NEURAL INFORMATION PROCESSING SYSTEMS

Diffusion Federated Dataset

[1]Seok-Ju Hahn and [2]Junghye Lee

[1]Argonne National Laboratory

[2]Seoul National University

TL; DR

~~Learning~~

- Synthetic Data from Federated Inference of Pre-trained Diffusion Models

TL; DR

~~Learning~~

- Synthetic Data from Federated Inference of Pre-trained Diffusion Models
  - No parameter exchange

  ✓ Other signals needs to be exchanged

TL; DR

- Synthetic Data from Federated ~~Learning~~ Inference of Pre-trained Diffusion Models

TL; DR

- Synthetic Data from Federated ~~Learning~~ Inference of Pre-trained Diffusion Models
  - No local update is required during communication
  - ✓ Local inference is required instead

TL; DR

- **Synthetic Data** from Federated ~~Learning~~ Inference of Pre-trained Diffusion Models
  - Synthetic data in FL be from **a mixture** of local distributions

TL; DR

- **Synthetic Data** from Federated ~~Learning~~ Inference of Pre-trained Diffusion Models
  - Synthetic data in FL be from **a mixture** of local distributions

$$p^\star(\boldsymbol{x}) = \sum_{i=1}^{K} w_i p_i(\boldsymbol{x})$$

TL; DR

- **Synthetic Data** from Federated ~~Learning~~ Inference of Pre-trained Diffusion Models
  - Synthetic data in FL be from **a mixture** of local distributions

$$p^\star(\boldsymbol{x}) = \sum_{i=1}^{K} w_i p_i(\boldsymbol{x})$$

mixing coefficient (unknown)

TL; DR

- **Synthetic Data** from Federated ~~Learning~~ Inference of Pre-trained Diffusion Models
  - Synthetic data in FL be from **a mixture** of local distributions

$$p^\star(\boldsymbol{x}) = \sum_{i=1}^{K} w_i p_i(\boldsymbol{x})$$

Usually set as

$w_i \propto n_i$ or $w_i = \frac{1}{K}$

Scope

- Cross-silo FL setting

Scope

- Cross-silo FL setting
  - Where a moderate number of reliable, and known, and addressable clients participate[1]

[1] Advances and Open Problems in Federated Learning (Kairouz et al., 2019)

Scope

- Cross-silo FL setting
  - Where a moderate number of reliable, and known, and addressable clients participate[1]
  - Where each client has sufficient computing power

[1] Advances and Open Problems in Federated Learning (Kairouz et al., 2019)

