



Trinity College Dublin

Coláiste na Tríonóide, Baile Átha Cliath

The University of Dublin

A Coopetitive-Compatible Data Generation Framework for Cross-silo Federated Learning

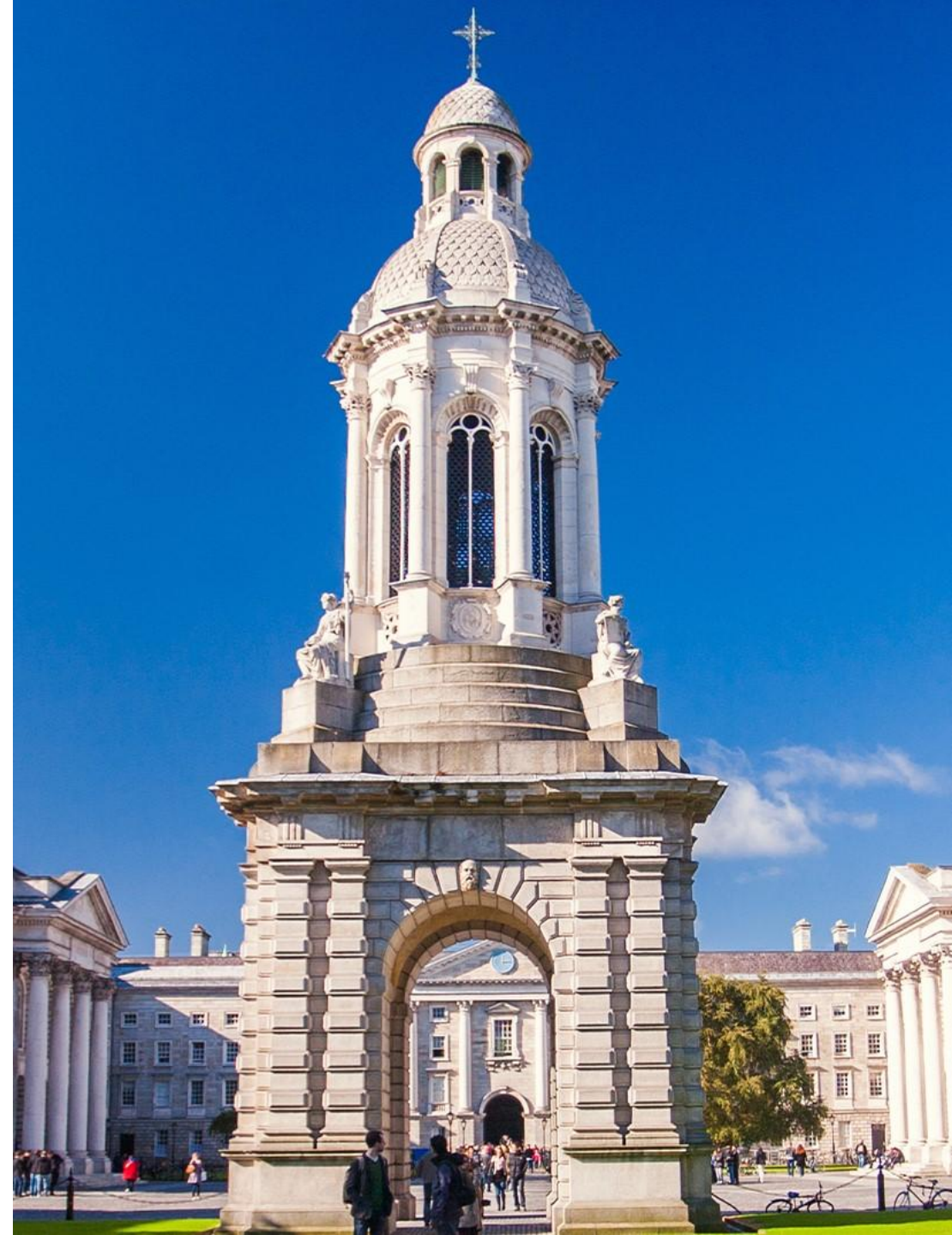
IEEE GLOBECOM 2025

Thanh Linh Nguyen, Quoc-Viet Pham

School of Computer Science and Statistics



Taipei, December 2025



Background: Cross-silo Federated Learning (CFL)

❖ Definition

- A privacy-preserving data sharing/learning paradigm [1]

