FLHetBench: Benchmarking Device and State Heterogeneity in Federated Learning

Junyuan Zhang[2*], Shuang Zeng[1*], Miao Zhang[3], Runxi Wang[2], Feifei Wang[1], Yuyin Zhou[5], Paul Pu Liang[4], Liangqiong Qu[1†]

[1]The University of Hong Kong, [2]Beihang University, [3]New York Univerisity
[4]Carnegie Mellon University, [5]UC Santa Cruz

Project Page: https://carkham.github.io/FL_Het_Bench/
Github: https://github.com/Carkham/FLHetBench

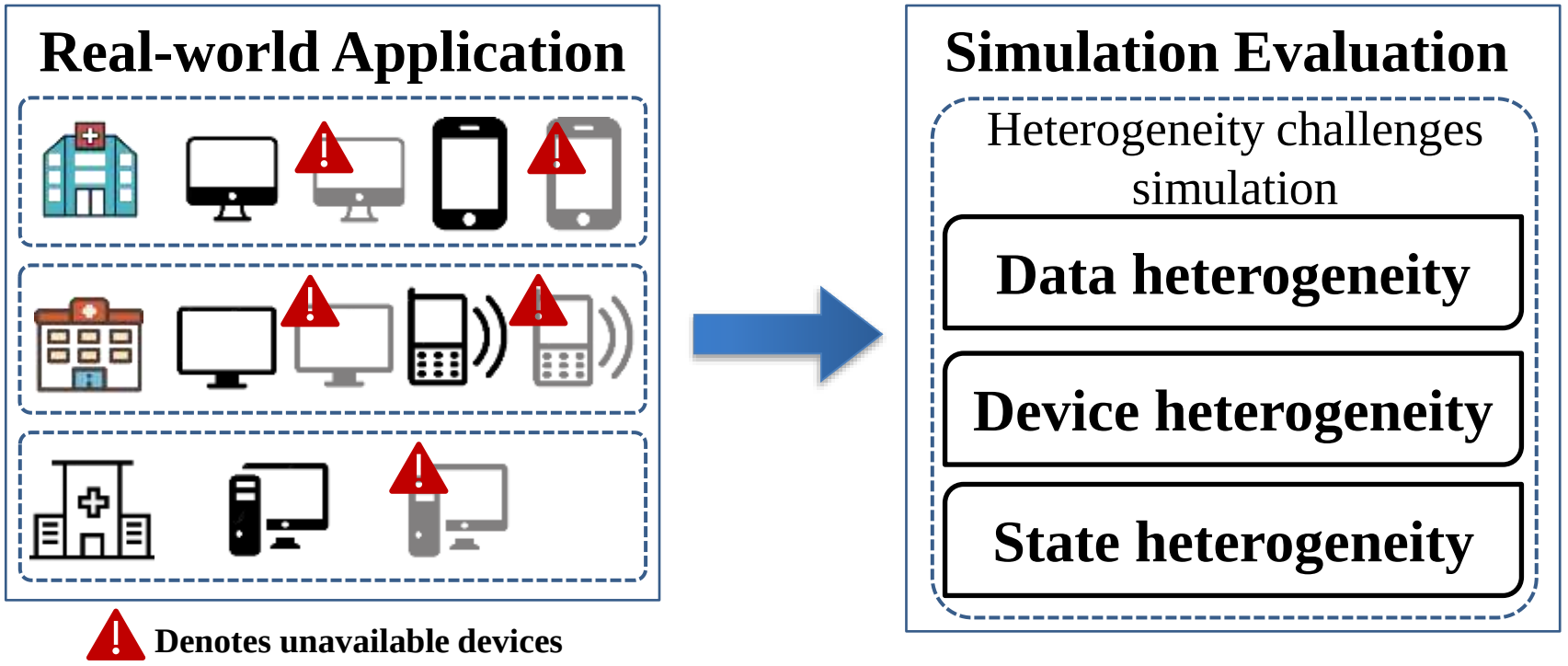
Background

Federated learning enables collaborative training without sharing private data among clients.

	Centralized training	Federated learning
Data Transfer	Raw data	Model
	Bring the data to the model	Bring the model to the data