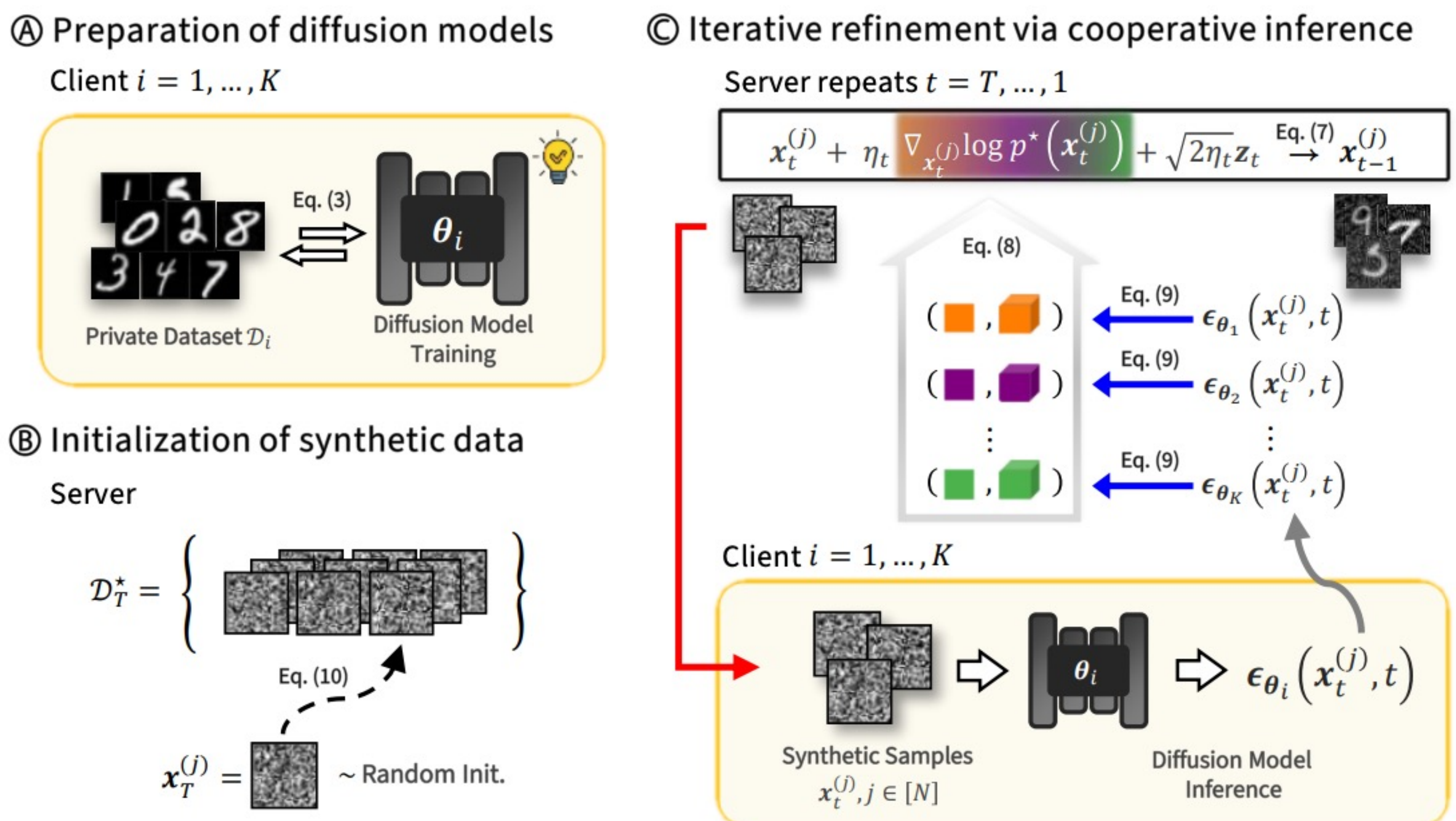
Scope

- Cross-silo FL setting
  - Where a moderate number of reliable, and known, and addressable clients participate[1]
  - Where each client has sufficient computing power
  - Where each client suffers from low diversity of data or low sample size[2]

[1] Advances and Open Problems in Federated Learning (Kairouz et al., 2019)
[2] Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings (du Terrail et al., 2022)

