812.01718, 2018. URL http://arxiv.org/abs/1812.01718.

Cohen, G., Afshar, S., Tapson, J., and van Schaik, A. EMNIST: extending MNIST to handwritten letters. In *2017 International Joint Conference on Neural Networks, IJCNN 2017, Anchorage, AK, USA, May 14-19, 2017*, pp. 2921–2926. IEEE, 2017. doi: 10.1109/IJCNN.2017.7966217.

Deng, Y., Lyu, F., Ren, J., Chen, Y., Yang, P., Zhou, Y., and Zhang, Y. FAIR: quality-aware federated learning with precise user incentive and model aggregation. In *40th IEEE Conference on Computer Communications, INFOCOM 2021, Vancouver, BC, Canada, May 10-13, 2021*, pp. 1–10. IEEE, 2021. doi: 10.1109/INFOCOM42981.2021.9488743. URL https://doi.org/10.1109/INFOCOM42981.2021.9488743.

Deng, Y., Lyu, F., Ren, J., Wu, H., Zhou, Y., Zhang, Y., and Shen, X. AUCTION: automated and quality-aware client selection framework for efficient federated learning. *IEEE Trans. Parallel Distributed Syst.*, 33(8):1996–2009, 2022. doi: 10.1109/TPDS.2021.3134647. URL https://doi.org/10.1109/TPDS.2021.3134647.

Ghari, P. M. and Shen, Y. Personalized federated learning with mixture of models for adaptive prediction and model fine-tuning. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/a7a6465b9344c5ecc691c02af40661a7-Abstract-Conference.html.

Guo, X., Yu, K., Cui, L., Yu, H., and Li, X. Federated causally invariant feature learning. In Walsh, T., Shah, J., and Kolter, Z. (eds.), *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pp. 16978–16986. AAAI Press, 2025. doi: 10.1609/AAAI.V39I16.33866. URL https://doi.org/10.1609/aaai.v39i16.33866.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778, 2016.

Hou, X., Wang, J., Du, J., Jiang, C., Ren, Y., and Niyato, D. Lightweight federated learning over wireless edge networks. *IEEE Trans. Mob. Comput.*, 25(1):300–312, 2026. doi: 10.1109/TMC.2025.3594102. URL https://doi.org/10.1109/TMC.2025.3594102.

Huang, J., Ma, B., Wu, Y., Chen, Y., and Shen, X. A hierarchical incentive mechanism for federated learning. *IEEE Trans. Mob. Comput.*, 23(12):12731–12747, 2024. doi: 10.1109/TMC.2024.3423399. URL https://doi.org/10.1109/TMC.2024.3423399.

Kingma, D. P., Welling, M., et al. Auto-encoding variational bayes, 2013.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Lee, G., Jeong, M., Shin, Y., Bae, S., and Yun, S. Preservation of the global knowledge by not-true distillation in federated learning. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/fadec8f2e65f181d777507d1df69b92f-Abstract-Conference.html.

Lei, S., Tang, H., Li, C., Zhang, X., Xu, C., and Wu, H. Federated maddpg-based collaborative scheduling strategy in vehicular edge computing. *IEEE Trans. Mob. Comput.*, 25(1):54–66, 2026. doi: 10.1109/TMC.2025.3590747. URL https://doi.org/10.1109/TMC.2025.3590747.

Li, G., Cai, J., Lu, J., and Chen, H. Incentive mechanism design for cross-device federated learning: A reinforcement auction approach. *IEEE Trans. Mob. Comput.*, 24(4):3059–3075, 2025. doi: 10.1109/TMC.2024.3508260. URL https://doi.org/10.1109/TMC.2024.3508260.

Li, T., Hu, S., Beirami, A., and Smith, V. Ditto: Fair and robust federated learning through personalization. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6357–6368. PMLR, 2021. URL http://proceedings.mlr.press/v139/li21h.html.

Li, Z., Ding, B., Yao, L., Li, Y., Xiao, X., and Zhou, J. Performance-based pricing for federated learning via auction. *Proceedings of the VLDB Endowment*, 17(6):1269–1282, 2024.

Lim, S. S., Udomcharoenchaikit, C., Limkonchotiwat, P., Chuangsuwanich, E., and Nutanong, S. Identifying and mitigating annotation bias in natural language understanding using causal mediation analysis. In Ku, L., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 11548–11563. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.686. URL https://doi.org/10.18653/v1/2024.findings-acl.686.

Lin, Y., Gao, Z., Du, H., Niyato, D., Kang, J., and Liu, X. Incentive and dynamic client selection for federated unlearning. In Chua, T., Ngo, C., Kumar, R., Lauw, H. W., and Lee, R. K. (eds.), *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pp. 2936–2944. ACM, 2024. doi: 10.1145/3589334.3645462. URL https://doi.org/10.1145/3589334.3645462.

Lu, R., Zhang, W., Wang, Y., Li, Q., Zhong, X., Yang, H., and Wang, D. Auction-based cluster federated learning in mobile edge computing systems. *IEEE Trans. Parallel Distributed Syst.*, 34(4):1145–1158, 2023. doi: 10.1109/TPDS.2023.3240767. URL https://doi.org/10.1109/TPDS.2023.3240767.

Lu, R., Yang, H., Wang, Y., He, H., Li, Q., Zhong, X., and Zhang, W. Multi-attribute auction-based grouped federated learning. *IEEE Trans. Serv. Comput.*, 17(3):1056–1071, 2024. doi: 10.1109/TSC.2024.3387734. URL https://doi.org/10.1109/TSC.2024.3387734.

Mamada, R. The market for lemons and information theory. *Math. Soc. Sci.*, 120:107–112, 2022. doi: 10.1016/J.MATHSOCSCI.2022.10.002. URL https://doi.org/10.1016/j.mathsocsci.2022.10.002.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, 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, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017.

Murhekar, A., Yuan, Z., Chaudhury, B. R., Li, B., and Mehta, R. Incentives in federated learning: Equilibria, dynamics, and mechanisms for welfare maximization. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/39b77b5e422b4e070e2811b73ea9bcf7-Abstract-Conference.html.

Pang, J., Yu, J., Zhou, R., and Lui, J. C. S. An incentive auction for heterogeneous client selection in federated learning. *IEEE Trans. Mob. Comput.*, 22(10):5733–5750, 2023. doi: 10.1109/TMC.2022.3182876. URL https://doi.org/10.1109/TMC.2022.3182876.