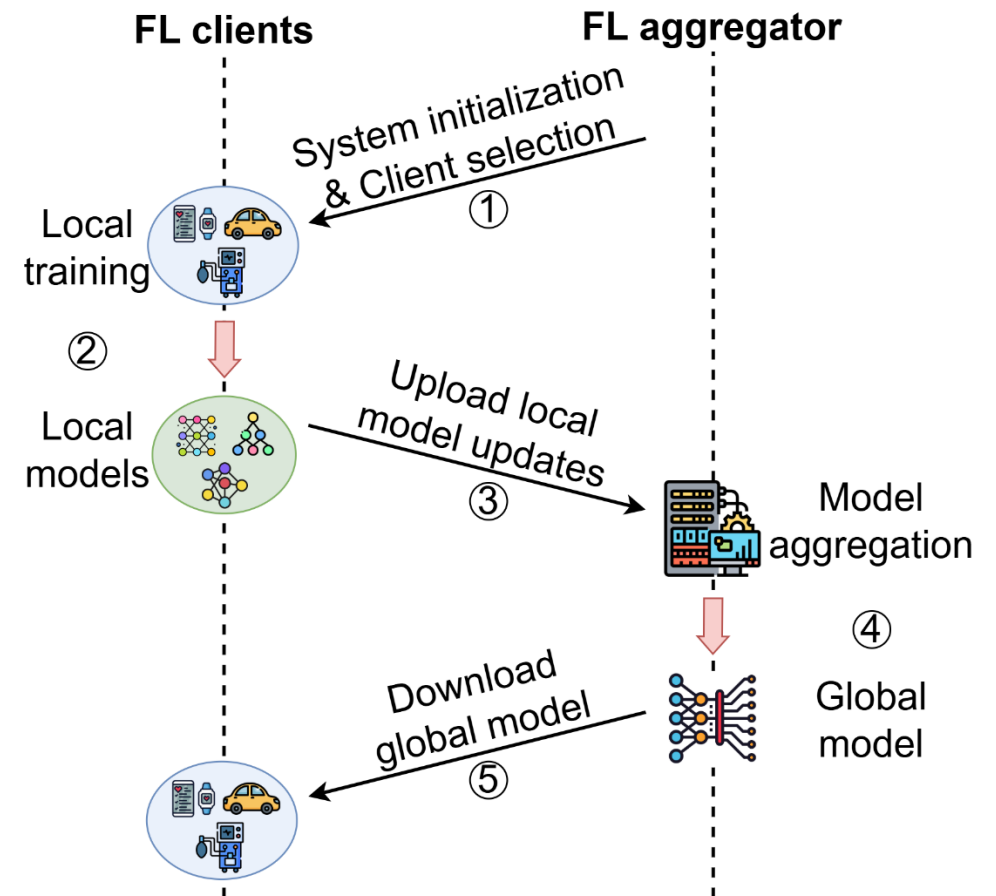
❖ Components

- Distributed FL organizations
- A central server

❖ Workflow

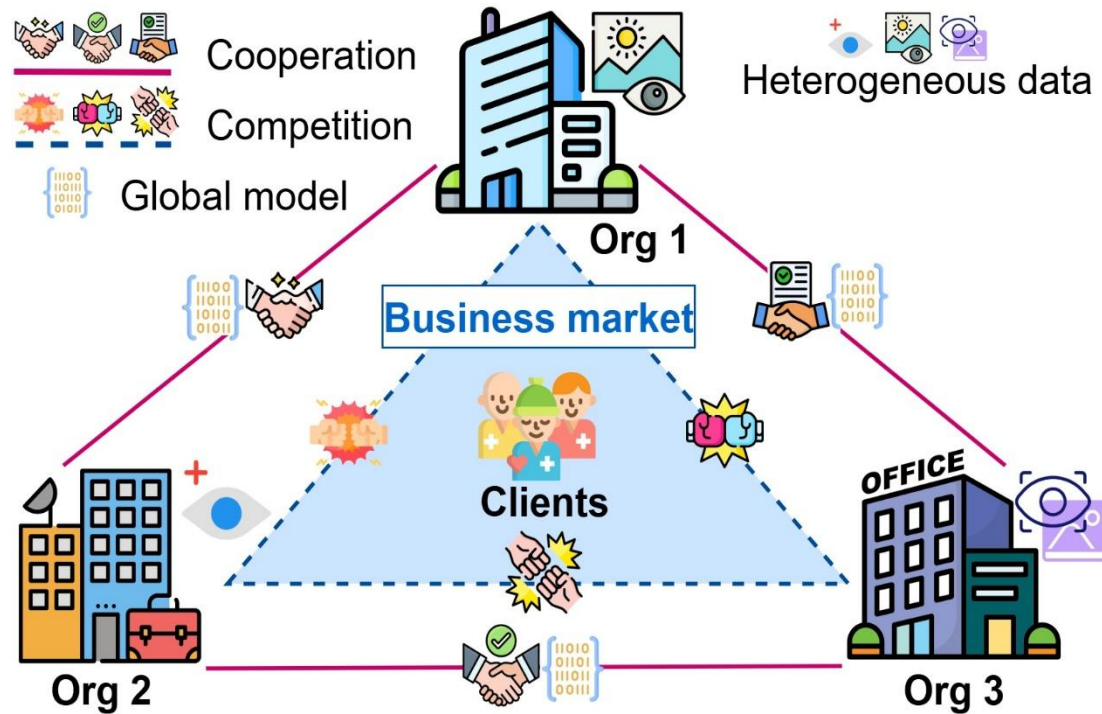
- Instead of sharing raw data, FL organizations only share the trained model parameters

Data Stays, Intelligence Unites



[1]. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Ar- cas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.

