Privacy-Preserving Federated Active Learning for Data Completion in Sparse Crowdsensing

Zhifei Wang, Wenbin Liu, En Wang, and Jie Wu

College of Computer Science and Technology, Jilin University
China Telecom Cloud Computing Research Institute
Department of Computer and Information Sciences, Temple University

Outline

I. Background and Motivation

II. Problem and Framework

III. Method

IV. Performance Evaluation

V. Conclusion

Outline

- I. **Background and Motivation**
- II. Problem and Framework
- III. Method
- IV. Performance Evaluation
- V. Conclusion

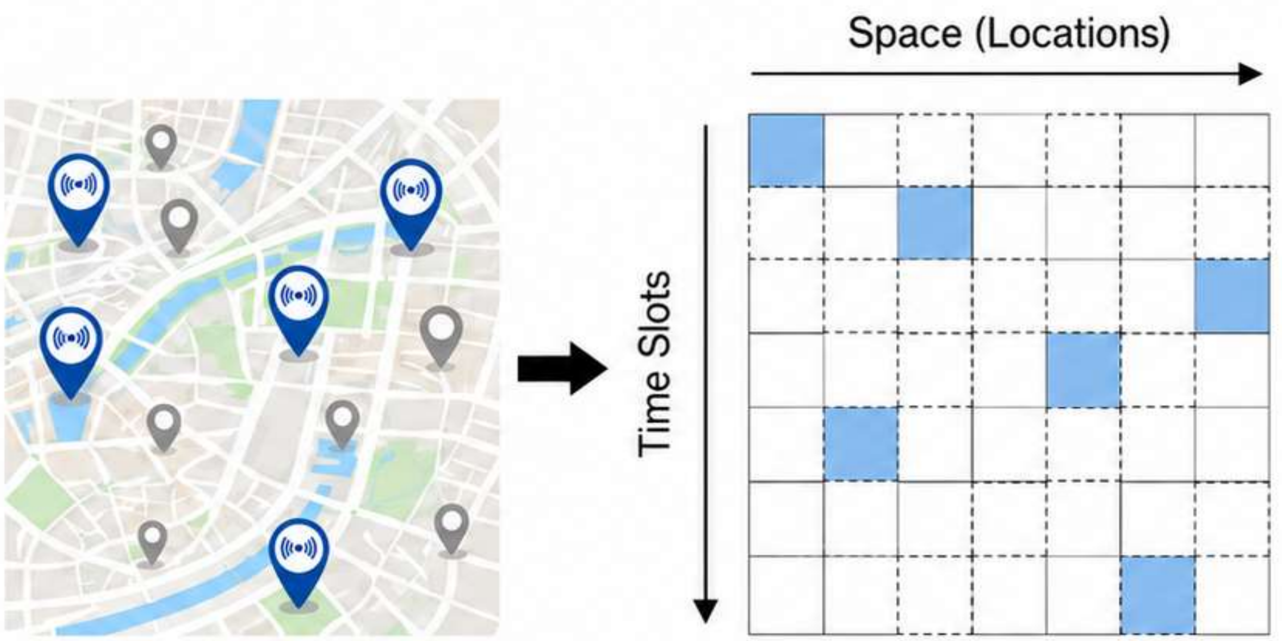
Mobile CrowdSensing (MCS)

- Recruit users with mobile devices
- Large-scale urban sensing at low cost
- Applications: air quality, traffic, noise, humidity
- Data collected across time and space

MCS enables large-scale sensing via mobile users.

Sparse Crowdsensing

- Only a few locations and time slots are sensed
- Budget and device constraints limit observations
- Sparse data contain many missing entries
- Accurate completion is therefore essential