Pearl, J. Interpretation and identification of causal mediation. *Psychological methods*, 19(4):459, 2014.

Ren, K., Liao, G., Ma, Q., and Chen, X. Differentially private auction design for federated learning with non-iid data. *IEEE Trans. Serv. Comput.*, 17(5):2236–2247, 2024. doi: 10.1109/TSC.2023.3332102. URL https://doi.org/10.1109/TSC.2023.3332102.

Tang, X. and Yu, H. Competitive-cooperative multi-agent reinforcement learning for auction-based federated learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pp. 4262–4270. ijcai.org, 2023. doi: 10.24963/IJCAI.2023/474. URL https://doi.org/10.24963/ijcai.2023/474.

Tang, X., Yu, H., Li, X., and Kraus, S. Intelligent agents for auction-based federated learning: A survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pp. 8253–8261. ijcai.org, 2024a. URL https://www.ijcai.org/proceedings/2024/912.

Tang, X., Yu, H., Li, Z., and Li, X. A bias-free revenue-maximizing bidding strategy for data consumers in auction-based federated learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pp. 4991–4999. ijcai.org, 2024b. URL https://www.ijcai.org/proceedings/2024/552.

Tang, Z., Zhang, Y., Dong, P., Cheung, Y., Zhou, A. C., Han, B., and Chu, X. Fusefl: One-shot federated learning through the lens of causality with progressive model fusion. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024c. URL http://papers.nips.cc/paper_files/paper/2024/hash/31e6e0c09325a3be16d93f84e40e0c7e-Abstract-Conference.html.

Wadleigh, J., Drew, J., and Moore, T. The e-commerce market for "lemons": Identification and analysis of websites selling counterfeit goods. In Gangemi, A., Leonardi, S., and Panconesi, A. (eds.), *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, pp. 1188–1197. ACM, 2015. doi: 10.1145/2736277.2741658. URL https://doi.org/10.1145/2736277.2741658.

Wang, D., Ren, J., Wang, Z., Wang, Y., and Zhang, Y. Privaim: A dual-privacy preserving and quality-aware incentive mechanism for federated learning. *IEEE Trans. Computers*, 72(7):1913–1927, 2023. doi: 10.1109/TC.2022.3230904. URL https://doi.org/10.1109/TC.2022.3230904.

Wang, H., Chen, H., Ma, M., and Cui, L. Dealing with noisy data in federated learning: An incentive mechanism with flexible pricing. In Long, G., Blumestein, M., Chang, Y., Lewin-Eytan, L., Huang, Z. H., and Yom-Tov, E. (eds.), *Proc. of WWW*, pp. 4432–4441, Sydney, NSW, Australia, 2025. ACM. doi: 10.1145/3696410.3714961. URL https://doi.org/10.1145/3696410.3714961.

Weng, J., Weng, J., Huang, H., Cai, C., and Wang, C. Fedserving: A federated prediction serving framework based on incentive mechanism. In *40th IEEE Conference on Computer Communications, INFOCOM 2021, Vancouver, BC, Canada, May 10-13, 2021*, pp. 1–10. IEEE, 2021. doi: 10.1109/INFOCOM42981.2021.9488807. URL https://doi.org/10.1109/INFOCOM42981.2021.9488807.

Wu, L., Guo, S., Hong, Z., Liu, Y., Xu, W., and Zhan, Y. Long-term adaptive VCG auction mechanism for sustainable federated learning with periodical client shifting. *IEEE Transactions on Mobile Computing*, 23(5):6060–6073, 2024a. doi: 10.1109/TMC.2023.3317063. URL https://doi.org/10.1109/TMC.2023.3317063.

Wu, X., Liu, X., Niu, J., Zhu, G., and Tang, S. Bold but cautious: Unlocking the potential of personalized federated learning through cautiously aggressive collaboration. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 19318–19327. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01775. URL https://doi.org/10.1109/ICCV51070.2023.01775.

Wu, Z., Amiri, M. M., Raskar, R., and Low, B. K. H. Incentive-aware federated learning with training-time model rewards. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024b. URL https://openreview.net/forum?id=FlY7WQ2hWS.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017. URL http://arxiv.org/abs/1708.07747.

Xu, S., Liu, L., and Liu, Z. Deepmed: Semiparametric causal mediation analysis with debiased deep learning. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b57939005a3cbe40f49b66a0efd6fc8c-Abstract-Conference.html.

Xu, Y., Zhu, Y., Wang, Z., Xu, H., and Liao, Y. Enhancing federated learning through layer-wise aggregation over non-iid data. *IEEE Trans. Serv. Comput.*, 18(2):798–811, 2025. doi: 10.1109/TSC.2025.3536309. URL https://doi.org/10.1109/TSC.2025.3536309.

Xu, Z., Cheng, D., Li, J., Liu, J., Liu, L., and Wang, K. Disentangled representation for causal mediation analysis. In Williams, B., Chen, Y., and Neville, J. (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 10666–10674. AAAI Press, 2023. doi: 10.1609/AAAI.V37I9.26266. URL https://doi.org/10.1609/aaai.v37i9.26266.

Yang, X., Huang, W., and Ye, M. Fedas: Bridging inconsistency in personalized federated learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 11986–11995. IEEE, 2024a. doi: 10.1109/CVPR52733.2024.01139. URL https://doi.org/10.1109/CVPR52733.2024.01139.

Yang, Y., Hu, M., Zhou, Y., Liu, X., and Wu, D. CSRA: robust incentive mechanism design for differentially private federated learning. *IEEE Trans. Inf. Forensics Secur.*, 19:892–906, 2024b. doi: 10.1109/TIFS.2023.3329441. URL https://doi.org/10.1109/TIFS.2023.3329441.

Zhan, Y., Li, P., Qu, Z., Zeng, D., and Guo, S. A learning-based incentive mechanism for federated learning. *IEEE Internet of Things Journal*, 7(7):6360–6368, 2020.

Zhang, J., Wu, Y., and Pan, R. Incentive mechanism for horizontal federated learning based on reputation and reverse auction. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pp. 947–956. ACM / IW3C2, 2021. doi: 10.1145/3442381.3449888. URL https://doi.org/10.1145/3442381.3449888.

