FedCP: Separating Feature Information for Personalized Federated Learning via Conditional Policy

Jianqing Zhang[1] Yang Hua[2] Hao Wang[3]

Tao Song[1] Zhengui Xue[1] Ruhui Ma[1] Haibing Guan[1]

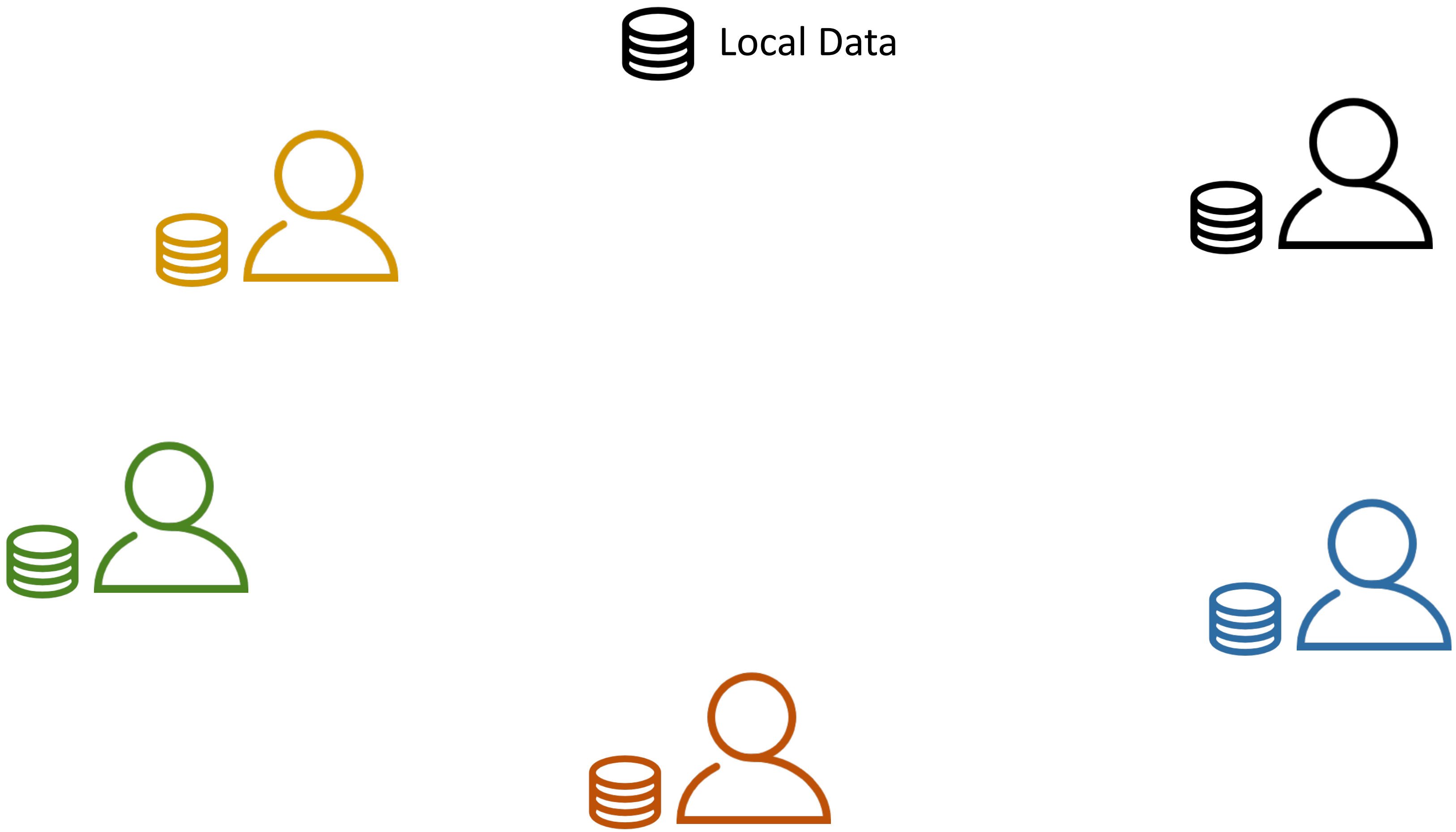
1 上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