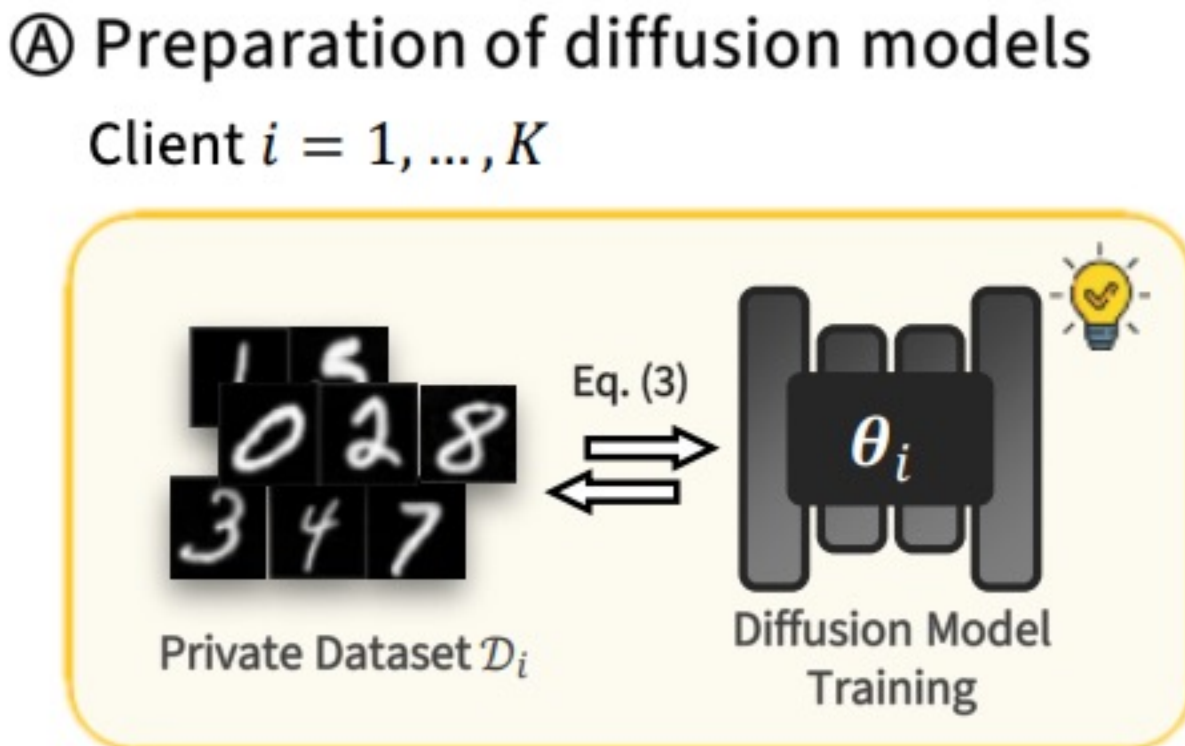
Overview

- Synthetic Data from Federated Inference of Pre-trained Diffusion Models



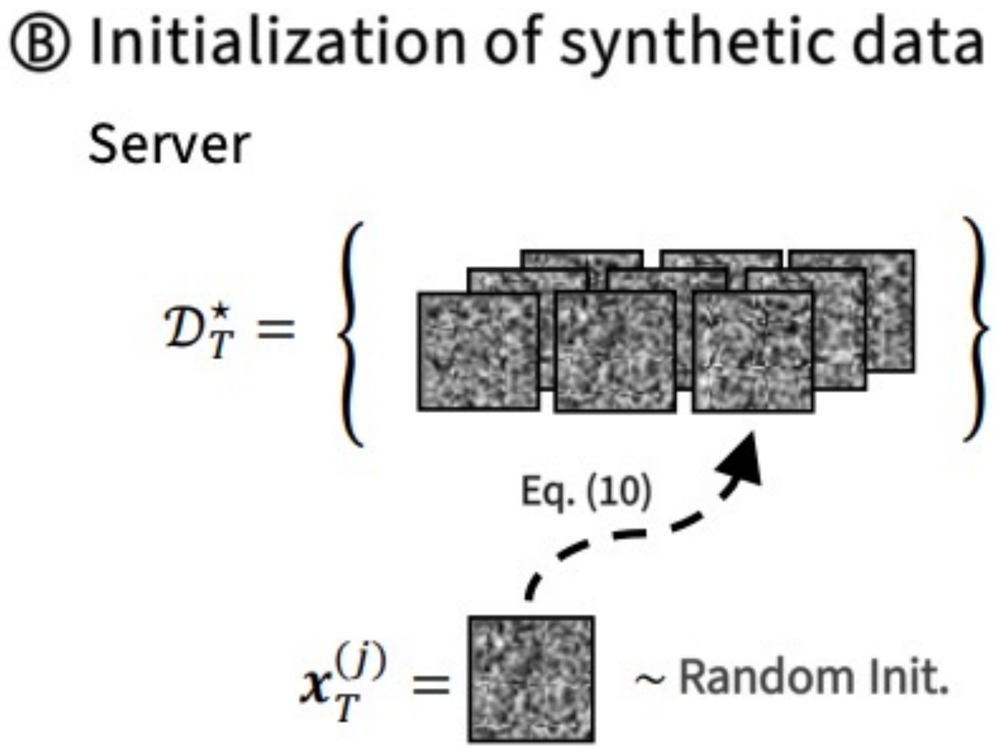
Breakdown

- A) Each client trains its own local diffusion model