Zhang, J., Hua, Y., Cao, J., Wang, H., Song, T., Xue, Z., Ma, R., and Guan, H. Eliminating domain bias for federated learning in representation space. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual*

Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/2e0d3c6ad1a4d85bef3cfe63af58bc76-Abstract-Conference.html.

A Theoretical Analysis

A.1 Analysis on Causal Effects

Theorem 1. *Conditional Model-Based Identification of Causal Mediation Effects. Under Assumptions 1–3, the NDE and NIE$_r$ admit the mediation functionals in Eqs. 22 and 23 within the specified latent variable model. Their estimation depends on the model and proxy informativeness conditions; the result does not establish identification from arbitrary proxies or unique recovery of the latent variables (Pearl, 2014).*

Assumption 1. *Latent Mediation Identification Conditions. Let $Z = (Z_{TM}, Z_{MY})$ denote pre-treatment latent adjustment variables. Consistency and positivity hold for the treatment levels being compared, with mediator support under the reference treatment contained in the support needed to evaluate the outcome kernel at both treatment levels. For all such t, t' and m, the potential outcomes satisfy:*

$$\{Y(t,m), M(t')\} \perp T \mid Z, \qquad Y(t,m) \perp M(t') \mid Z \tag{27}$$

The second condition is a cross-world independence assumption for natural effects. These conditions require sufficient latent adjustment and exclude an unaccounted treatment-induced mediator–outcome confounder. They do not require the observed variables T and M, or Y and (T, M), to be conditionally independent, and therefore preserve both causal paths.

Assumption 2. *Proxy Informativeness Assumption. The observed proxies $X = \{X_{TM}, X_{MY}\}$, representing dataset size and data usage cost, retain information relevant to the latent adjustment variables in the specified model. We assume the model kernels needed for the mediation functional can be estimated from (X, T, M, Y), with the variational approximation:*

$$q(Z_{TM}, Z_{MY} \mid X) \approx p(Z_{TM}, Z_{MY} \mid T, M, Y, X) \tag{28}$$

This is a modeling and inference condition, not a proof that the two proxies capture every latent common cause. Approximate posterior inference can introduce estimation error; the approximation does not itself establish exact identification or unique latent parameter recovery.

Assumption 3. *Latent Causal Factorization Assumption. The specified causal model has the factorization*

$$p(X, Z, T, M, Y) = p(X, Z)\, p(T \mid Z_{TM}) \cdot p(M \mid T, Z_{TM}, Z_{MY})\, p(Y \mid T, M, Z_{MY}) \tag{29}$$

where $p(X, Z) = p(X)p(Z \mid X)$, and the corresponding causal mechanisms remain invariant under interventions on T and M. The outcome kernel retains direct dependence on T. The mediator decoder $p(M \mid T, Z_{TM})$ used in the main text is a restricted instance of the mediator kernel above, imposing no additional dependence on Z_{MY} given (T, Z_{TM}). The population-level identity requires the chosen kernels to correctly represent the relevant data-generating process; approximation error affects empirical estimates.

Proof. For treatment levels t, t', define the nested counterfactual mean $\Psi(t, t') = \mathbb{E}[Y(t, M(t'))]$. Conditioning on the same latent variables in both kernels, Assumption 1 gives:

$$\begin{aligned} \Psi(t,t') &= \int \mathbb{E}[Y(t,m) \mid Z = z]\, p(M(t') = m \mid Z = z)\, p(z)\, dm\, dz \\ &= \int \mathbb{E}[Y \mid T = t, M = m, Z = z]\, p(m \mid T = t', Z = z)\, p(z)\, dm\, dz \end{aligned} \tag{30}$$

For continuous variables, the notation for the mediator denotes its conditional density. Consistency and treatment exchangeability identify the mediator kernel, while the cross-world condition and exchangeability identify the outcome kernel. Assumption 3 then yields:

$$\begin{aligned} \Psi(t,t') = \int & \mathbb{E}[Y \mid T = t, M = m, Z_{MY} = z_{MY}] \\ & \cdot p(m \mid T = t', Z_{TM} = z_{TM}, Z_{MY} = z_{MY})\, p(z_{TM}, z_{MY})\, dm\, dz_{TM}\, dz_{MY} \end{aligned} \tag{31}$$

The natural effects follow by subtraction:

$$\text{NDE}(t_1, t_0) = \Psi(t_1, t_0) - \Psi(t_0, t_0) \tag{32}$$

$$\text{NIE}_r(t_1, t_0) = \Psi(t_1, t_0) - \Psi(t_1, t_1) \tag{33}$$

Substituting the expression for Ψ gives Eqs. 22 and 23. The distribution $p(Z)$ can equivalently be written as $\int p(Z \mid X)p(X)\,dX$ within the model. Assumption 2 supplies the proxy-based estimation premise for these kernels. Thus, the result is a conditional model-based representation of the mediation effects, rather than a nonparametric guarantee independent of proxy quality.

The shared integration over Z preserves the dependence between the mediator and outcome induced by latent common causes. In particular, $\int p(m \mid T = t, Z = z)p(z)\,dz$ describes the mediator under intervention on T under these assumptions; it must not be substituted for the observational conditional distribution $p(m \mid T = t)$. Dataset size and data usage cost offer complementary information about training scale and data heterogeneity, but additional common causes outside the treatment and proxy coverage can invalidate the identification conditions. □

A.2 Analysis on Auction

Theorem 2. *Conditional Individual Rationality. Suppose the winner set $\mathcal{W}^t$ is nonempty and $Con_{j,t} > 0$ for every $j \in \mathcal{W}^t$. Under the proportional payment rule, a winning seller i has nonnegative utility if:*

$$\textit{Budget} \geq \textit{Cost}^s_{i,t} \frac{\sum_{j \in \mathcal{W}^t} Con_{j,t}}{Con_{i,t}} \tag{34}$$

If this sufficient budget condition holds for every winner, individual rationality holds throughout the winner set.