Background

Federated learning faces heterogeneity challenges



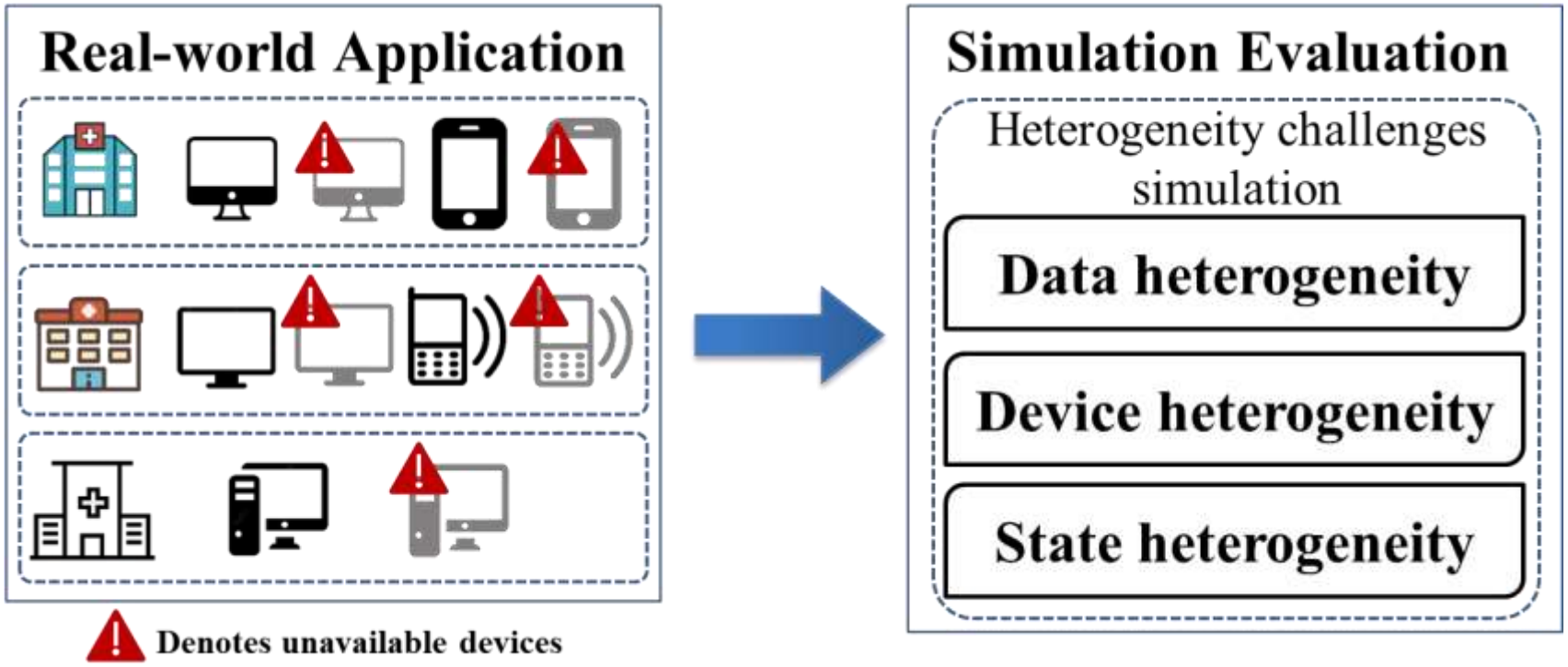
Real-world application challenges:

- Inconsistent local data distribution （data heterogeneity)
- Different device capability (device heterogeneity)
- Asynchronous online states (state heterogeneity)

Background

What happens to different FL algorithms when they are employed in real-world FL environments with varying degrees of device and state heterogeneity?

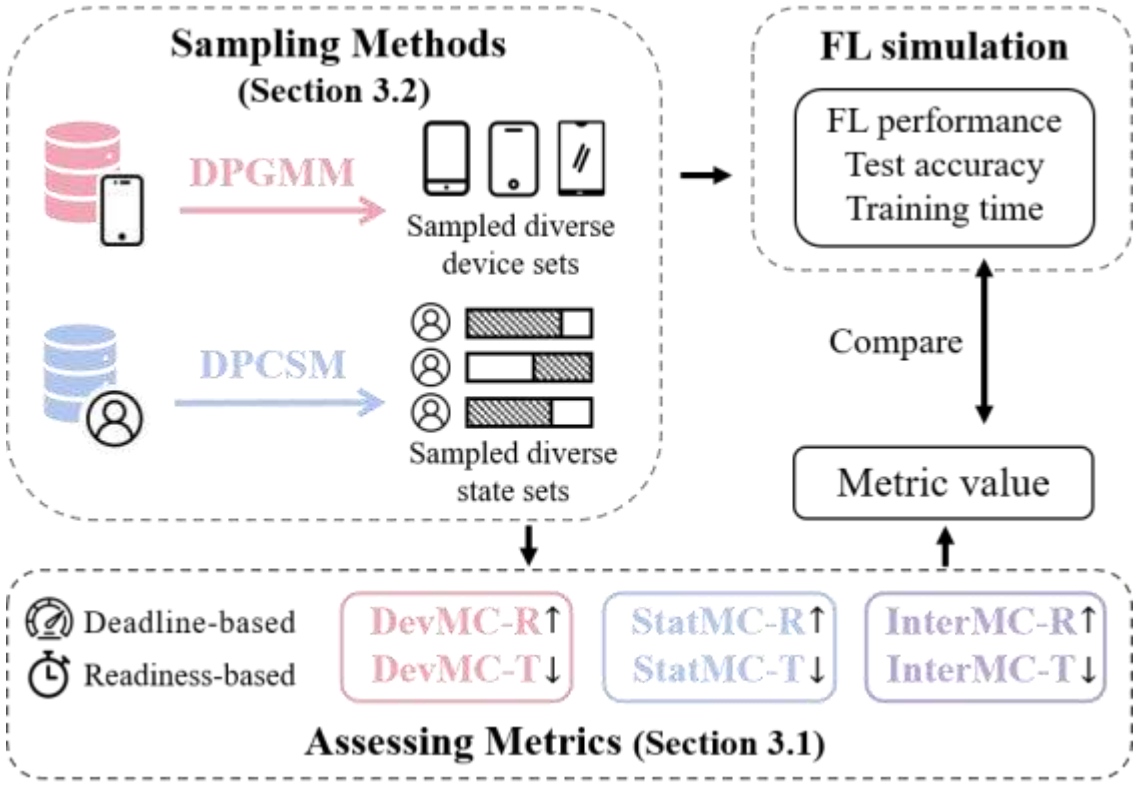
Federated learning faces heterogeneity challenges



Efforts to investigate device and state heterogeneity are limited due to the lack of benchmarks and evaluation metrics reflecting these dimensions of heterogeneity in the real world. Prior studies either use simulated environments or are limited to a small set of real datasets.