Background: Coopetition = Cooperation + Competition



Definition 1 (CFL Coopetition) [2]. CFL coopetition characterizes the dual strategy of **cooperative engagement** in joint model training to develop a shared global model, **competitive utilization of this model** to maximize their revenue outcomes or market performance, reflecting their inherent market rivalry.

[2]. D. R. Gnyawali and B.-J. Park, "Co-opetition and technological innovation in small and medium-sized enterprises: A multilevel conceptual model," *Journal of Small Business Management*, vol. 47, no. 3, pp. 308–330, 2009.

Motivation

▪ Research Gaps and Challenges (RGC):

- **#RGC1:** *Statistical heterogeneity* and its influence have been underexamined under the **competitive** nature among organizations.
- **#RGC2:** Lacks an explicit theoretical **characterization of statistical heterogeneity and competitive intensity**, nor quantify their effects on learning performance, social welfare, and organizational decision-making.
- **#RGC3:** Given that data generation is resource-intensive, organizations need to **optimize the use of GenAI-based augmented data to sustain CFL cooperation** (e.g., maximizing the global FL model performance and social welfare).

Research Question #1 (RQ1): *How does the coexistence of statistical heterogeneity and competition affect organizations' decision-making and social welfare?*

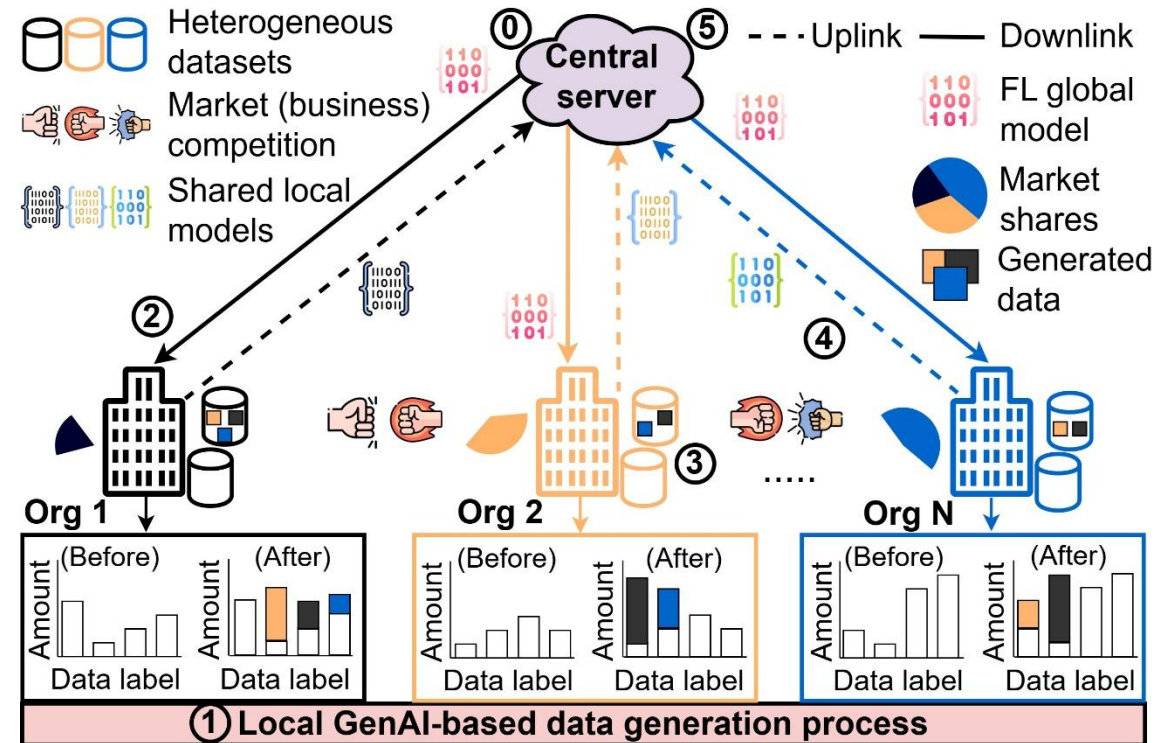
Research Question #2 (RQ2): *In the presence of statistical heterogeneity and competition, how can we design an effective GenAI-based data generation mechanism for organizations?*

Proposed Solution – System Architecture

We propose a **Co**opetitive-**Co**mpatible Data **Ge**neration (**CoCoGen**) Framework for CFL.

System Architecture

- Competitive organizations
 - Contribute the local and synthetic data
- Central server
 - Data contribution strategy profile allocation
 - Model aggregation
 - Payment



CoCoGen system with learning workflow comprising of (0) *Global model initialization & Organization strategy profile allocation*, (1) *Local GenAI-based data generation*, (2) *Global model downloading*, (3) *local training*, (4) *local upload updates*, (5) *receive & aggregate model updates*.