Proof. The payment rule gives:

$$p_{i,t} = \frac{Con_{i,t}}{\sum_{j \in \mathcal{W}^t} Con_{j,t}} \text{Budget} \tag{35}$$

Both the numerator and denominator of the contribution share are positive. Multiplying Eq. 34 by that share and using Eq. 35 therefore preserves the inequality and yields:

$$p_{i,t} \geq \text{Cost}^s_{i,t}, \qquad \text{Utility}^s_{i,t} = p_{i,t} - \text{Cost}^s_{i,t} \geq 0 \tag{36}$$

The budget condition is a sufficient feasibility premise, not a consequence of selecting the lowest adjusted bids. A winner set containing a nonpositive contribution lies outside the positive-contribution domain analyzed here, and a zero total contribution makes the proportional formula undefined. The theorem does not establish individual rationality in those cases or prescribe a different selection or payment rule. □

Theorem 3. *Causally Guided Bid Consistency. For a fixed learned distribution, let δ denote a posterior mean displacement induced by a bid deviation in one latent coordinate. This displacement is distinct from the strategic bid component $\epsilon_{i,t}$. Let:*

$$q_\delta = \mathcal{N}(\mu_q + \delta, \sigma_q^2), \qquad p = \mathcal{N}(\mu_p, \sigma_p^2) \tag{37}$$

where the means and positive variances are fixed in this conditional analysis. The server-side consistency penalty $R(\delta) = D_{\text{KL}}(q_\delta \| p)$ is strictly convex in the centered residual $r = \mu_q + \delta - \mu_p$ and is uniquely minimized at $r = 0$, equivalently $\delta^* = \mu_p - \mu_q$.

Proof. The complete Gaussian KL divergence is:

$$R(\delta) = \frac{1}{2}\left[\frac{\sigma_q^2 + (\mu_q + \delta - \mu_p)^2}{\sigma_p^2} - 1 + \log\frac{\sigma_p^2}{\sigma_q^2}\right] \tag{38}$$

Consequently,

$$\frac{\partial R}{\partial \delta} = \frac{\mu_q + \delta - \mu_p}{\sigma_p^2}, \qquad \frac{\partial^2 R}{\partial \delta^2} = \frac{1}{\sigma_p^2} > 0 \tag{39}$$

The stationary displacement is $\mu_p - \mu_q$, and the increase relative to the minimum is:

$$R(\delta) - R(\delta^*) = \frac{r^2}{2\sigma_p^2} \geq 0 \tag{40}$$

The result extends coordinatewise to the diagonal Gaussian posteriors by summing these penalties.

The calculation describes the consistency penalty for a posterior displacement induced by an inconsistent bid; it does not assume that every bid deviation produces such a displacement. Random market noise is not an optimization variable in this fixed-distribution calculation. The result concerns the KL term, not the complete reconstruction objective or a bound on changes in the final ranking. In particular, a zero centered residual need not imply a zero strategic bid deviation.

Client utility is a separate object. With b_i denoting client i's submitted bid and b_{-i} the other bids, it is:

$$U_{i,t}(b_i, b_{-i}) = a_{i,t}(b_i, b_{-i}) \left[p_{i,t}(\mathcal{W}^t(b_i, b_{-i})) - \text{Cost}_{i,t}^s \right] \tag{41}$$

where $a_{i,t} \in \{0, 1\}$ indicates selection by adjusted-bid ranking. Dominant-strategy incentive compatibility would require:

$$U_{i,t}(\text{Cost}_{i,t}^s, b_{-i}) \geq U_{i,t}(\widetilde{b}_i, b_{-i}) \tag{42}$$

The KL penalty does not enter this client utility directly and does not prove the inequality. The current proportional payment rule does not by itself exclude a profitable unilateral bid deviation. Theorem 3 therefore provides a conditional server-side calibration property, without asserting truthful bidding as a best response or a dominant strategy. □

Theorem 4. *Budget Feasibility. For a nonempty winner set in the positive-contribution domain of Theorem 2, the proportional payments exhaust the budget:* $\sum_{i \in \mathcal{W}^t} p_{i,t} = \textit{Budget}$.

Proof. By construction of the payment rule:

$$\begin{aligned} \sum_{i \in \mathcal{W}^t} p_{i,t} &= \sum_{i \in \mathcal{W}^t} \left(\frac{Con_{i,t}}{\sum_{j \in \mathcal{W}^t} Con_{j,t}} \times \text{Budget} \right) \\ &= \frac{\text{Budget}}{\sum_{j \in \mathcal{W}^t} Con_{j,t}} \times \sum_{i \in \mathcal{W}^t} Con_{i,t} \\ &= \text{Budget} \end{aligned} \tag{43}$$

Therefore, the budget is feasible. □

A.3 Analysis on Convergence

Theorem 5. *Conditional Convergence under Controlled Residual Selection Bias. Under Assumptions 4–7, the update* $w_{t+1} = w_t - \eta_t g_t$ *from a fixed initial point satisfies:*

$$\lim_{T \to \infty} \frac{\sum_{t=0}^{T-1} \eta_t \mathbb{E} \|\nabla F(w_t)\|^2}{\sum_{t=0}^{T-1} \eta_t} = 0, \qquad \liminf_{t \to \infty} \mathbb{E} \|\nabla F(w_t)\|^2 = 0 \tag{44}$$

These conclusions express asymptotic first-order stationarity; they do not assert convergence of the entire parameter sequence or unconditional elimination of selection bias.

Assumption 4. *L-smoothness. The global objective* $F(w) = \mathbb{E}_i[f_i(w)]$ *satisfies* $F(w') \leq F(w) + \nabla F(w)^\top (w' - w) + \frac{L}{2} \|w' - w\|^2$ *for all* $w, w' \in \mathbb{R}^d$.

Assumption 5. *Controlled Residual Selection Bias. Let* $\mathcal{F}_t$ *denote the training history before the round-*t *update, so that* w_t *is* $\mathcal{F}_t$*-measurable, and let* $b_t = g_t - \nabla F(w_t)$. *Almost surely,*

$$\|\mathbb{E}[b_t \mid \mathcal{F}_t]\| \leq B_t, \qquad \mathbb{E}[\|b_t\|^2 \mid \mathcal{F}_t] \leq \sigma_b^2, \qquad \sum_{t=0}^{\infty} B_t^2 < \infty \tag{45}$$

where $B_t \geq 0$ *is deterministic and* $\sigma_b^2 < \infty$. *This condition permits nonzero transient bias but requires its conditional mean to decay sufficiently, and a uniform constant bias bound alone is insufficient.*

Assumption 6. *Step Size Rule. The step size is* $\eta_t = c/(t+1)$ *with* $0 < c < 1/L$. *Hence* $1 - L\eta_t \geq \rho := 1 - Lc > 0$, $\sum_t \eta_t = \infty$, *and* $\sum_t \eta_t^2 < \infty$.