FLHetBench: Device and State Evaluation Benchmark

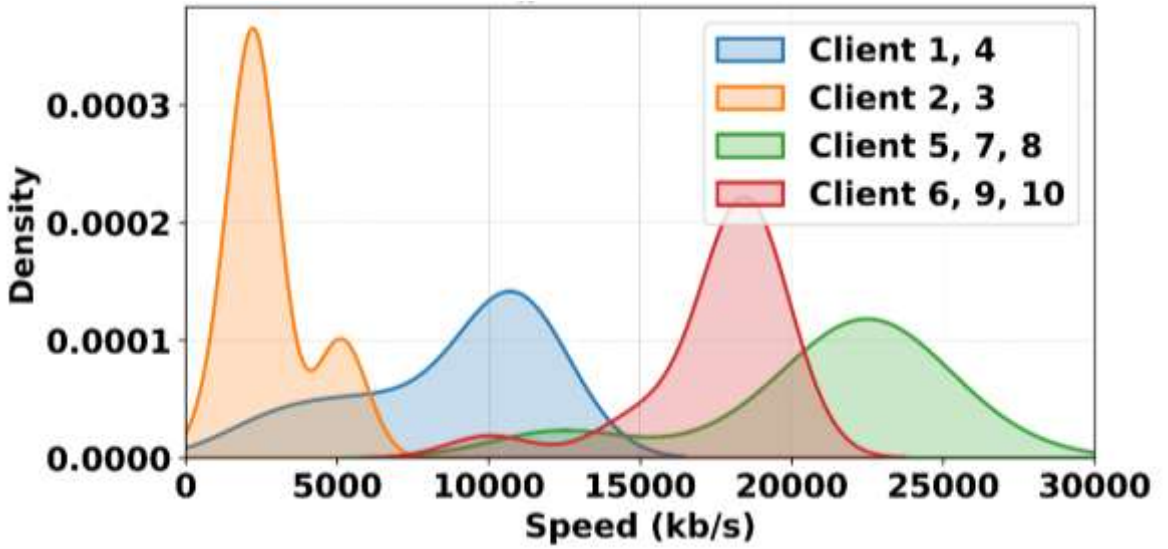
- We introduce **FLHetBench**, the first real-world device and state heterogeneity evaluation benchmark in FL. Our FLHetBench involves:
- **Methods to simulate various device and state heterogeneity of real-world FL**
  - Two innovative Dirichlet process-based sampling methods
  - DPGMM for continuous device data
  - DPCSM for discrete state data.
- **Metrics to quantitively assess degrees of device and state heterogeneity**
  - Several isolated and interplay metrics, based on Monte Carlo (MC) simulations and clients' successful participating ratio, to assess device/state heterogeneity in FL.



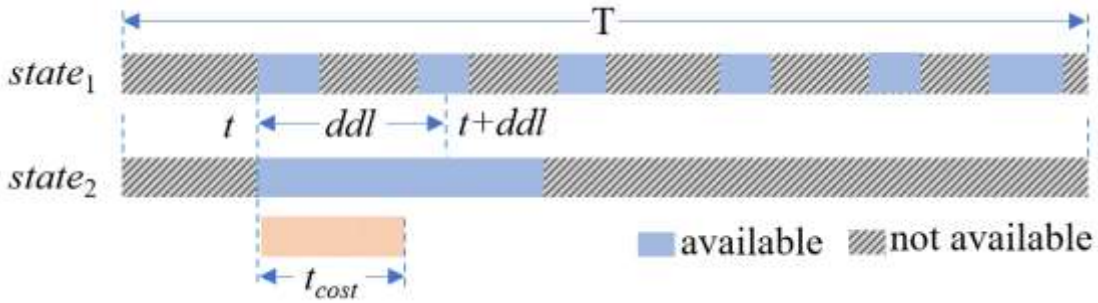
Metrics for Assessing Device/State Heterogeneity

Metrics for Assessing Device/State Heterogeneity

How to quantitively assess degrees of device and state heterogeneity?