Proposed Solution – Problem Formulation [1/4]

- Set of clients: $N := \{1, \dots, n, \dots, N\}$
- Set of training rounds: $T := \{1, \dots, t, \dots, T\}$
- Original (local) & GenAI-augmented datasets of org n : $D_n^{loc} \& D_n^{gen}, \forall n \in N$, respectively
- Original data size: $d_n^{loc} := |D_n^{loc}|$
- GenAI-augmented data size: $d_n^{gen} := |D_n^{gen}|$
- Total data size: $d_n^{mix} := d_n^{loc} + d_n^{gen}$
- GenAI-augmented data size strategy profile: $\mathbf{d}^{gen} = \{d_n^{gen} : \forall n \in N\}$
- GenAI-augmented data size strategy profile excluding n : $\mathbf{d}_{-n}^{gen} = \{d_{n'}^{gen} : \forall n', n \in N, n' \neq n\}$
- **Power-law function based n 's local model performance & global model performance** given $\mathbf{d}^{gen}$, respectively:

$$\epsilon_n(d_n^{gen}) \approx \alpha (d_n^{loc} + d_n^{gen})^{-\beta} - \delta, \quad \forall \alpha, \beta > 0, \delta \geq 0, \quad (1)$$

$$\epsilon(\mathbf{d}^{gen}) = \exp\left(\frac{\frac{1}{N} \sum_{n \in N} \epsilon_n}{\varrho}\right). \quad (2)$$

Proposed Solution – Problem Formulation [2/4]

▪ **Local cost of client n in one training round t :**
$$C_n = C_n^{cmp} \underbrace{\kappa_n(\eta_n d_n^{mix})}_{\text{For training}} + \underbrace{\mu_n d_n^{gen}}_{\text{For generating}} f_n^2, \quad (3)$$
 where f_n is the processing capacity of n .

▪ **The revenue of n from contributing d_n^{gen} :**
$$r_n(d_n^{gen}, \mathbf{d}_{-n}^{gen}) = \psi_n \underbrace{[\epsilon_0]}_{\substack{\text{Global model} \\ \text{performance} \\ \text{before training}}} - \underbrace{\epsilon(\mathbf{d}^{gen})}_{\substack{\text{Global model} \\ \text{performance} \\ \text{trained on } \mathbf{d}^{gen}}}, \quad (4)$$

where ψ_n is the organization n 's revenue per the global model's performance unit.

▪ **Total payoff redistribution for n :**
$$P_n(d_n^{gen}, \mathbf{d}_{-n}^{gen}) = \sum_{n' \in N} \xi \gamma_{n,n'} \left[\underbrace{\epsilon_n(d_n^{gen}, \mathbf{d}_{-n}^{gen})}_{\substack{\text{\textit{n's local model} \\ \text{performance}}}} - \underbrace{\epsilon_{n'}(d_{n'}^{gen}, \mathbf{d}_{-n'}^{gen})}_{\substack{\text{\textit{n's local model} \\ \text{performance}}} \right], \quad (5)$$

where $\xi \geq 0$ is the payoff compensation for each contribution gap unit,
 $\gamma_{n,n' \in (0,1)}$ is the competitive intensity between n and n' .

Proposed Solution – Problem Formulation [3/4]

- Loss incurred by coopetition for n :

$$R_n(d_n^{gen}, \mathbf{d}_{-n}^{gen}) = \sum_{n' \in N} \phi_{n'} \gamma_{n,n'} [\epsilon_n(d_n^{gen}, \mathbf{d}_{-n}^{gen}) - \epsilon_{n'}(d_{n'}^{gen}, \mathbf{d}_{-n'}^{gen})], \quad (6)$$

where $\phi_{n'} \geq 0$ is the organization n 's revenue per unit of contribution gap.

- Client n 's utility* per training round:

$$U_n(d_n^{gen}, \mathbf{d}_{-n}^{gen}) = \underbrace{r_n(d_n^{gen}, \mathbf{d}_{-n}^{gen})}_{\text{Revenue from contributing } d_n^{gen}} + \underbrace{P_n(d_n^{gen}, \mathbf{d}_{-n}^{gen})}_{\text{Payoff redistribution}} - \underbrace{C_n - C_0}_{\text{Local costs and fees for the server}} - \underbrace{R_n(d_n^{gen}, \mathbf{d}_{-n}^{gen})}_{\text{Coopetition loss}}, \quad (7)$$

where C_0 is the fee of calculating the data generation strategy profile and aggregating the global model, which is paid by all orgs.

- Social welfare**: $\sum_{n \in N} U_n(d_n^{gen}, \mathbf{d}_{-n}^{gen})$

*Utility is the difference between gain and loss.