Assumption 7. *Lower-bounded Objective. There exists* $F^* > -\infty$ *such that* $F(w) \geq F^*$ *for all* w, *and* $F(w_0) < \infty$.

Proof. Smoothness and conditional expectation yield:

$$\mathbb{E}[F(w_{t+1}) \mid \mathcal{F}_t] \leq F(w_t) - \eta_t \nabla F(w_t)^\top \mathbb{E}[g_t \mid \mathcal{F}_t] + \frac{L\eta_t^2}{2} \mathbb{E}[\|g_t\|^2 \mid \mathcal{F}_t] \tag{46}$$

Substituting $g_t = \nabla F(w_t) + b_t$, applying $\|u+v\|^2 \leq 2\|u\|^2 + 2\|v\|^2$, and using Assumption 5 gives:

$$\mathbb{E}[F(w_{t+1}) \mid \mathcal{F}_t] \leq F(w_t) - \eta_t(1 - L\eta_t)\|\nabla F(w_t)\|^2 + \eta_t B_t \|\nabla F(w_t)\| + L\eta_t^2 \sigma_b^2 \tag{47}$$

Young's inequality absorbs the mixed term:

$$\eta_t B_t \|\nabla F(w_t)\| \leq \frac{\eta_t(1 - L\eta_t)}{2} \|\nabla F(w_t)\|^2 + \frac{\eta_t B_t^2}{2(1 - L\eta_t)} \tag{48}$$

Taking full expectation and summing from $t = 0$ to $T - 1$ therefore yields:

$$\frac{1}{2} \sum_{t=0}^{T-1} \eta_t(1 - L\eta_t) \mathbb{E}\|\nabla F(w_t)\|^2 \leq F(w_0) - \mathbb{E}[F(w_T)] + \frac{1}{2} \sum_{t=0}^{T-1} \frac{\eta_t B_t^2}{1 - L\eta_t} + L\sigma_b^2 \sum_{t=0}^{T-1} \eta_t^2 \tag{49}$$

The right-hand side is uniformly bounded: $F(w_T) \geq F^*$, $\eta_t \leq c$, $1 - L\eta_t \geq \rho > 0$, and the series in Assumptions 5 and 6 are finite. Consequently,

$$\frac{\rho}{2} \sum_{t=0}^{T-1} \eta_t \mathbb{E}\|\nabla F(w_t)\|^2 \leq F(w_0) - F^* + \frac{c}{2\rho} \sum_{t=0}^{\infty} B_t^2 + L\sigma_b^2 \sum_{t=0}^{\infty} \eta_t^2 < \infty \tag{50}$$

Dividing by $\sum_{t=0}^{T-1} \eta_t \to \infty$ proves the weighted-average limit. Non-negativity then implies the stated $\liminf$ conclusion.

Assumption 5 is a premise of the optimization analysis, not an unconditional consequence of the causal model. The adjusted bids influence the selected set and hence the residual bias, but this theorem characterizes training only when that bias satisfies the stated conditional bounds. Evidence that causal calibration reduces selection bias and a proof of optimization under controlled bias are distinct forms of support. □

B Supplemental Experiments

Table 3: Ablation results of various causal training epochs in CausalAFL on CIFAR-10 with $\alpha = 0.01$.

	Acc.	Welf.	NDE	NIE_r	ATE
5 Epochs	52.12	354.74	0.3636	0.1270	0.4906
10 Epochs	54.68	399.54	0.2015	1.1204	1.3219
15 Epochs	52.03	382.24	0.2156	0.9291	1.1448
20 Epochs	51.55	380.57	0.3761	0.7146	1.0907
25 Epochs	50.75	377.12	0.9923	0.0758	1.0681

B.1 Ablation Study

Causal Epoch Ablation. We explore the effect of training the causal model on the server for different numbers of epochs, as reported in Table 3. When the server trains for only 5 epochs, the NDE is 0.3636 and the NIE_r is only 0.1270, showing that the indirect path remains under-trained. Increasing training to 10 epochs brings a significant boost, the causal model both fully amplifies the bid-to-contribution indirect path and controls the direct effect, achieving the best balance of accuracy and social welfare. Further extending training to 15, 20 and 25 epochs causes both accuracy and welfare

Table 4: Ablation results of various causal layers in CausalAFL on CIFAR-10 with $\alpha = 0.01$.

	Acc.	Welf.	NDE	NIE_r	ATE
0 Layer	51.34	388.44	-	-	-
3 Layers	51.35	370.82	0.5470	0.9727	1.0137
4 Layers	53.34	386.14	0.2571	1.0176	1.2748
5 Layers	54.68	399.54	0.2015	1.1204	1.3219
6 Layers	53.06	383.21	0.4560	0.7641	1.2201
7 Layers	52.18	381.12	0.7130	0.3765	1.0894

to decline. Although social welfare briefly rebounds at 20 epochs, over-training leads to dramatic increases in NDE and NIE_r, indicating that the causal model begins to overfit noise or redundant paths. Therefore, 10 epochs of server-side training are sufficient to learn the causal paths needed for confounding adjustment most effectively, maximizing the use of indirect paths while keeping the total effect reasonable and yielding optimal results.

Causal Layer Ablation. We compare the impact of varying the number of causal layers, where each layer comprises both an inference network and a generative network, as shown in Table 4. When no causal layers are used (layer = 0), the model relies solely on the original bidding mechanism. As the number of causal layers increases to 5, CausalAFL reaches 54.68% accuracy and 399.54 social welfare on CIFAR-10, marking an average gain of about 3.34% accuracy and 11.1 social welfare over the noncausal model. At this point, the ratio of NDE to NIE_r is well balanced, indicating the most effective amplification of indirect contributions. Adding 6 or 7 layers leads to declines in both accuracy and welfare, with a sharp rise in NDE and increased fluctuation in NIE_r, suggesting that overly deep causal networks introduce redundancy or begin to overfit noise. Overall, 5 causal layers strike the best balance between denoising and modeling complex causal paths and thus provide the most reliable contribution signals for the auction mechanism.