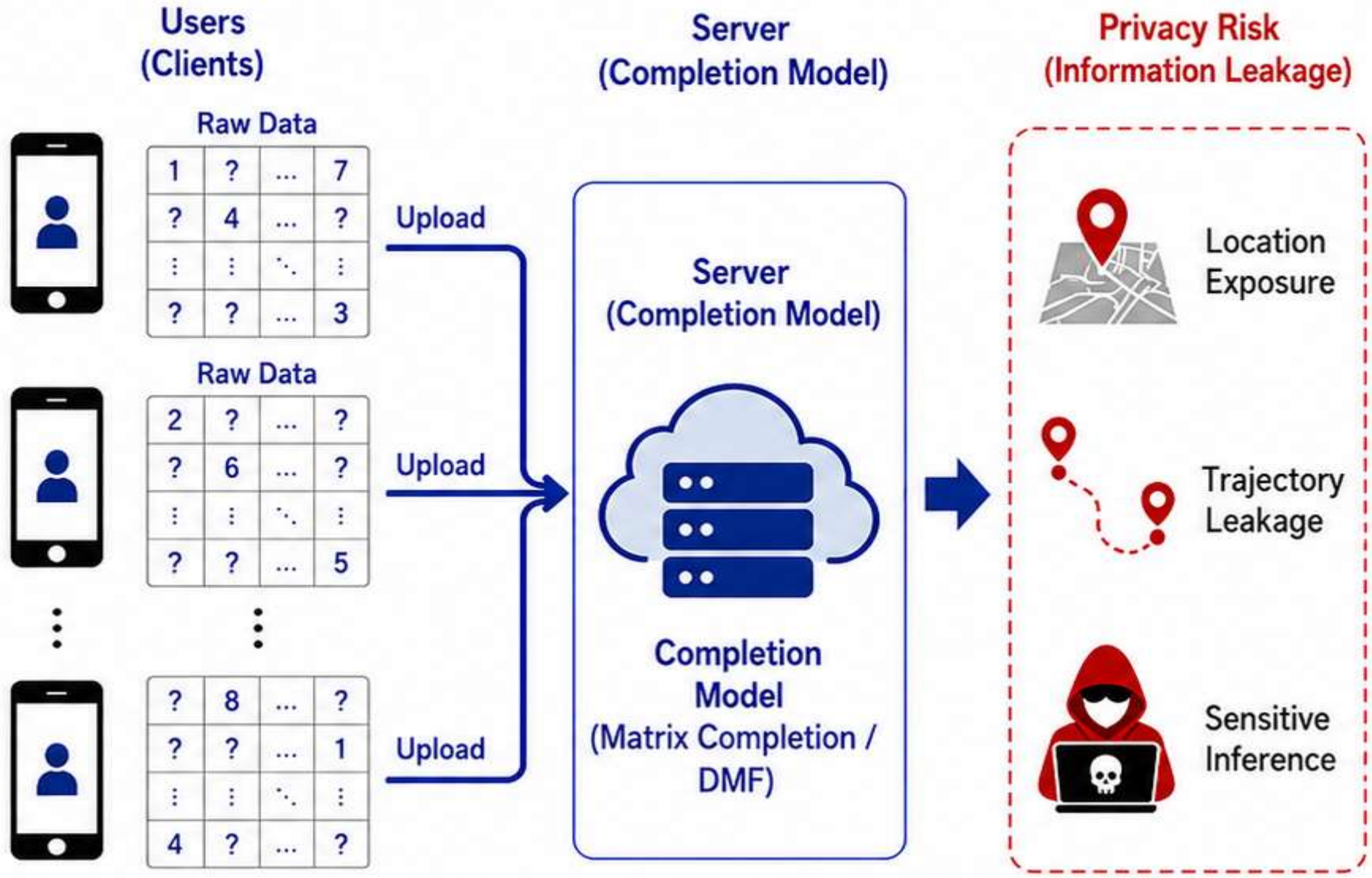
Sparse sensing in the city

Sparse spatiotemporal observations

Sparse observations make direct sensing incomplete.

Centralized Completion: Privacy Risk

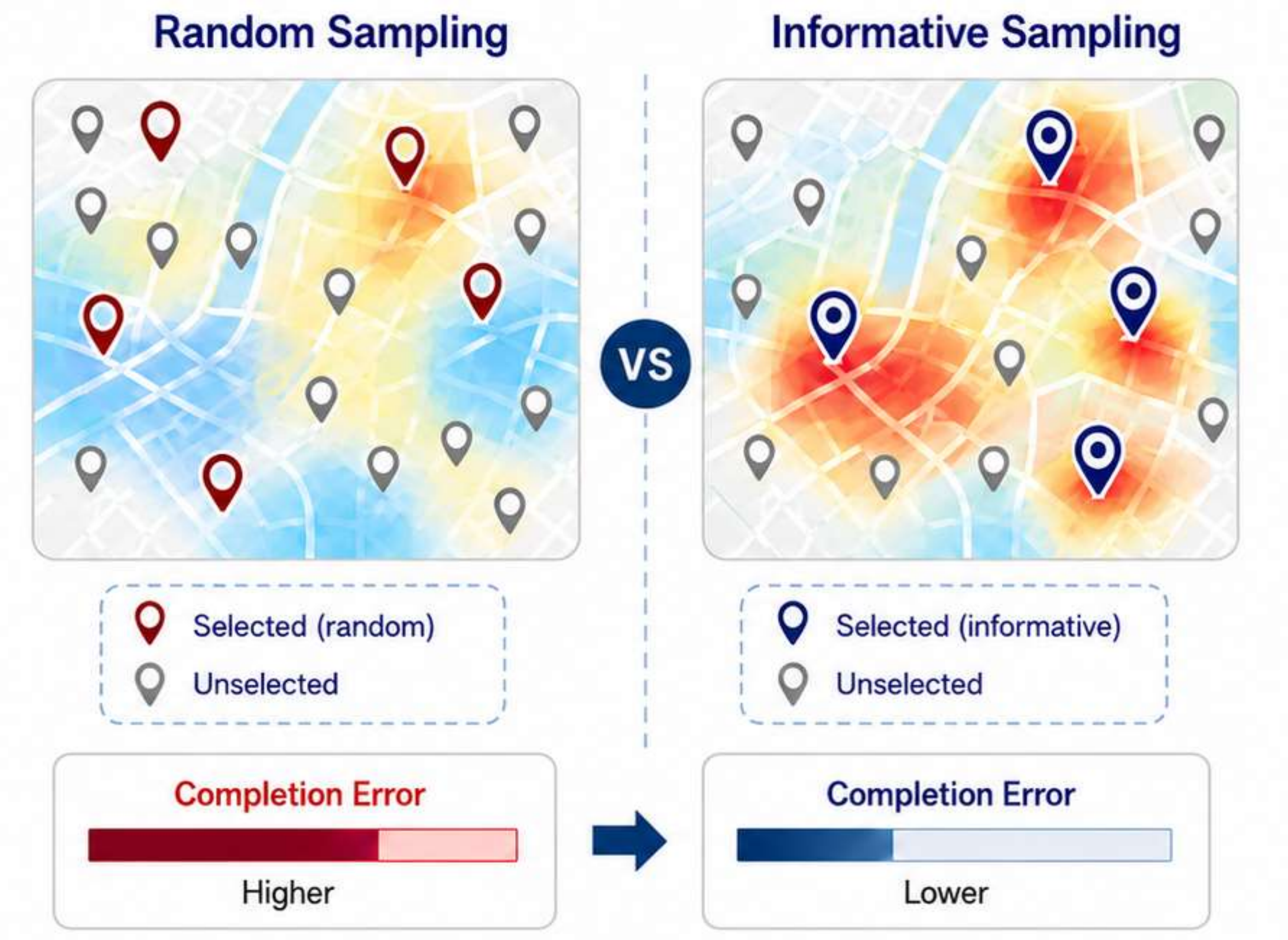
- Centralized completion can recover missing data
- Users upload raw sensing data to a server
- The server aggregates data for completion
- Locations and trajectories may be exposed



Raw uploads improve completion but expose privacy.

Inefficient Random Sampling

- Only a few locations can be sensed
- Random/fixed sampling may pick low-value points
- It ignores uncertainty and informativeness
- Accuracy improves only slightly under limited budget



Sampling should focus on informative locations.

Challenge

- **Privacy-preserving completion**
  - How to train completion models without uploading raw sensing data?
- **Local sparse modeling**
  - How to capture useful spatiotemporal dependencies from client-local sparse data?
- **Budget-efficient sampling**
  - How to select high-value observations under a limited sensing budget?

Outline

Problem Formulation

- A server and K distributed clients participate in Sparse Crowdsensing.
- Each client k observes a sparse matrix: $Y_{obs}^{(k)} = Y^{(k)} \odot C^{(k)}$.
- Goal: collaboratively learn a global completion model $f_G(\cdot)$.
- Constraints:
  1. raw data remain local,
  2. uploaded updates satisfy differential privacy,
  3. and active sampling is limited by budget B.