Device data : gaussian distribution



State data: discrete available duration

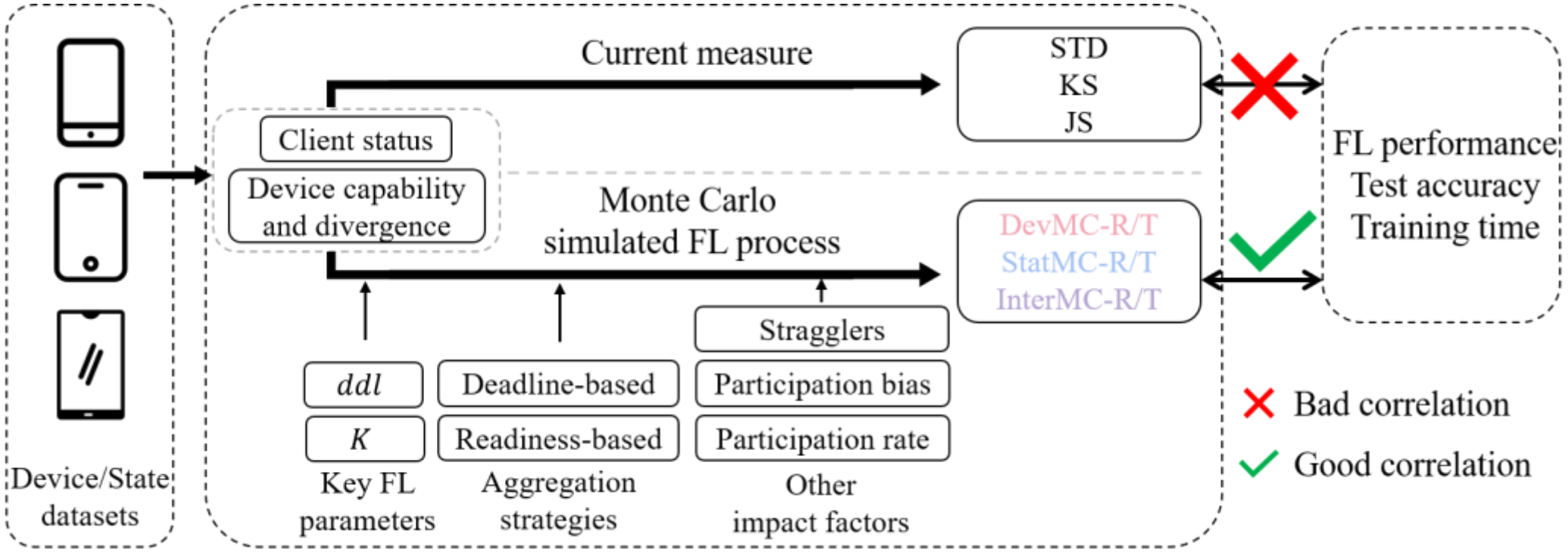
Q: Could we directly use data heterogeneity assessment metric?
Such as Jensen–Shannon divergence , pair-wise KS statistics, STD?
A: **FL performance is not directly related to the statistical divergence!**

STD	KS	Training cost of FedAVG
18.78	0.380	183h
1763.71	0.607	534h
4370.31	0.801	64,901h

Metrics for Assessing Device/State Heterogeneity

How to quantitively assess degrees of device and state heterogeneity?

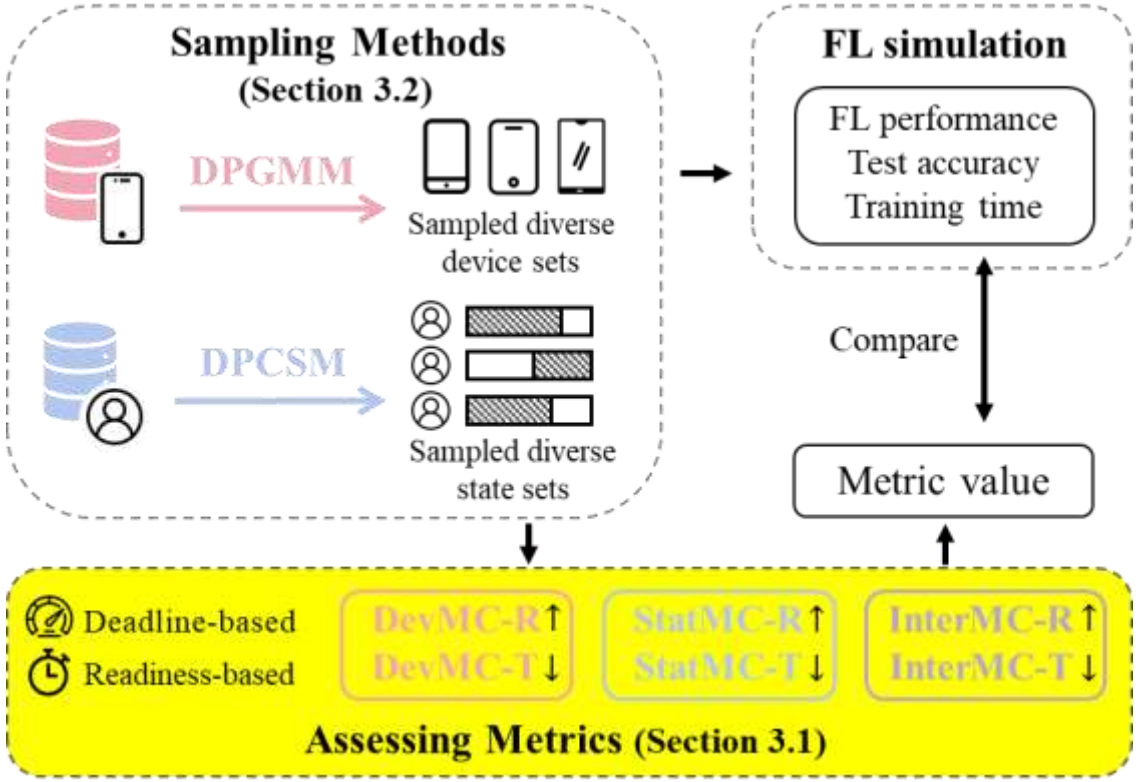
- **In fact, the impact of a real-world device/state database on FL is shaped by various confounding factors, such as device capacities, device divergence, state status, and server aggregation strategies, etc.**



Metrics for Assessing Device/State Heterogeneity

Isolated and interplay metrics accounting for various factors

- **Monte Carlo based metrics**
  - Therefore, we introduce **DevMC-R/T, StatMC-R/T and InterMC-R/T** to assess isolated device, state and interplay heterogeneity.
  - We propose using Monte Carlo (MC) simulations to mimic the real FL training process, enabling metric estimation while reducing computational costs and accounting for various **confounding factors**.