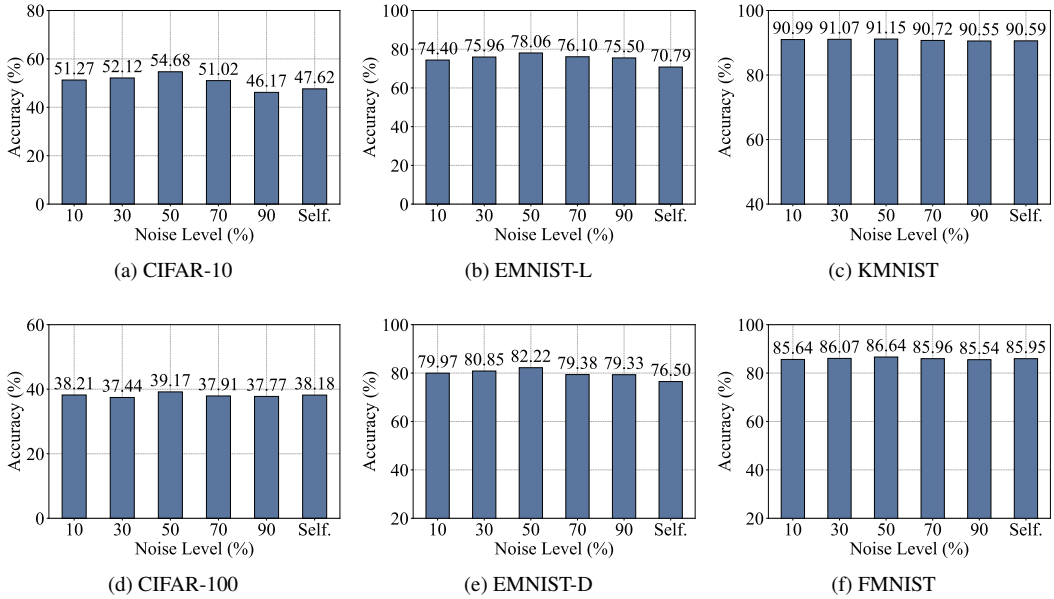
VAE Ablation. To evaluate the necessity of VAE, we conduct an ablation study by replacing the VAE with a traditional Linear Structural Equation Model (LSEM), as shown in Table 5. This variant models the relationship between latent and observed variables using linear regression, without capturing the distributional characteristics of latent confounders. The results demonstrate that the simpler bias correction method leads to a clear drop in both accuracy and social welfare. This suggests that linear causal adjustments perform poorly in complex, noisy environments with high non-IID settings, while VAE more effectively models latent confounding, enhancing the precision and robustness of causal effect estimation.

Table 5: Comparison results of VAE and LSEM on CIFAR-10 and CIFAR-100 with $\alpha = 0.01$.

	CIFAR-10		CIFAR-100	
	Acc.	Welf.	Acc.	Welf.
VAE	**54.68**	**399.54**	**39.17**	**278.25**
LSEM	17.52	163.98	19.59	186.14

B.2 Noise Study

Figure 5 shows how different noise levels affect model accuracy across datasets, where "self." represents the random value between the mean and variance of the bids made by the clients as noise, and the rest represents the random value between the mean and the noise level. For example, on CIFAR-10, accuracy steadily rises at 10% to 50% noise, then falls slightly at 70% to 90% noise, demonstrating adaptability to moderate noise. On EMNIST-L, accuracy peaks at about 78% when noise is 50% and stays above 75% at higher noise levels, dropping to around 70% under "self." noise, which highlights robust performance under distribution heterogeneity. The KMNIST curve is almost flat, fluctuating around 90% and exceeding 91% at its highest, and even under extreme noise it stays above 90%, showing exceptional noise resistance. On CIFAR-100, accuracy reaches about 39% at 50% noise and holds around 37% at 70 and 90% noise, while under "self." noise it sits near 38%, indicating stable performance even in complex tasks. These results jointly show that CausalAFL leverages causal correction to improve performance in moderate noise environments and maintains high accuracy even under extreme or adaptive noise, demonstrating its stability.



(a) CIFAR-10 (b) EMNIST-L (c) KMNIST

(d) CIFAR-100 (e) EMNIST-D (f) FMNIST

Figure 5: Noise results of CausalAFL on six datasets with $\alpha = 0.01$.

Table 6: Budget Results of CausalAFL on CIFAR-10 with $\alpha = 0.01$.

	Acc.	Welf.	NDE	NIE$_r$	ATE
10 Budgets	50.38	372.72	0.3551	0.5014	0.8565
15 Budgets	53.19	386.90	0.2801	0.8279	1.1079
20 Budgets	54.68	399.54	0.2015	1.1204	1.3219
25 Budgets	53.89	389.34	0.3522	0.9142	1.2664
30 Budgets	52.75	381.54	0.5596	0.5162	1.0758

B.3 Budget Study

We explore how increasing the server's per-round budget from 10 to 30 affects performance, as shown in Table 6. The model's performance rises and then falls as the budget increases. When the budget is 10, test accuracy on CIFAR-10 is 50.38% with social welfare of 372.72. Raising the budget to 15 yields slight gains in both accuracy and welfare, and when the budget reaches 20, CausalAFL achieves its best results on both datasets, with CIFAR-10 accuracy climbing to 54.68%, social welfare to 399.54. At this level, the causal model still effectively selecting high-quality clients, producing the high NDE and NIE$_r$. As the budget increases further to 25 and 30, each winning seller receives substantially larger payments, which amplifies NDE in the causal model and diminishes the relative importance of the indirect pathway, causing the total effect to move away from its optimum and leading to declines in both accuracy and social welfare. This shows that an excessive budget actually weakens the auction's incentive power and invites more noisy participants. Overall, a budget of 20 best balances the selection of high-quality clients with efficient resource allocation, delivering the optimal combination of model performance and social welfare.