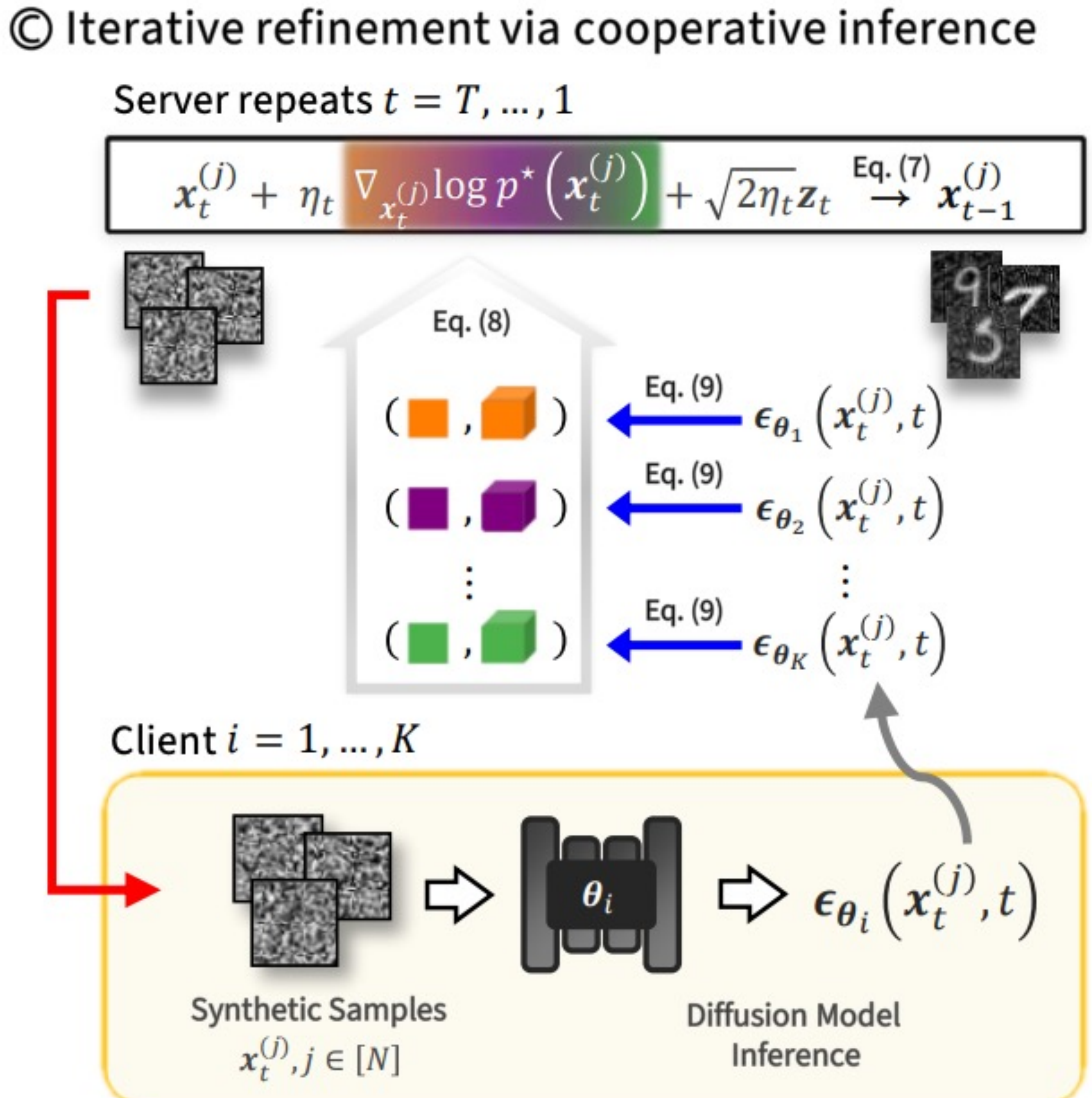
Breakdown

- B) Server randomly initializes synthetic dataset



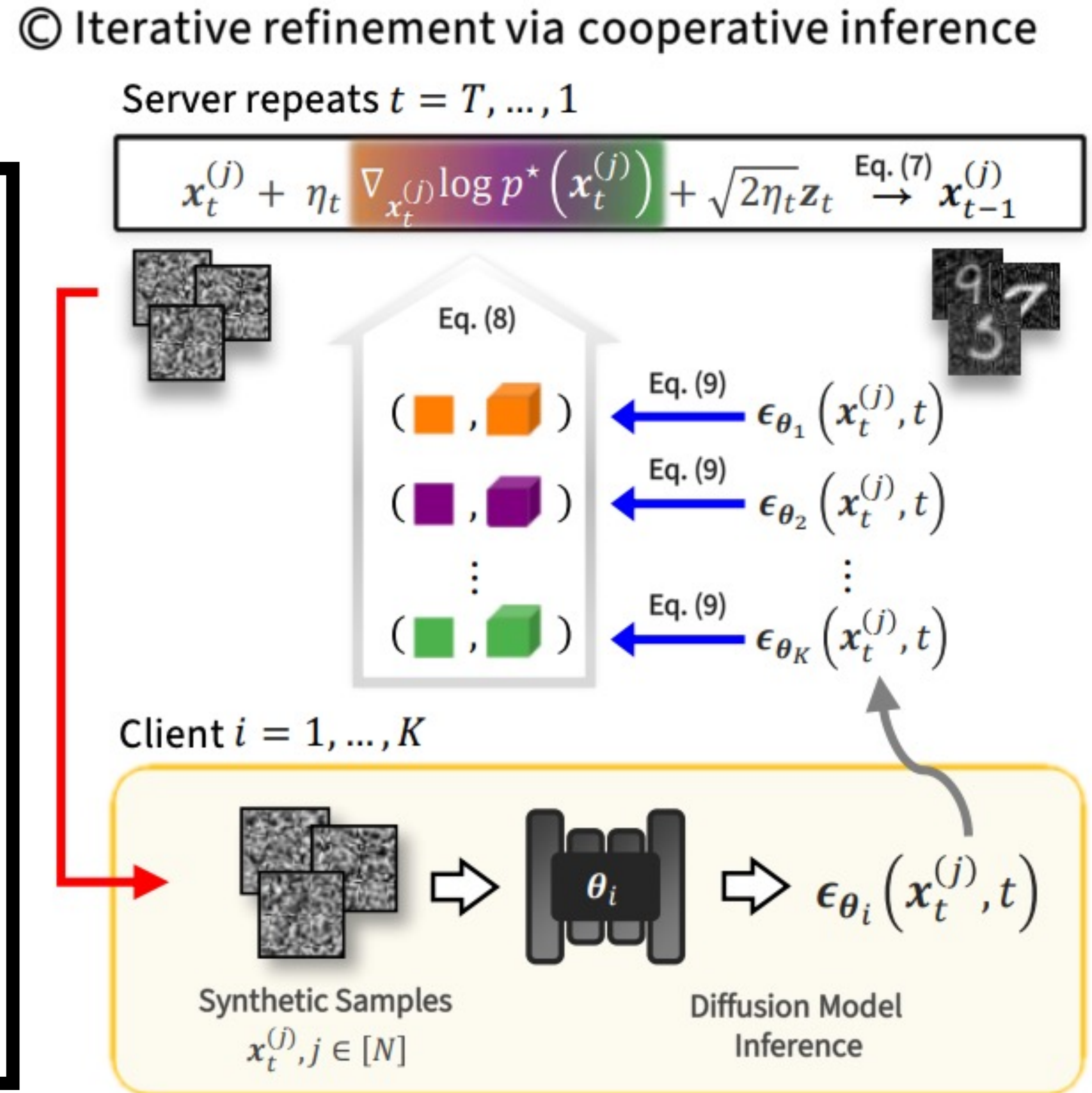
Breakdown

- C) Cooperative sampling of the synthetic dataset
  - Each local diffusion model represents a local distribution



Breakdown

- Eq. (7) (Black box itself)
    - Unadjusted Langevin Algorithm (ULA[3]) (first-order MCMC sampler)