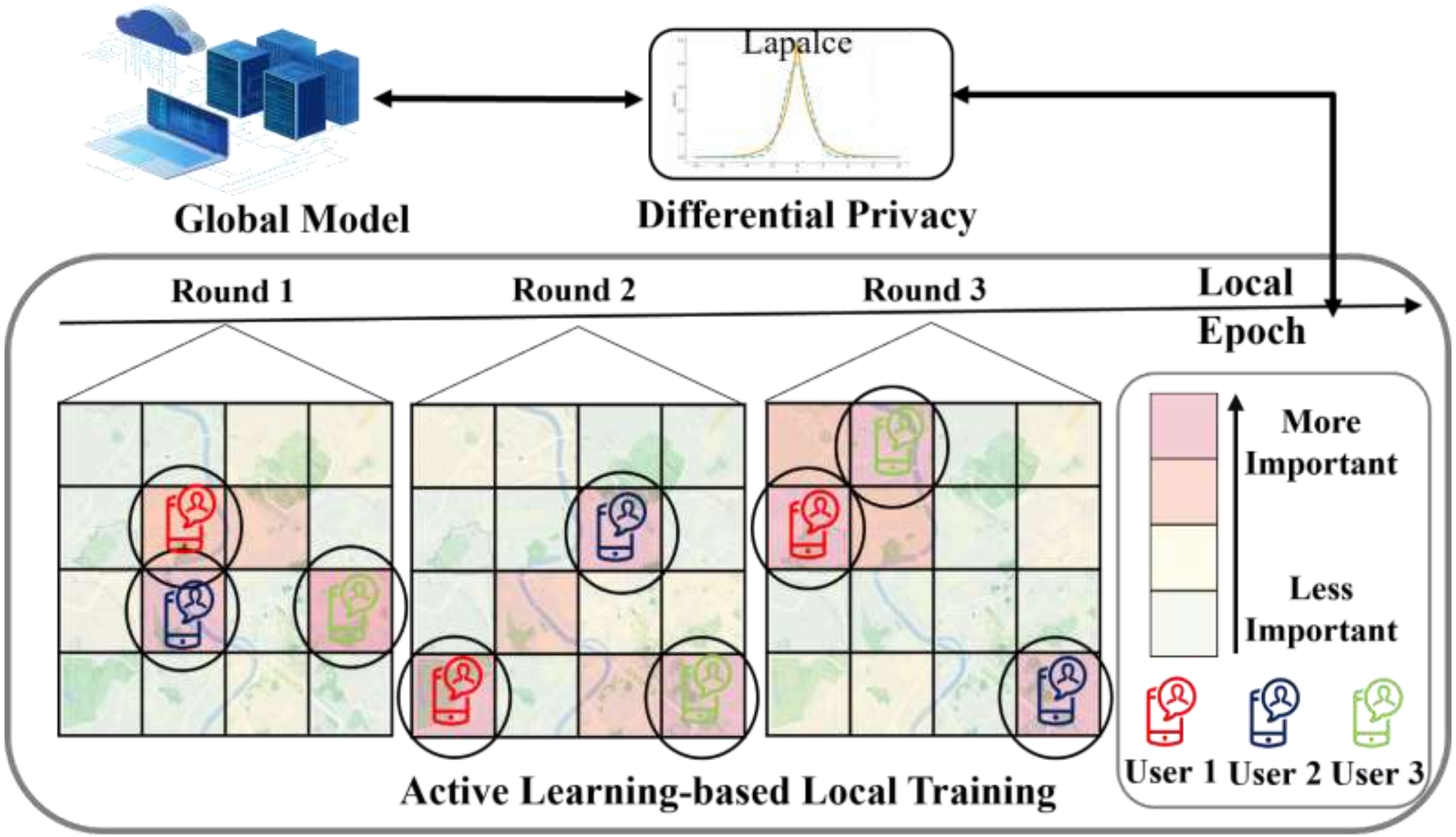
Distributed objective

minimize completion error
+
privacy regularization
under active learning budget

$\longrightarrow$

1. accurate completion
2. privacy preservation
3. sampling efficiency

Framework Overview