B.4 Cost and Payment

We compare each winning seller's cost with the payment they receive in the final auction round, as shown in Figure 6. It shows that every winner, whether the lowest-cost or highest-cost participant, receives a payment that exceeds their cost, ensuring individual rationality and providing a positive incentive to each seller. Those sellers with medium-level costs often contribute more to global model improvement and thus receive the larger rewards, demonstrating that the server can precisely allocate its limited budget according to each seller's marginal contribution. In contrast, sellers with lower costs and smaller contributions still receive enough payment to cover their participation expenses,

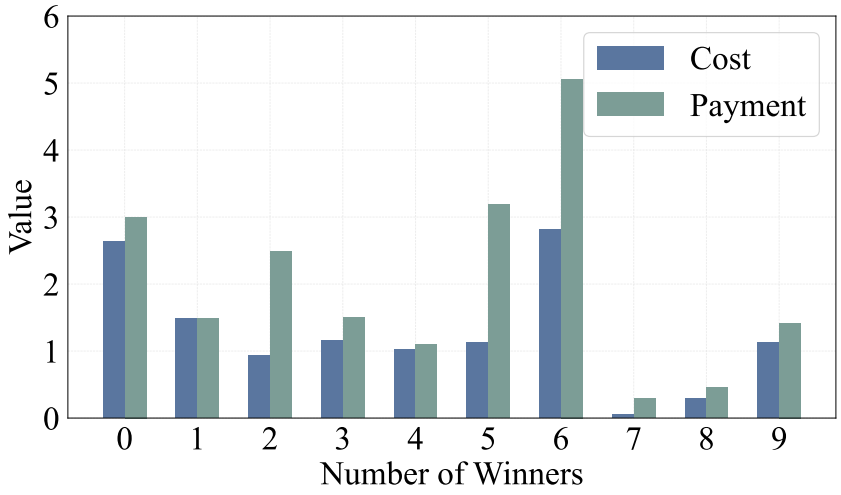


Figure 6: Cost and payment results of winning sellers on EMNIST-L with $\alpha = 0.01$.

preventing high-quality sellers from withdrawing due to insufficient compensation. This payment scheme balances efficiency with fairness and effectively deters free-riding. CausalAFL aligns every unit of budget with genuine model contributions by incorporating both NDE and NIE$_r$ into a causally driven social welfare objective, supporting contribution-aware payments in the evaluated setting. The individual-rationality guarantee is conditional on the positive-contribution and budget premises of Theorem 2.

B.5 Asynchronous Study

Table 7: Impact of different asynchronous factors on CIFAR-10 with $\alpha = 0.01$.

	Acc.	Welf.
Dropout	47.91	387.89
Communication Latency	49.36	393.61
Stragglers	45.90	384.92
All	44.39	377.48

We explore an asynchronous auction setting, covering dropout, communication latency, or stragglers, as shown in Table 7. Specifically, in each auction round, only responding sellers can participate, with response determined by random response. Each round includes a random number of client dropouts (random number can be 0), discontinuing auction and FL collaboration. Clients below 90% response speed are considered laggards and excluded from the auction. For stragglers, we set their operating frequency below a threshold, rendering them uncompetitive. The top 50% of participating clients win each auction.

Although these conditions increase auction difficulty and reduce client performance, the decline remains acceptable while maintaining stable social welfare. This occurs because the causal guidance term ATE $\cdot\, t$ dynamically adjusts bidding strategies in asynchronous environments. And high-latency clients exhibit reduced ATE, decreasing $\text{bid}^s_{i,t}$ to lower winning probability and avoid training interruption risks. Meanwhile, the social welfare objective quantifies the causal transmission efficiency from bids M to marginal contributions Y through NIE$_r$. This prioritizes clients with high data distribution alignment, maintaining model performance and social welfare balance in asynchronous auctions.

B.6 Client Evolution Study

To explore the effectiveness of ATE in CausalAFL, we supplement scenarios where client data distributions and computational conditions continuously evolve during auction rounds, as shown

Table 8: Comparison of static and evolving scenarios on CIFAR-10 and CIFAR-100 with $\alpha = 0.01$.

	CIFAR-10		CIFAR-100	
	Acc.	Welf.	Acc.	Welf.
Static	54.68	399.54	39.17	278.25
Evolving	49.42	347.65	42.43	296.62

in Table 8. It can be observed that although accuracy and social welfare decrease on CIFAR-10, model performance and social welfare are ultimately enhanced on the more complex CIFAR-100. This occurs because the ATE guidance mechanism essentially functions as a dynamic accumulator of causal effects, relying not on static historical contributions but on the client's ability to maintain causal chain stability. In complex tasks, CausalAFL more effectively identifies resource-stable clients, counteracting negative impacts from environmental fluctuations and ultimately enhancing performance.

Table 9: Client utility over rounds under rare-data setting with $\alpha = 0.01$.

	40	80	120	160	200
CIFAR-10	37.36	62.72	85.72	93.58	109.90
CIFAR-100	28.27	47.24	71.21	84.52	122.60

B.7 Lemon Market Study

The root of the lemon market problem lies in high-quality clients being forced to raise bids due to high costs, while low-quality clients strategically underbid to masquerade as cost-effective, resulting in adverse selection. we supplement a new case experiment: Client 0 possesses rare data but exhibits distributional deviation from the global model (i.e., it owns 2 classes in CIFAR-10 and 10 classes in CIFAR-100 absent from other clients), yet its distribution remains extremely heterogeneous. We track this client's auction wins, finding it is selected 96 rounds in 200 auctions on CIFAR-10 and 125 rounds in 200 auctions on CIFAR-100. Its individual utilities are shown in Table 9. Results demonstrate that CausalAFL maintains this client's competitiveness with no adverse selection, as validated by its sustained winning rate and utility.

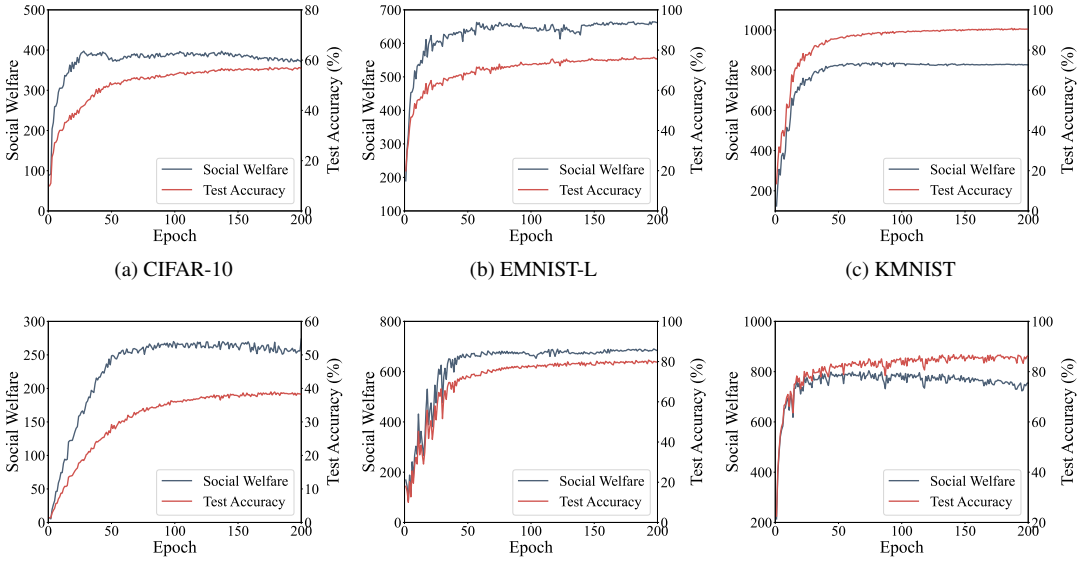
Table 10: Comparison with standard regression techniques on CIFAR-10 with $\alpha = 0.01$, where CausalAFL performs 54.68% accuracy, 399.50 social welfare and 1.5793 value of ATE.