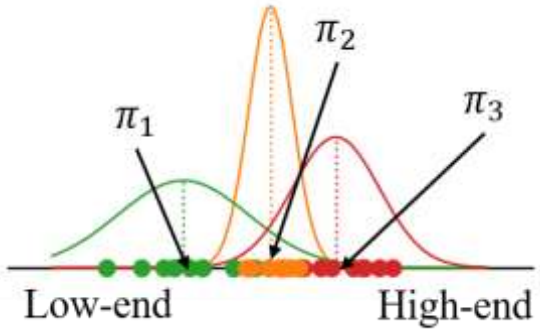


Methods for Sampling Device/State Heterogeneity

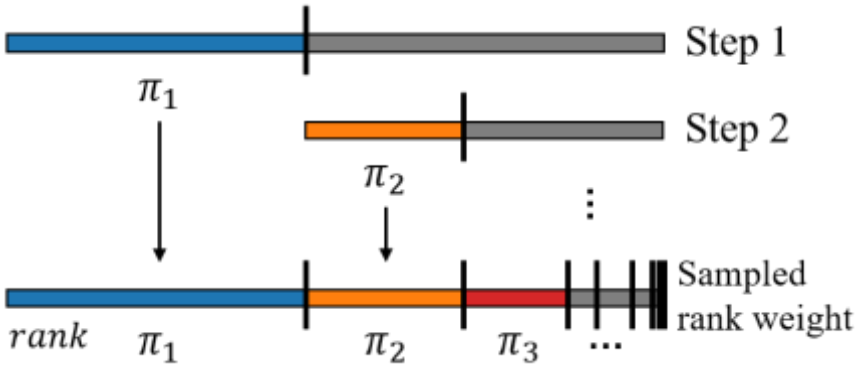
Sampling Methods for various heterogeneity

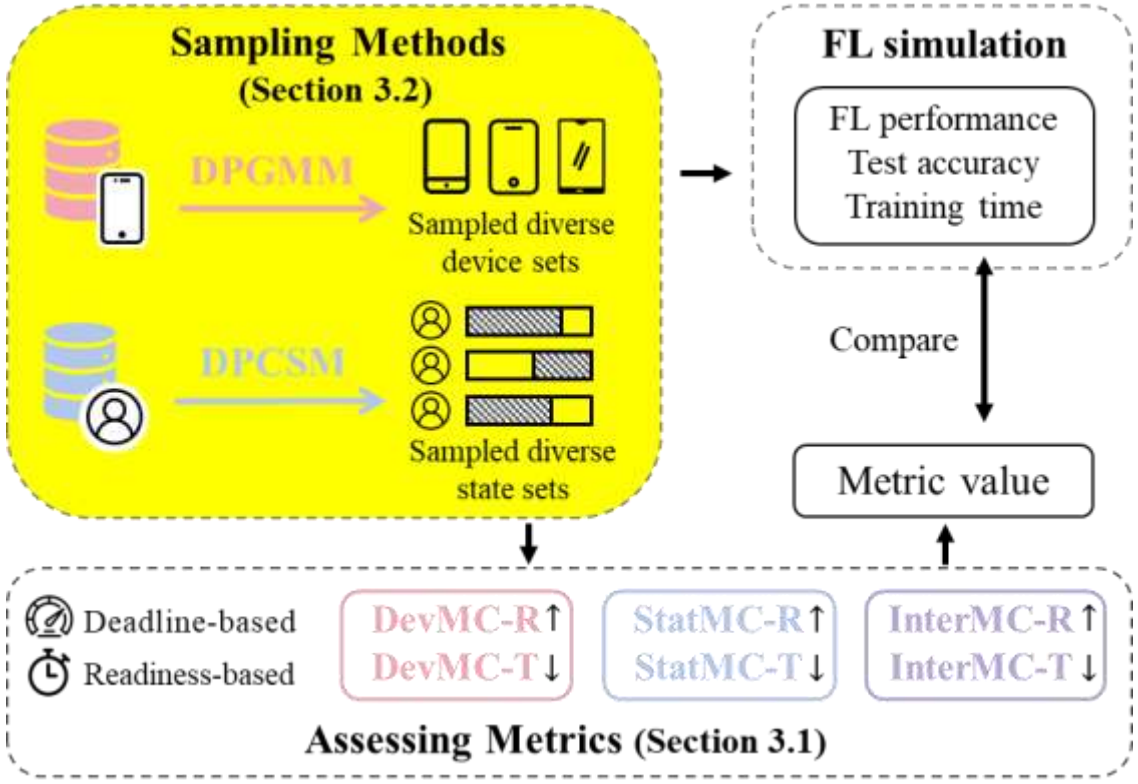
How to simulate varying degrees of device/state heterogeneity?