Federated training + Differential privacy + Active learning-based local sampling

Outline

I. Background and Motivation

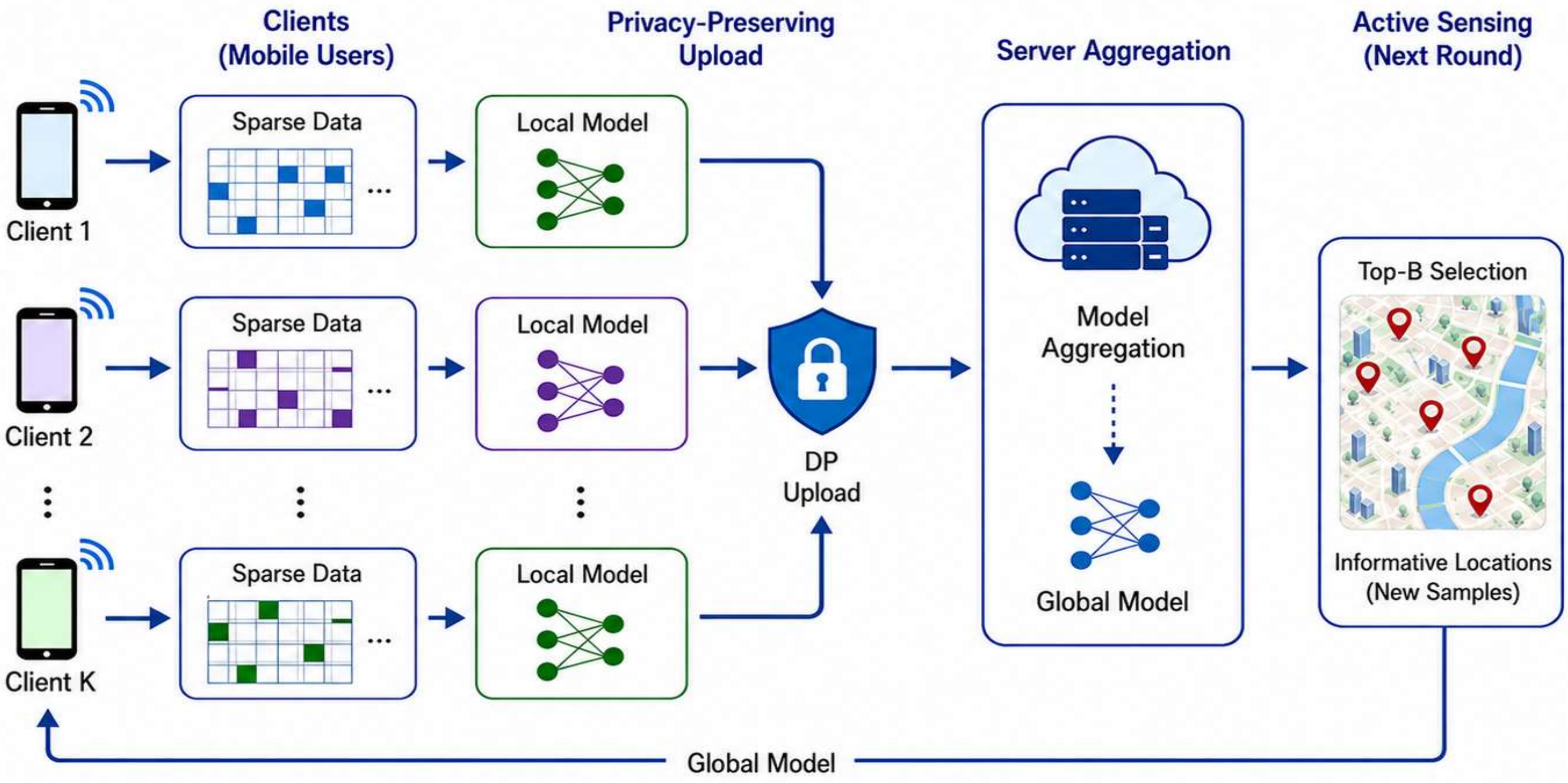
II. Problem and Framework

III. Method

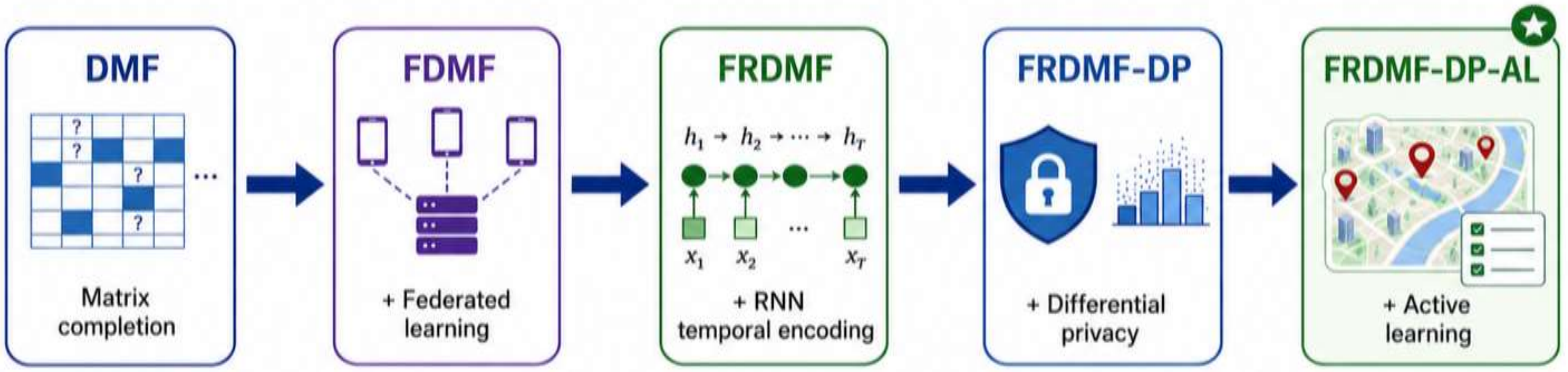
IV. Performance Evaluation

V. Conclusion

Overall Structure