**Social welfare is a sum of the utility of organizations.

Proposed Solution – Problem Formulation [4/4]

- **Game G :**

- **Players:** N cooperative & heterogeneous organizations per training round.
- **Strategies:** organization n contributes d_n^{gen} in the training round t .
- **Objectives:** Each organization aims to **maximize** its utility expressed in (7).

- **Global Objective:** Social welfare **maximization**

$$\max_{d^{gen}} \sum_{n \in N} U_n(d_n^{gen}, d_{-n}^{gen}), \quad (8a)$$

Constraints:

1. $d_{\min}^{gen} \leq d_n^{gen} \leq d_{\max}^{gen}, \quad (8b)$

2. Individual Rationality (IR): $U_n(d_n^{gen}, d_{-n}^{gen}) \geq 0, \quad (8c)$

3. Budget Balance (BB): $P_n(d_n^{gen}, d_{-n}^{gen}) = 0. \quad (8d)$

Proposed Solution – Potential game-based mechanism [1/2]

Theorem 1: $\mathbf{G}$ is a weighted potential game with the potential function $F(\mathbf{d}^{gen})$, which is calculated by:

$$F(\mathbf{d}^{gen}) = \epsilon(\mathbf{d}^{gen}) - \sum_{n \in N} \frac{\kappa_n C_n^{cmp} (\eta_n + \mu_n) d_n^{gen} f_n^2}{z_n}, \quad (9)$$

where $z_n = \sum_{n' \in N} \gamma_{n,n'} (\xi - \phi_{n'}) - \psi_n < 0$.

→ The Nash Equilibrium of $\mathbf{G}$ will be the minimizer of $F(\mathbf{d}^{gen})$.

Proposed Solution – Potential game-based mechanism [2/2]

Theorem 2: $\mathbf{G}$ possesses a NE of $\mathbf{d}^{*,gen} = \{d_n^{*,gen} : \forall n \in N\}$ under different cases, where we define the following variables:

$$A_1 \triangleq \frac{1}{N} \sum_{n \in N} \alpha (d_n^{loc} + d_n^{*,gen})^{-\beta} - \delta; \quad A_2 \triangleq \frac{\kappa_n C_n^{cmp} (\eta_n + \mu_n) f_n}{z_n}; \quad A_3 \triangleq (d_n^{loc} + d_n^{*,gen})^{-\beta-1}.$$

□ **Case 1:** $d_n^{*,gen} = d_{min}^{gen}$ iff $\frac{A_3 \alpha \beta}{N \varrho} \exp\left(\frac{A_1 - 1}{\varrho}\right) < -A_2$.

□ **Case 2:** $d_n^{*,gen} = d_{max}^{gen}$ iff $\frac{A_3 \alpha \beta}{N \varrho} \exp\left(\frac{A_1 - 1}{\varrho}\right) > -A_2$.

□ **Case 3:** $d_{min}^{gen} < d_n^{*,gen} < d_{max}^{gen}$ and we have

$$d_n^{*,gen} = \left[-\frac{A_2 N \varrho}{\alpha \beta} \exp\left(\frac{A_1 - 1}{\varrho}\right) \right]^{-\frac{1}{\beta+1}} - d_n^{loc},$$

with the following update rule for $\{u_n^{(k)}\}_{n=1}^N$ using the fixed point iteration:

$$\forall n \in \{1, \dots, N\}, u_n \triangleq d_n^{loc} + d_n^{gen}, u_n^{(k+1)} = \left[-\frac{A_2 N \varrho}{\alpha \beta} \exp\left(-\frac{\frac{1}{N} \sum_{n \in N} \left[\alpha (u_n^{(k)})^{-\beta} - \delta \right] - 1}{\varrho} \right) \right]^{-\frac{1}{\beta+1}},$$

where $u_n^{(k)}$ converges to $(d_n^{loc} + d_n^{*,gen})$ as $k \rightarrow \infty$. We can then get the value of $d_n^{*,gen}$.

Numerical Results – Simulation Setup

Settings:

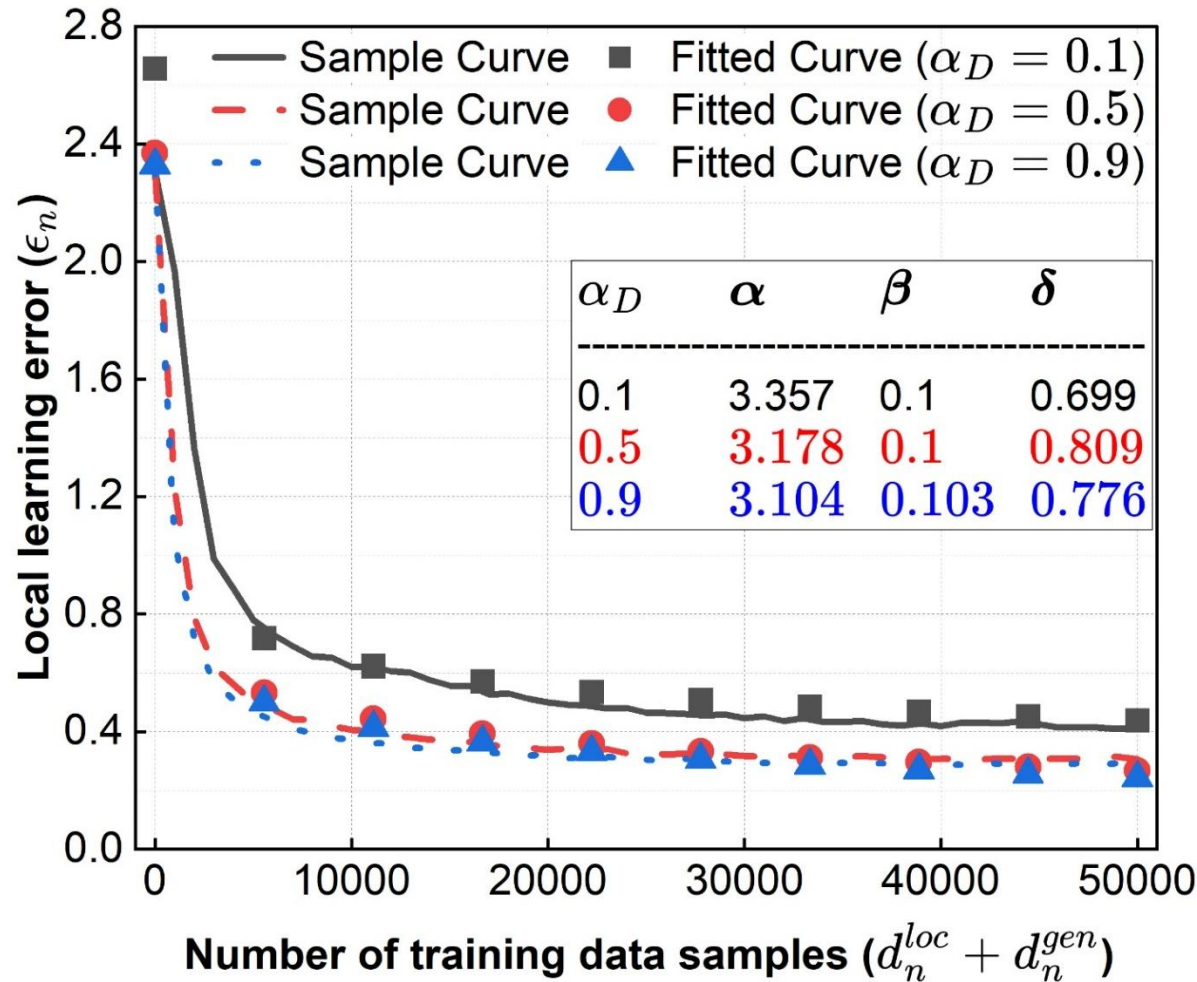
- The number of clients: $N = 10$
- The number of local data samples d_n^{loc} : $U(1000, 3000)$ samples
- The minimum number of GenAI-based augmented data d_{min}^{gen} : 0 samples
- The maximum number of GenAI-based augmented data d_{max}^{gen} : 3000 samples
- Electronic price: 0.3438 euro per kWh [3]

Benchmarks for comparisons:

- **Vanilla CFL (VCFL).** Traditional CFL without GenAI-based data augmentation approaches (i.e., $d_n^{gen} = 0$).
- **Without Competition among Organizations (WCO).** There is no competition among organizations (i.e., $\gamma_{n,n'} = 0, \forall n, n' \in N, n \neq n'$).
- **Random Data Generation (RaDG).** Organizations randomly generate the amount of GenAI-based augmented data for local training.

[3]. <https://selectra.ie/energy/guides/electricity-prices-ireland>

Numerical Results – Feasibility [1/2]