2 QUEEN'S UNIVERSITY BELFAST

3 LSU

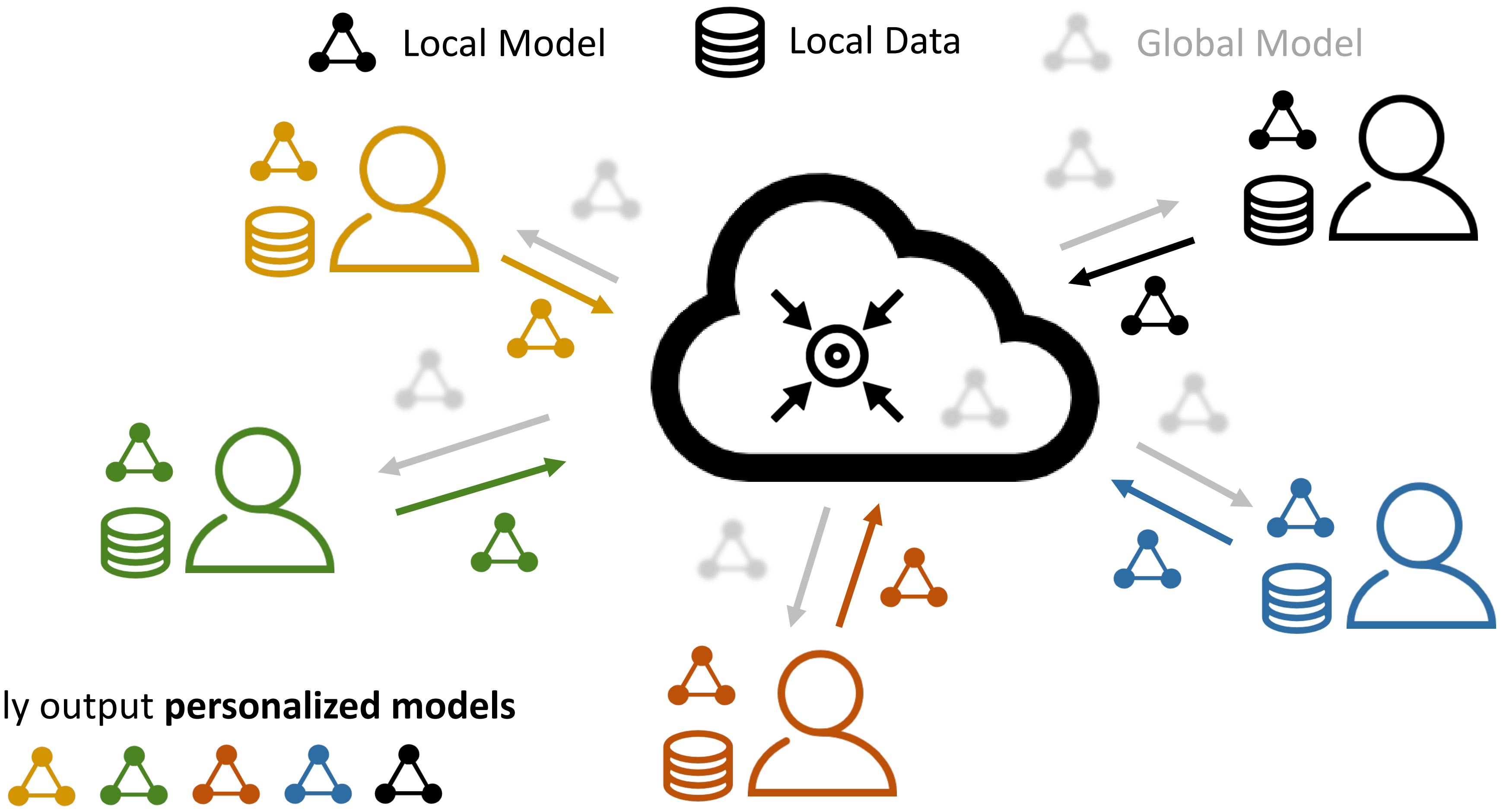
Existing Personalized Federated Learning (pFL)

- In practice, clients generate their specific private data, as shown by the colorful icons here.



Existing Personalized Federated Learning (pFL)

- **Goal:** address the *statistical heterogeneity* issue by learning personalized models.

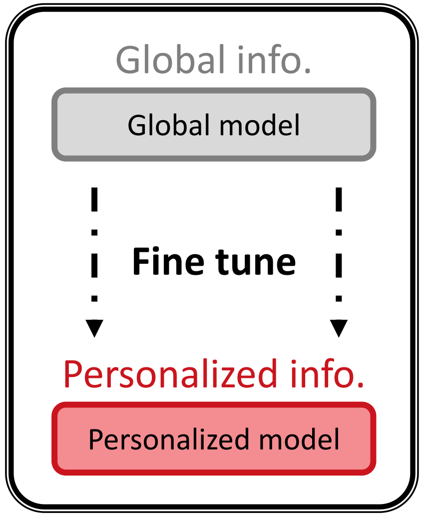


Existing Personalized Federated Learning (pFL)

- **Consensus:** reasonably utilizing global and personalized information is the key for pFL.