Method Evolution



+federation

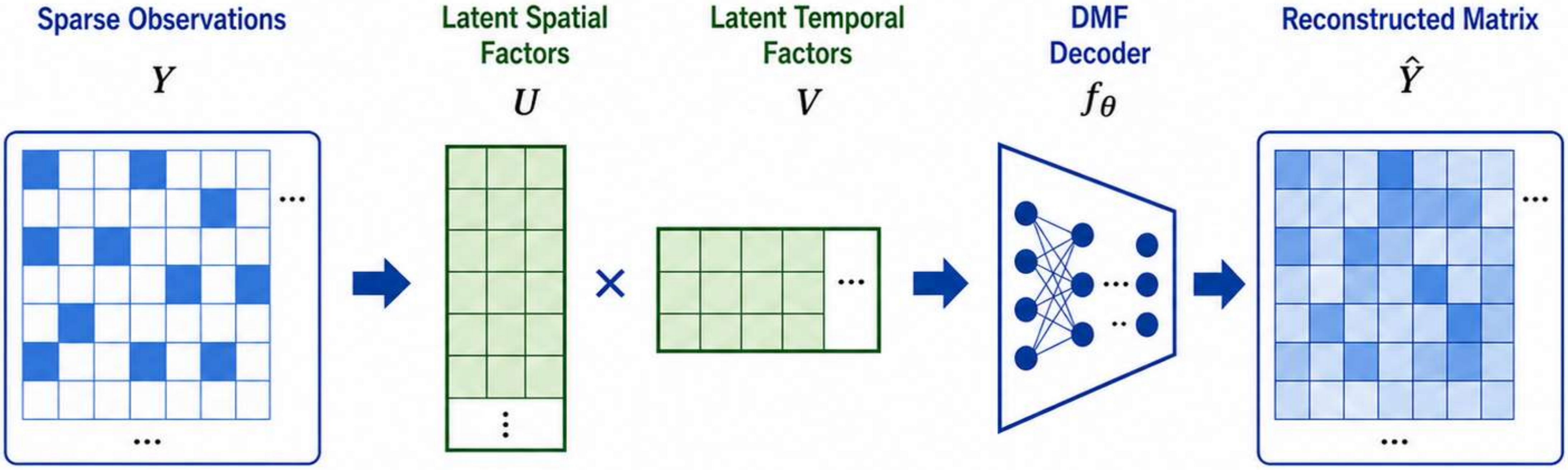
+temporal modeling

+privacy protection

+informative sampling

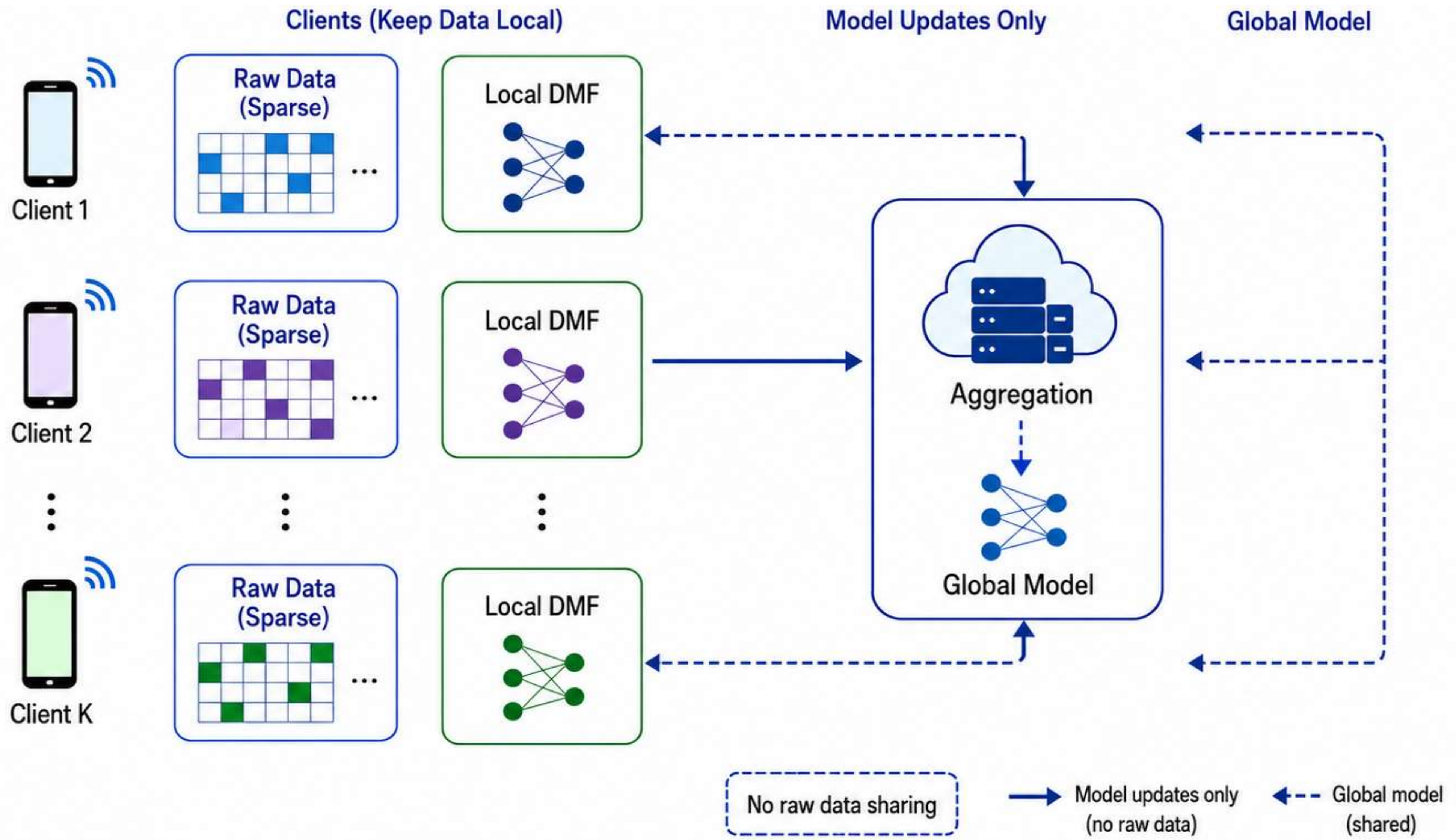
Each stage adds a new capability to the previous model

Deep Matrix Factorization (DMF)