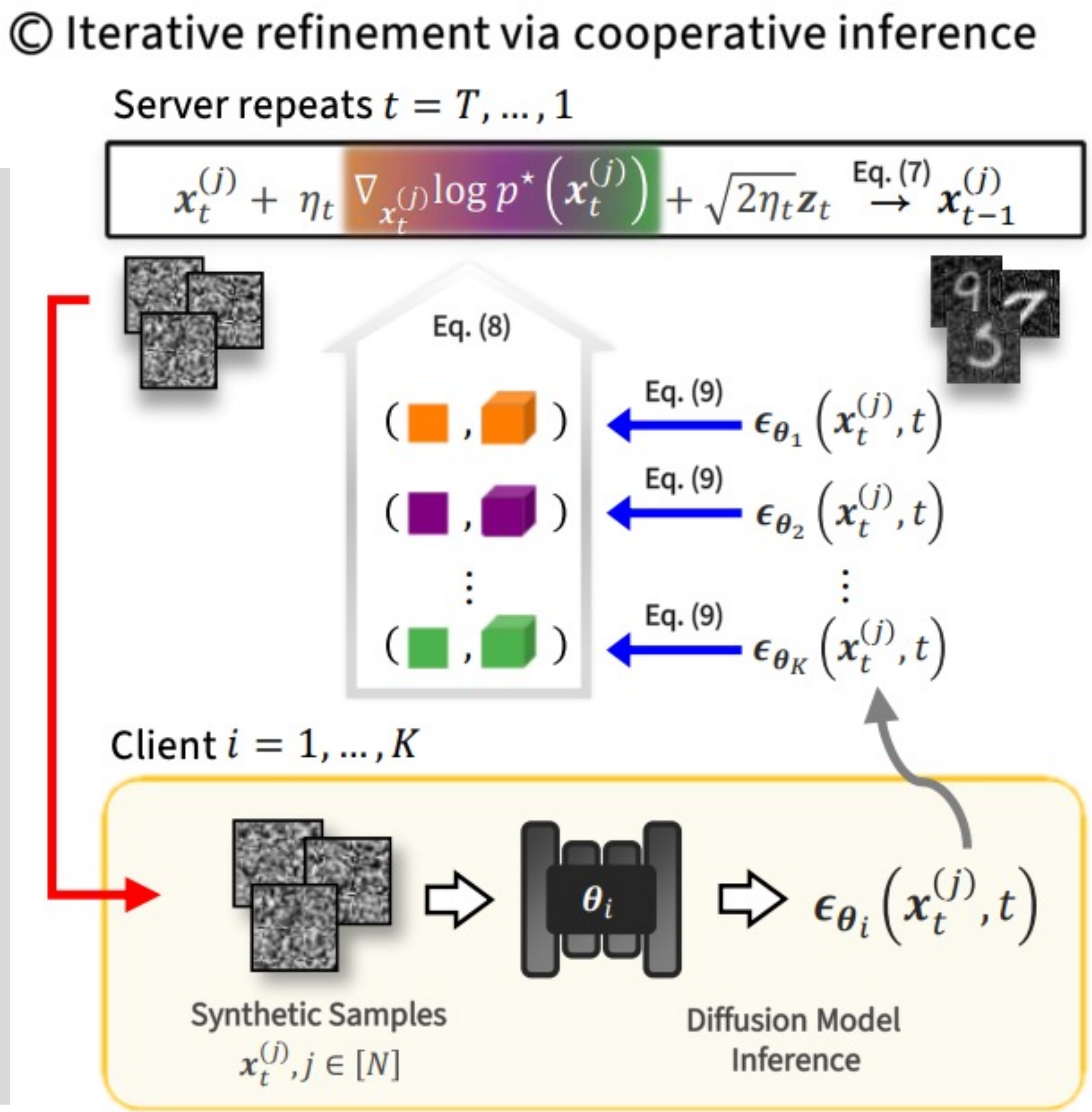
[3] Exponential Convergence of Langevin Distributions and Their Discrete Approximations (Roberts and Tweedie, 1996)

Breakdown

- Eq. (8) (grad. of log-prob. w.r.t. $\boldsymbol{x}$)

$$\nabla_{\boldsymbol{x}} \log p^{\star}(\boldsymbol{x}) = \sum_{i=1}^{K} \tilde{w}_i \nabla_{\boldsymbol{x}} \log p_{\boldsymbol{\theta}_i}(\boldsymbol{x})$$