- **Dirichlet Process Gaussian Mixture Model (DPGMM)**
  - Device database: gaussian mixture model



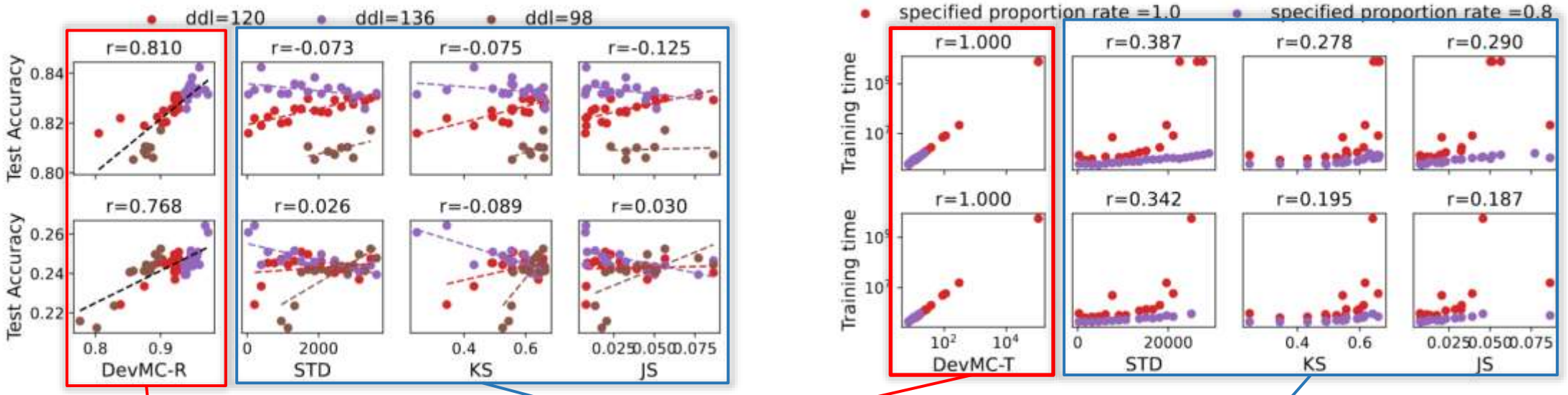
- **Dirichlet Process Construction-based Sampling Method (DPCSM)**





Experiments: Validating Heterogeneity Metrics

The Pearson correlation coefficient r is employed as a quantitative measure of the relationship