	Lasso		OLS		Ridge	
	5L-NN	VAE	5L-NN	VAE	5L-NN	VAE
Acc.	32.40	25.51	31.64	24.45	33.00	18.94
Welf.	312.69	242.90	301.66	233.32	315.83	177.91
ATE	0.0129	0.3061	0.0025	0.1768	0.0070	0.2499

B.8 ATE estimation Study

In auction environments, bidirectional causal pathways exist among client cost T, bid M, and contribution Y. For instance, data size X_{TM} simultaneously influences cost T and bid M. Data distribution divergence X_{MY} simultaneously affects bid M and contribution Y. Directly regressing $Y \sim T$ would yield biased ATE estimates due to confounding variables. Moreover, bid M acts as a strategically manipulated mediator. Clients artificially perturb bids ($\text{bid}_{i,t}^s = \text{Cost}_{i,t}^s + \varepsilon_{i,t}$), making M a strategic decision variable. Directly modeling $Y \sim T$ fails to distinguish NDE and NIE_r, where the former is the direct effect of cost changes on contribution and the later is the indirect effect of cost mediated by bidding strategies.

We evaluate the impact of standard regression techniques (Lasso, OLS, Ridge) for direct ATE estimation, including regression with M as a covariate using 5 layer neural networks (5L-NN), regression with Z_{TM}, Z_{MY} as covariates using VAE, as shown in Table 10. Results show that regression-based ATE estimates are near-zero or negligible, offering minimal performance gains. In contrast, CausalAFL yields significantly larger ATE estimates, leading to higher accuracy and



(a) CIFAR-10 (b) EMNIST-L (c) KMNIST

(d) CIFAR-100 (e) EMNIST-D (f) FMNIST

Figure 7: Convergence performance with social welfare and test accuracy on six datasets.

social welfare. This validates that only by fully eliminating confounding and correctly isolating direct/indirect effects can ATE effectively guide the auction mechanism.

B.9 Convergence Performance

We evaluate the convergence of CausalAFL on six datasets under $\alpha = 0.01$ by tracking both the global model's test accuracy and the value of social welfare, as shown in Figure 7. As communication rounds progress, the accuracy and social welfare both rise smoothly and rapidly before gradually leveling off after approximately fifty to one hundred rounds. On EMNIST-L, KMNIST and FMNIST the model's accuracy swiftly climbs from an initial 10% to 20% range to exceed 70% within the first fifty rounds. At the same time, social welfare increases in tandem, demonstrating that the causally driven bidding mechanism consistently selects high-contribution clients early on to deliver rapid gains to the global model. Even on the more challenging CIFAR-10, CIFAR-100 and EMNIST-D, CausalAFL achieves around 40% to 80% accuracy improvement within the first one hundred rounds and its social welfare curve mirrors the accuracy trend. Overall, the convergence curves confirm that the joint optimization of a generative causal graph and variational inference delivers both efficiency and robustness in AFL.

C Limitations

Although CausalAFL demonstrates promising results in mitigating causal bias in AFL, it still faces two notable limitations:

First, this work considers only two proxy confounders to approximate the latent confounders. However, in real-world auction settings, many other unobserved confounders may exist, such as device stability, data annotation quality, network latency, or strategic behaviors. These factors may simultaneously affect multiple variables in the causal graph, and failing to account for them could result in residual bias in the estimation of causal effects, potentially undermining the accuracy and fairness of bid adjustments.

Second, the method assumes a standard mediation structure among client information, bid, and marginal contribution, where all causal influence is transmitted through the bid. Nevertheless, in practice, marginal contribution may also be directly affected by client information. For example, high costs may implicitly indicate stronger computational resources or higher data quality, which can

directly enhance global model performance. Restricting the causal modeling to a single pathway may fail to capture such direct effects, thus limiting the expressiveness and generalizability of CausalAFL.

D Code of Ethics

This paper adheres to the principles of integrity, transparency and respect for privacy. The proposed CausalAFL framework safeguards client confidentiality by estimating causal mediation effects through proxy variables rather than by accessing sensitive distribution details, thereby greatly reducing the risk of exposing raw data. In our experiments, we simulate real bidding scenarios with synthetic noise to avoid any direct use or exposure of actual user data and to eliminate the potential for misuse or privacy breaches.

E Societal Impacts

CausalAFL brings CMA into AFL to boost the participation of high-quality clients and improve overall model performance, offering a new incentive mechanism for large-scale collaborative training in resource constrained environments. It shows strong potential in scenarios such as smart cities, healthcare and industrial IoT where data privacy must be maintained while multiple parties collaborate, striking a balance between participant benefits and system efficiency. At the same time the framework could be misused to craft sophisticated bidding strategies that disadvantage vulnerable participants or distort market fairness. To address these concerns we recommend performing comprehensive risk assessments before deployment and adopting additional privacy preserving techniques to reinforce data protection, so that ethical and societal risks are effectively mitigated as technological progress advances.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly stated that this paper is aimed at reduce information asymmetry and corrects for confounding factors in federated learning.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to Appendix.C.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Please refer to Appendix.A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please refer to Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We use open-source datasets and publicly disclose the experimental setup in Section 4.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Please refer to Section 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: These results were obtained under fixed hyper-parameters and a fixed train-test split, as reproducibility was prioritized for datasets.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to Section 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Please refer to Appendix.D.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to Appendix.E.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The datasets used in this paper are open source.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not important, original, or non-standard components of the core methods in this paper.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.