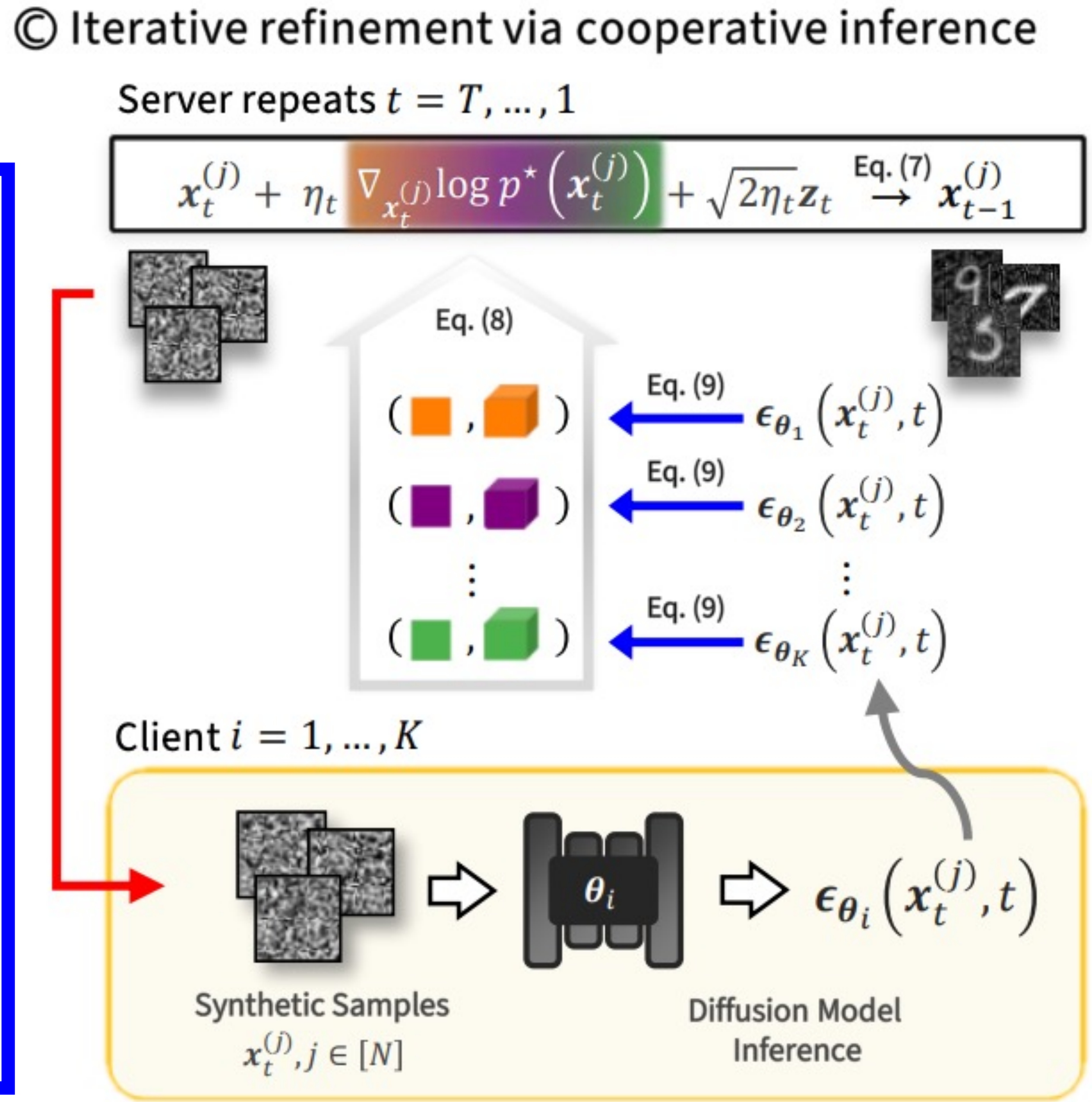
$$\tilde{w}_i = \frac{w_i \exp(-\lambda f_{\boldsymbol{\theta}_i}(\boldsymbol{x}))}{\sum_{j=1}^{K} w_j \exp(-\lambda f_{\boldsymbol{\theta}_j}(\boldsymbol{x}))}$$



Breakdown

- Eq. (9) (output to **score** conversion)

$$\nabla_{\boldsymbol{x}_t} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) \approx -\frac{\lambda}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$$



Wrap-up

- Data-centric federated synthetic data generation framework
- … as a cooperative inference of pre-trained diffusion models
- Thanks to their connection to energy-based models (EBMs)
- Non-asymptotic convergence guarantee
- Easy plug-and-play of differential privacy guarantee
- Lighter in communication compared to parameter averaging