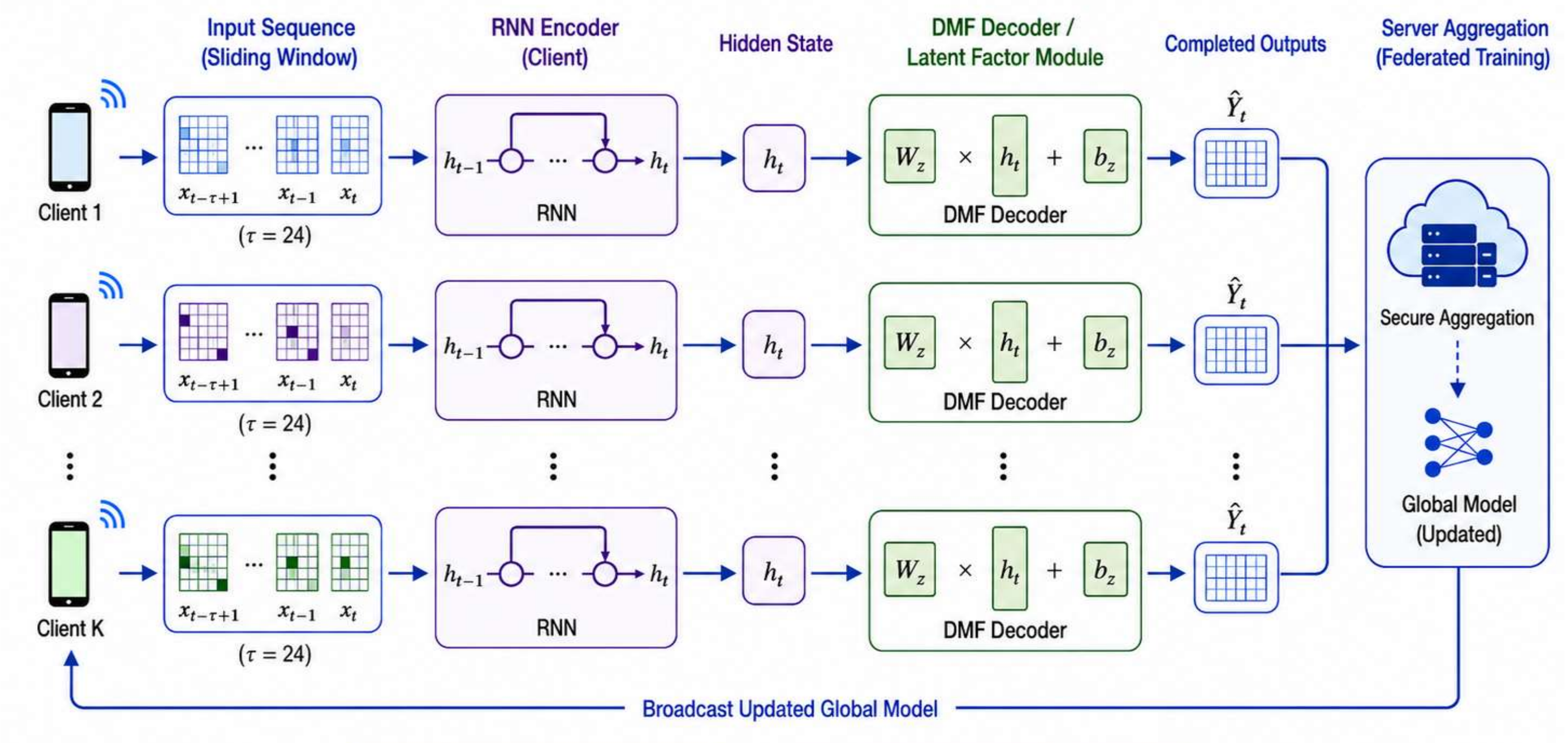


$$\mathcal{L} = \|(Y - \hat{Y}) \odot C\|_F^2$$

Federated Deep Matrix Factorization (FDMF)



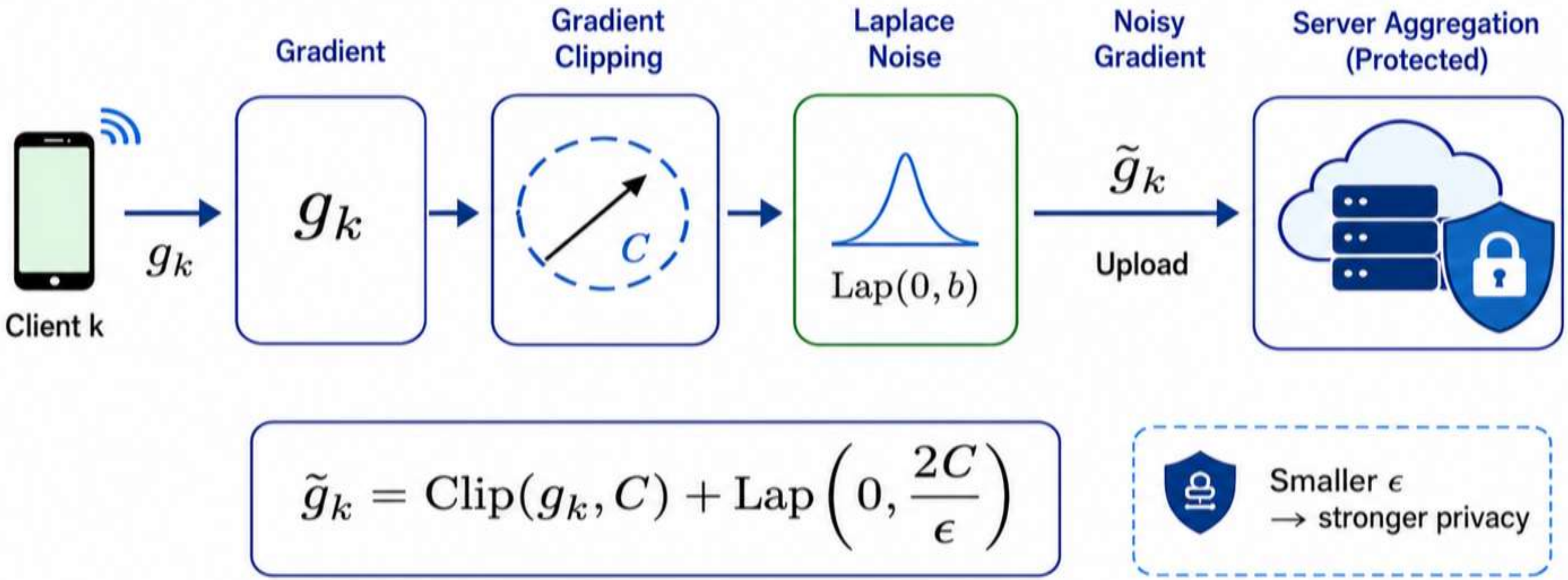
Federated RNN-enhanced DMF (FRDMF)