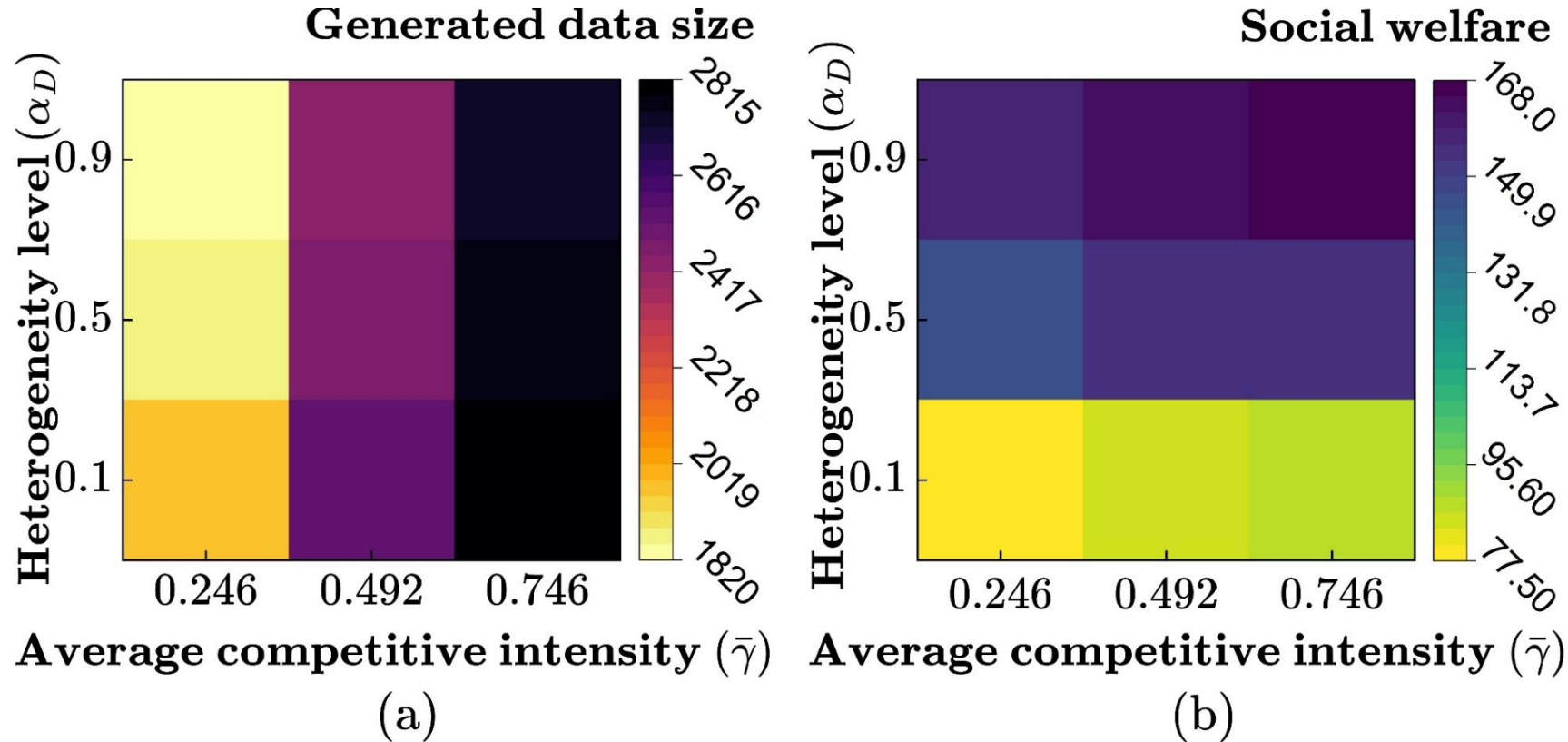


- **Obs. 1:** With FMNIST, as the amount of GenAI-generated data increases, the local learning error decreases significantly across different α_D values.
- **Obs. 2:** Understanding of scaling relationships between learning error, training set size ($d^{loc} + d^{gen}$), and α_D in the distributed settings.

Emphasize the importance to augment training data using GenAI in heterogeneous CFL settings.