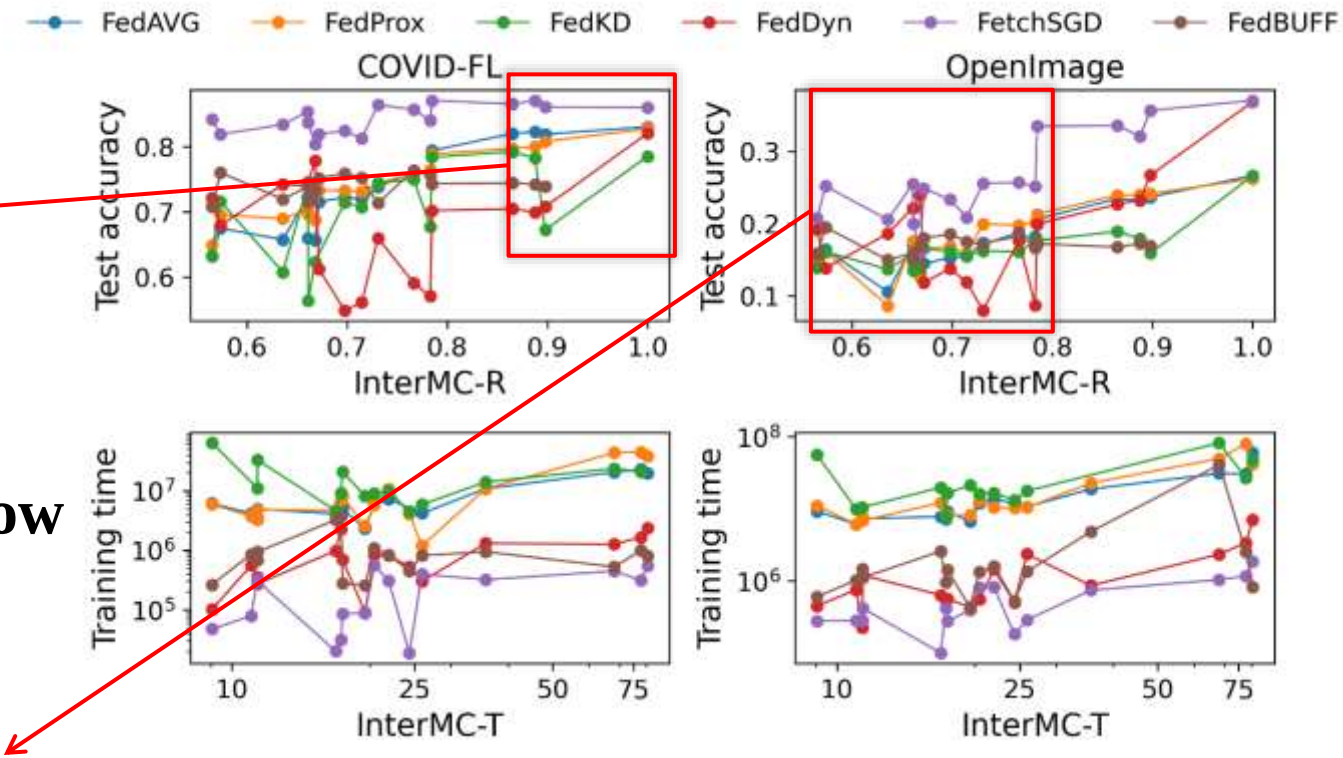


Empirical relationship between metrics and FedAVG test accuracy/training time using deadline-based/readiness-based strategy

DevMC-R/T exhibits a higher correlation (***r* > 0.76/0.95**) than other metrics (STD, KS, JS)

Benchmarking Existing FL Methods with FLHetBench.

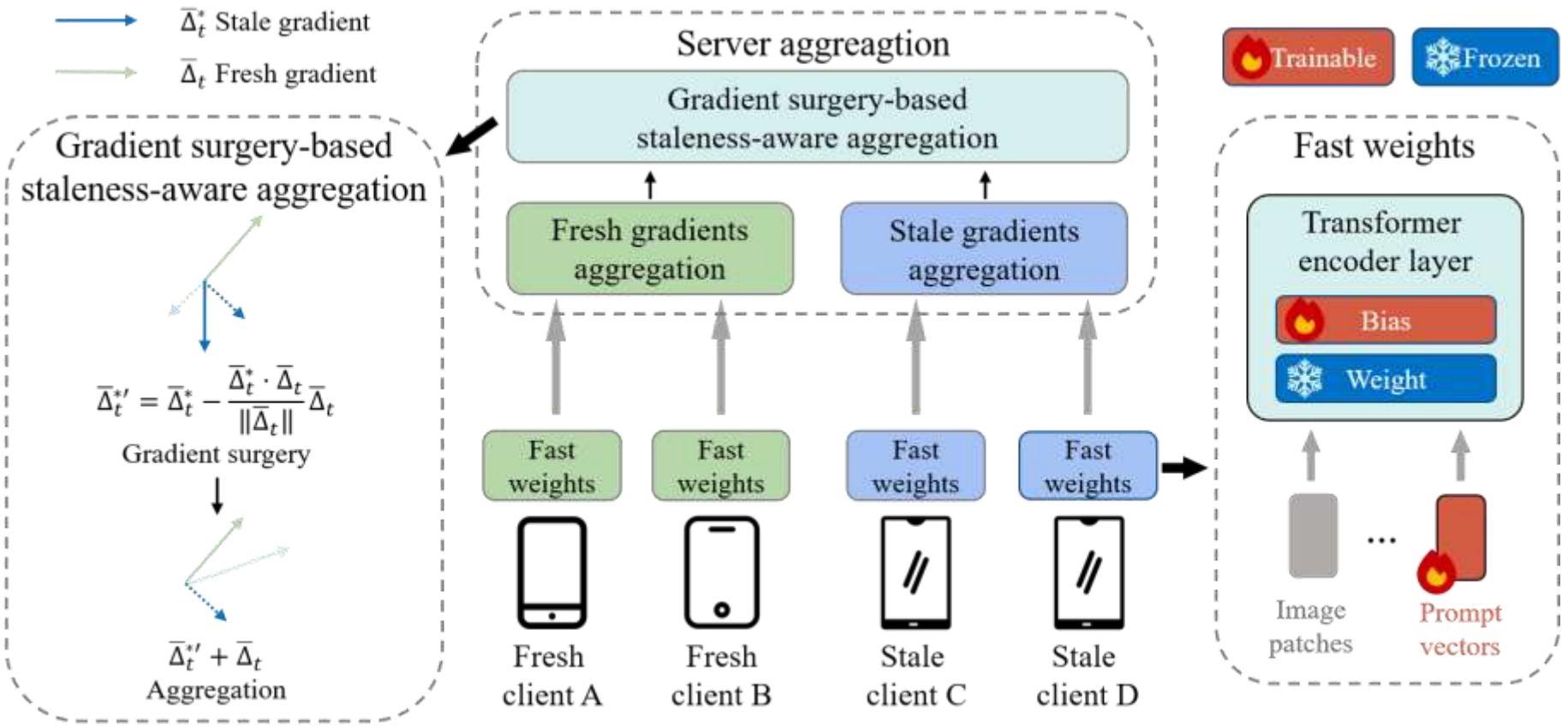
- **Observation 1:**
  - Most methods perform well under mild device/state heterogeneity but struggle with increased device and state heterogeneity.
- **Observation 2:**
  - **The increased wall-clock time** and **low resource utilization of participating clients**, caused by device/state heterogeneity, are the primary factors contributing to the performance degradation of current FL methods in heterogeneous real-world device/state scenarios.