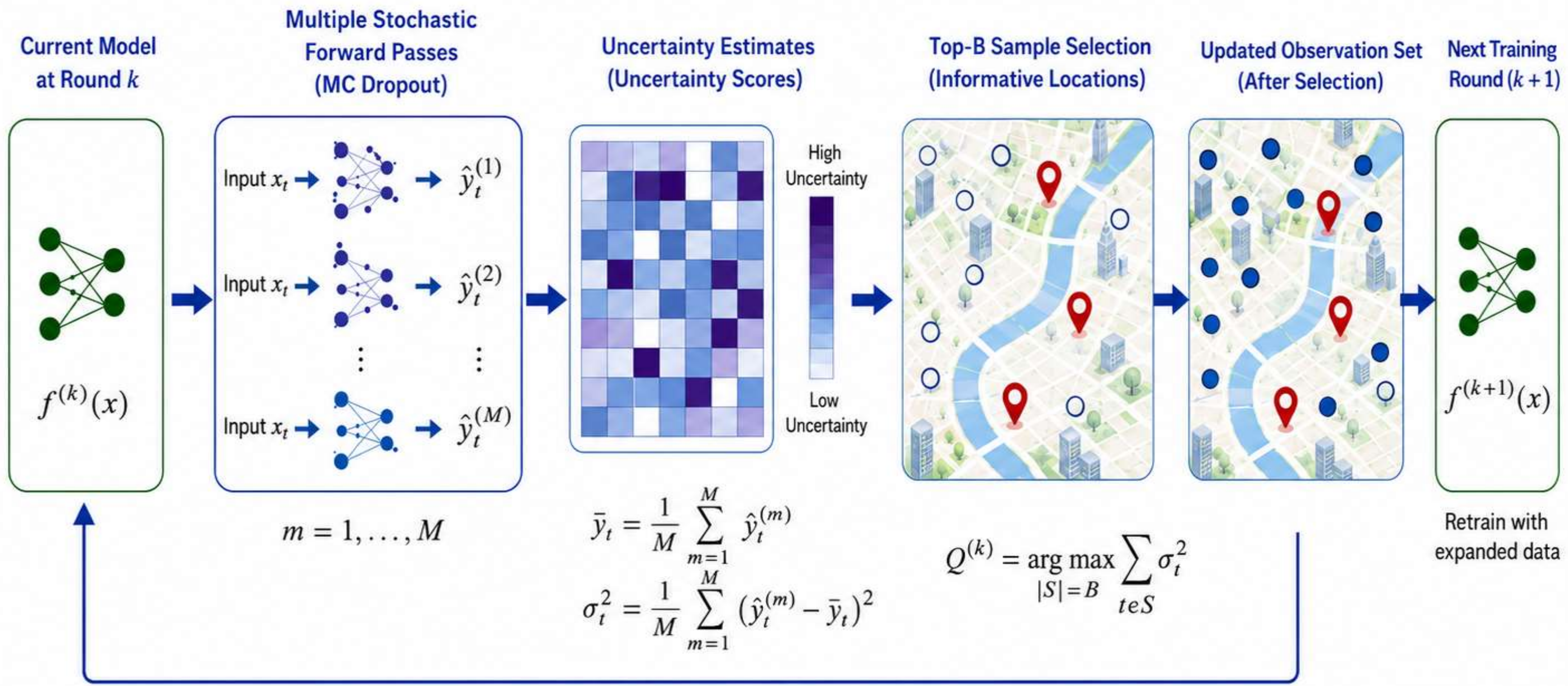
$$h_t = \mathrm{RNN}(x_t, h_{t-1})$$

$$\hat{Y}_t = f_\theta(W_z h_t + b_z)$$

Differential Privacy Mechanism (FRDMF-DP)



Active Learning-based Sample Selection (FRDMF-DP-AL)



Outline

I. Background and Motivation

II. Problem and Framework

III. Method

IV. Performance Evaluation

V. Conclusion

Datasets and Compared Models

Datasets

PM25	Othree	Humidity	MyTraffic
Beijing PM2.5 36 stations	Beijing ozone 36 stations	EPFL campus 0.5-hour interval	Taxi speed data 50 roads

Compared Models

Centralized	Federated	Privacy-preserving
DMF RNN-DMF	FDMF FRDMF	FRDMF-DP FRDMF-DP-AL

Main results

PERFORMANCE COMPARISON OF DIFFERENT MODELS.

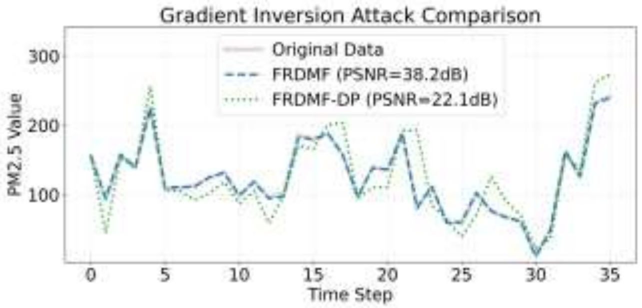
Model	RMSE				Time Cost (s)				ASR (%)
	MyTraffic	PM25	Othree	Humidity	MyTraffic	PM25	Othree	Humidity	
DMF	23.47	37.35	8.20	1.86	2.13	2.07	2.07	2.89	100*
FDMF	23.69	29.09	7.28	3.22	1.20	1.09	1.28	1.72	24
RNN-DMF	103.92	94.63	78.26	28.19	183.80	35.40	148.93	37.52	100*
FRDMF	24.44	32.48	16.31	4.35	952.50	171.14	670.93	55.01	31
FRDMF-DP	50.72	48.81	21.76	12.25	410.86	79.23	267.17	55.84	19
FRDMF-DP-AL	**37.10**	**40.27**	**16.26**	**7.15**	**822.09**	**112.86**	**607.93**	**206.22**	**19**

Key observations

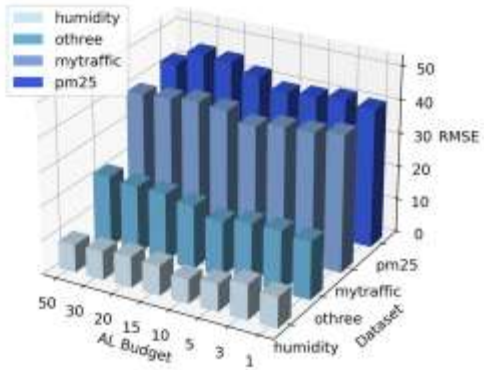
1. Federated models avoid centralized raw data storage.
2. Differential privacy reduces attack strength but introduces utility loss.
3. Active learning recovers accuracy under the same privacy mechanism.

Othree: FRDMF-DP → FRDMF-DP-AL
RMSE 21.76 → 16.26 (−25.3%)

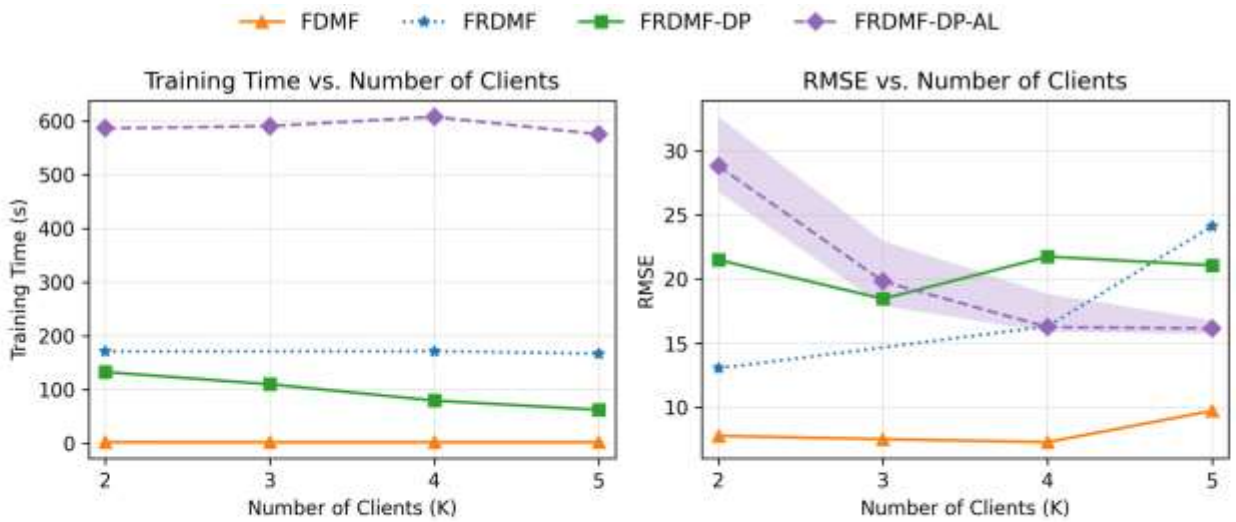
Privacy Robustness & AL Efficiency



(a) Gradient Inversion Attack Comparison.