**Our mechanism and weighted-potential-game analysis are dataset-agnostic.*

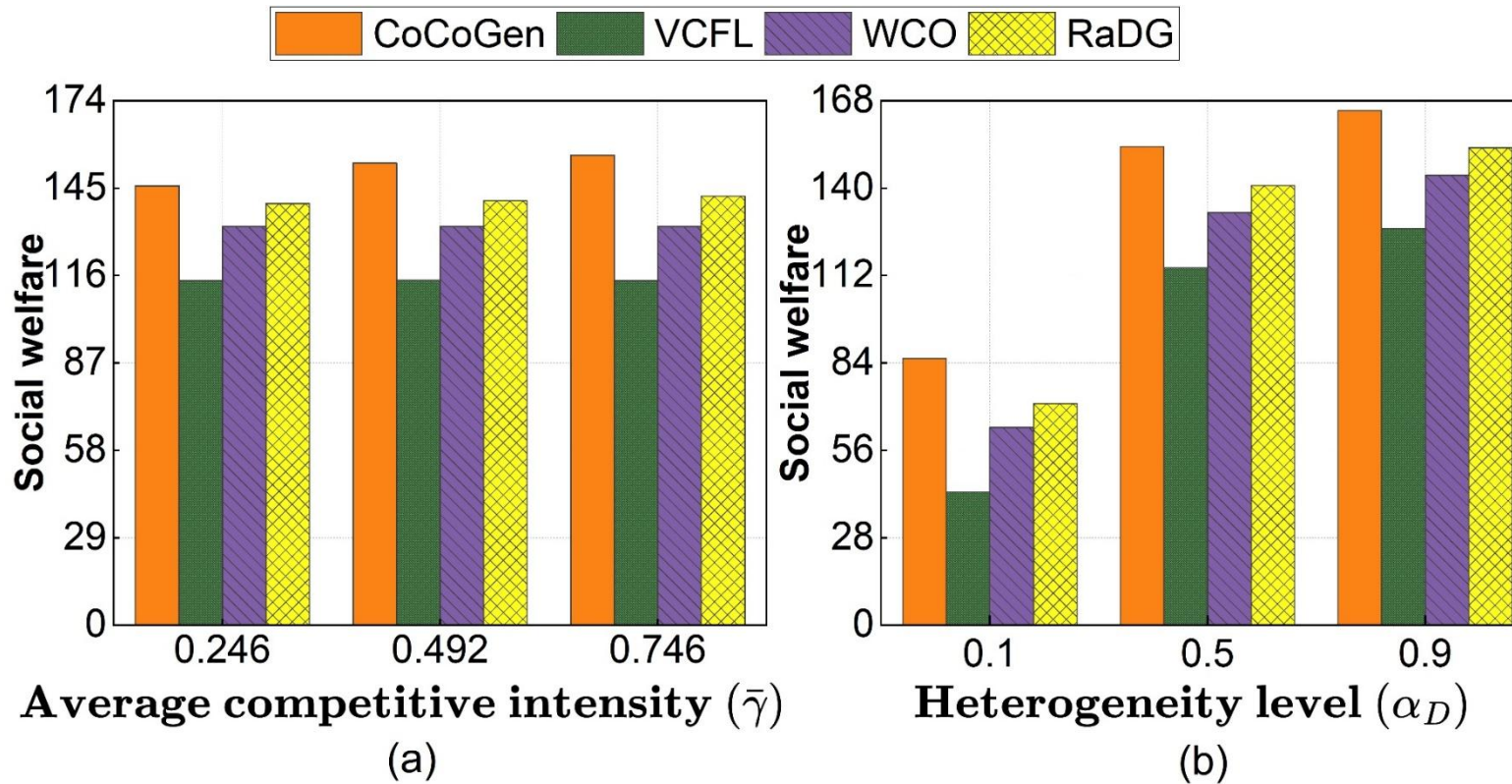
Numerical Results – Feasibility [2/2]



- **Obs. 3:** Under **high** α_D ($\alpha_D = 0.1$) with **increasing** $\bar{\gamma}$, **more** generated data required but **boost** social welfare.
- **Obs. 4:** Under **high** $\bar{\gamma}$ ($\bar{\gamma} = 0.746$) with **increasing** α_D , **more** generated data required but **reduce** social welfare.

Greater competition $\bar{\gamma}$ and heterogeneous level α_D require more generated data but reduce social welfare (#RQ1).

Numerical Results – Efficiency



- **Obs. 1:** Under a fixed α_D , **CoCoGen** and RaDG increase over $\bar{\gamma}$, VCFL yields the lowest, WCO maintains stable.
- **Obs. 2:** Under a fixed $\bar{\gamma}$, as data distributions become more heterogeneous (i.e., smaller α_D), the social welfare gap between **CoCoGen** and the baselines widens.

CoCoGen is more robust to $\bar{\gamma}$ and α_D , outperforming other baselines in terms of social welfare (#RQ2).

Conclusion and Future Work

- We proposed **CoCoGen**, a cooperative-compatible data generation framework, that characterizes **competition** and **statistical heterogeneity** through learning loss and utility-based functions.
- We have proved that each training round constitutes a weighted potential game, enabling the derivation of GenAI-based data generation strategies that maximize social welfare.
- Higher competitive intensity (i.e., $\bar{\gamma}$) and statistical heterogeneity (i.e., α_D) significantly raise GenAI-based data generation demands while reducing social utility and vice versa.
- Compared with other benchmarks, **CoCoGen** achieves **better social welfare**.



Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Thank You

Thanh Linh Nguyen
Email: tnguyen3@tcd.ie

APPENDIX

A. Proof of Theorem 1

Proof. We adopt the forward method [21], where we consider the case where an organization n can deviate from strategy $\{d_n^{\text{gen}}\}$ to $\{d_n'^{\text{gen}}\}$, while other organizations' strategy set $\{d_{-n}^{\text{gen}}\}$ keeps unchanged. We have the following equality (20), shown at the bottom of the page.

According to formulas (14) and (20), we obtain

$$F(d^{\text{gen}}) - F(d'^{\text{gen}}) = \frac{1}{z_n} [U_n(d_n^{\text{gen}}, d_{-n}^{\text{gen}}) - U_n(d_n'^{\text{gen}}, d_{-n}^{\text{gen}})] \quad (21)$$

The Eq. (20) shows that any change of the $F(d^{\text{gen}})$ is equal to the $\frac{1}{z_n}$ -scaling change of the organization utility function caused by any unilateral deviation of the organization, which indicates that the $\mathcal{G}$ is a weighted potential game. This completes the proof. $\square$