First row: InterMC-R vs. test accuracy for FL algorithms on COVID-FL/OpenImage with deadline-based strategy. Second row: InterMC-T vs. FL training time using readiness-based strategy. InterMC-R=1 denotes no device/state heterogeneity

Solution: Addressing Heterogeneity with BiasPromt+

- Motivated by the above investigation, we introduce **BiasPrompt+**, a novel method employing **gradient surgery-based staleness-aware aggregation** (maximizing resource utility) and **fast weights (minimizing communication/computation costs)** to address device and state heterogeneity in FL.



Solution: Addressing Heterogeneity with BiasPromt+

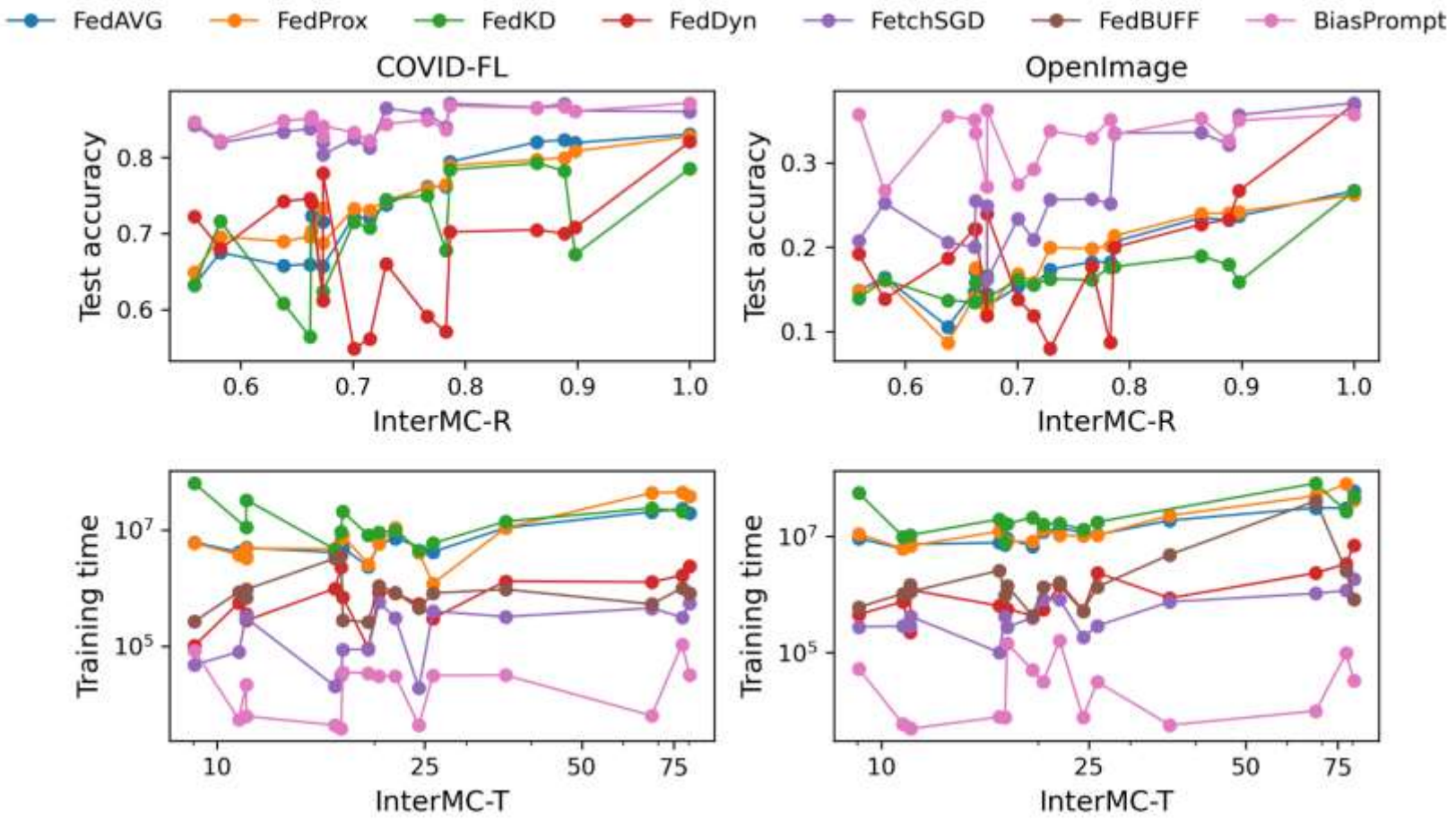


Figure. First row: InterMC-R vs. test accuracy for FL algorithms on COVID-FL/OpenImage with deadline-based strategy. Second row: InterMC-T vs. FL training time using readiness-based strategy. InterMC-R=1 denotes no device/state heterogeneity. **BiasPrompt+ consistently surpasses competing methods.**

Conclusion

- We introduce FLHetBench, a *pioneering* benchmark for evaluating device and state heterogeneity in FL. Our real-world databases, sampling methods, and metrics are released at https://github.com/FLHetBench/code, facilitating future exploration of this pivotal field.

- We conduct the ***first*** comprehensive evaluation of FL on varying degrees of device and state heterogeneity using FLHetBench, revealing that long wall-clock time and low resource utilization of participating clients contribute to the performance degradation of current FL methods in heterogeneous real-world device/state scenarios.

- We propose a simple and efficient method, BiasPrompt+, to mitigate device/state heterogeneity challenges. Extensive experimental results validate the superiority of our BiasPrompt+ over competing methods.