(b) Performance evaluations on AL Budget

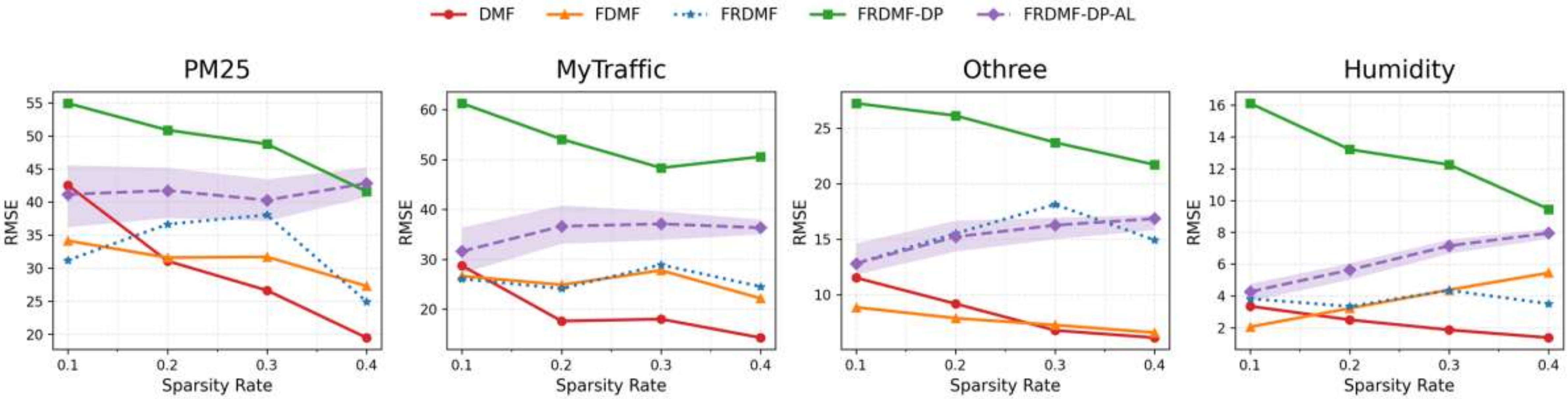


Gradient inversion attack

Scalability with different clients

1. Differential privacy lowers PSNR from 38.2 dB to 22.1 dB, reducing gradient leakage risk.
2. Active learning improves accuracy without increasing privacy risk or communication cost.
3. The federated framework scales to more clients with stable RMSE.

Robustness under Sparsity



Sparse adaptability

1. Federated models capture localized spatiotemporal features through client-side modeling.
2. On PM25, FRDMF-DP-AL fluctuates only 4.08% when sparsity changes from 0.1 to 0.4.
3. RNN-based temporal modeling is important for sparse scenarios.

Ablation Study

ABLATION EXPERIMENT: THE NECESSITY OF RNN, DP AND AL FOR MODEL PERFORMANCE (ε=18, SPARSITY=0.3).

Model	RMSE				ASR (%)			
	MyTraffic	PM2.5	Othree	Humidity	MyTraffic	PM2.5	Othree	Humidity
RNN-DMF	103.92	94.63	78.26	28.19	100	99	100	97
FDMF-DP	25.16	33.12	12.87	12.41	25	26	12	19
FRDMF	24.44	32.48	16.31	4.35	32	35	31	33
FRDMF-DP	50.72	48.81	21.76	12.25	26	28	19	22
FRDMF-DP-AL	37.10	40.27	16.26	7.15	26	28	19	22

Necessity of RNN, DP and AL

1. RNN is crucial for temporal dependencies and sparse adaptability.
2. DP reduces privacy leakage by perturbing uploaded updates.
3. AL compensates for the utility loss caused by privacy noise.

Outline

I. Background and Motivation

II. Problem and Framework

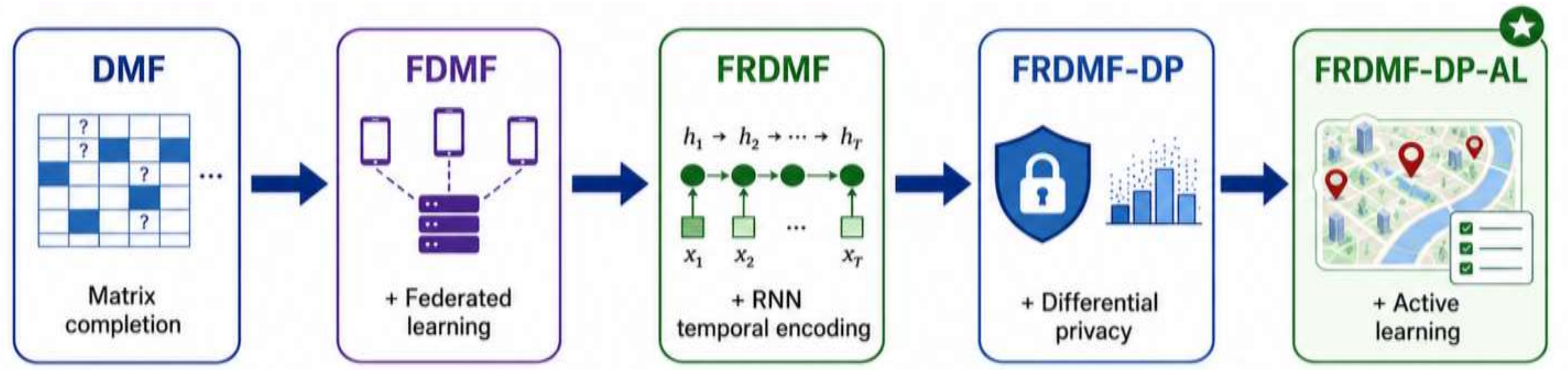
III. Method

IV. Performance Evaluation

V. Conclusion

Conclusion

- **Problem:** ***Privacy-preserving Data Completion*** in Sparse Crowdsensing
- **Framework:** Federated learning with localized temporal modeling and global matrix factorization.
- **Privacy:** ε-differential privacy suppresses gradient inversion attacks.
- **Sampling:** Active learning selects high-value observations under budget.
- **Evaluation:** Four datasets show a strong privacy–utility–efficiency trade-off.



Thank you!

Q&A

zfwang22@mails.jlu.edu.cn