Existing Personalized Federated Learning (pFL)

- **Consensus:** reasonably utilizing global and personalized information is the key for pFL.
  - E.g., meta-learning-based (Per-FedAvg)

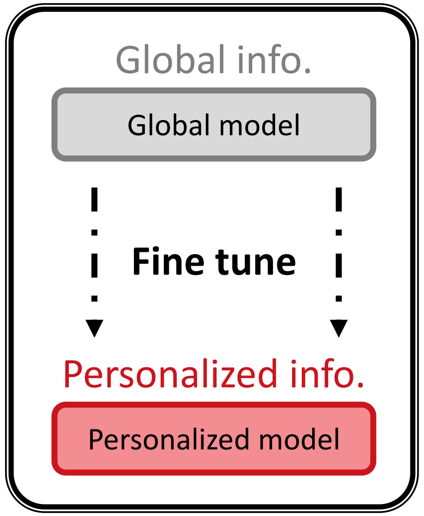


Per-FedAvg[1]

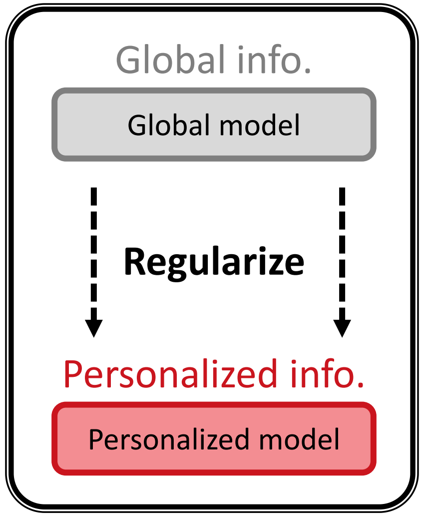
[1] Fallah A, Mokhtari A, Ozdaglar A. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. NeurIPS, 2020.

Existing Personalized Federated Learning (pFL)

- **Consensus:** reasonably utilizing global and personalized information is the key for pFL.
  - E.g., meta-learning-based (Per-FedAvg), regularization-based (Ditto)



Per-FedAvg[1]

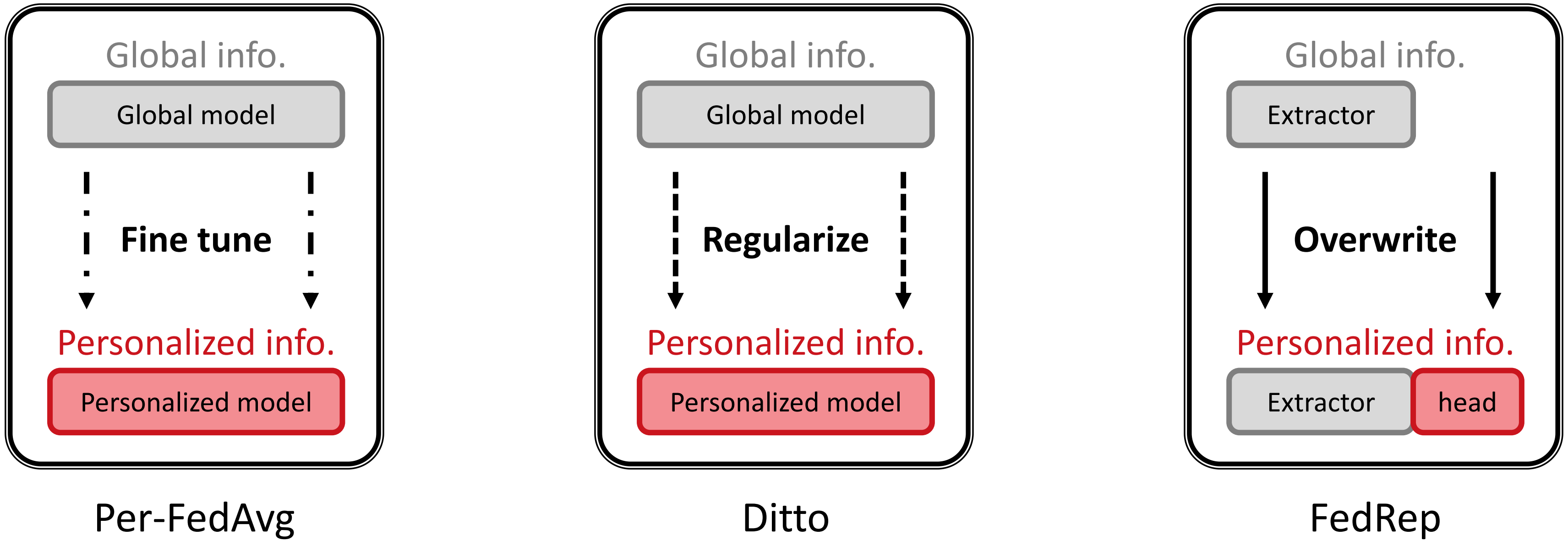


Ditto[2]

[1] Fallah A, Mokhtari A, Ozdaglar A. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. NeurIPS, 2020.
[2] Li T, Hu S, Beirami A, et al. Ditto: Fair and robust federated learning through personalization. ICML, 2021.

Existing Personalized Federated Learning (pFL)

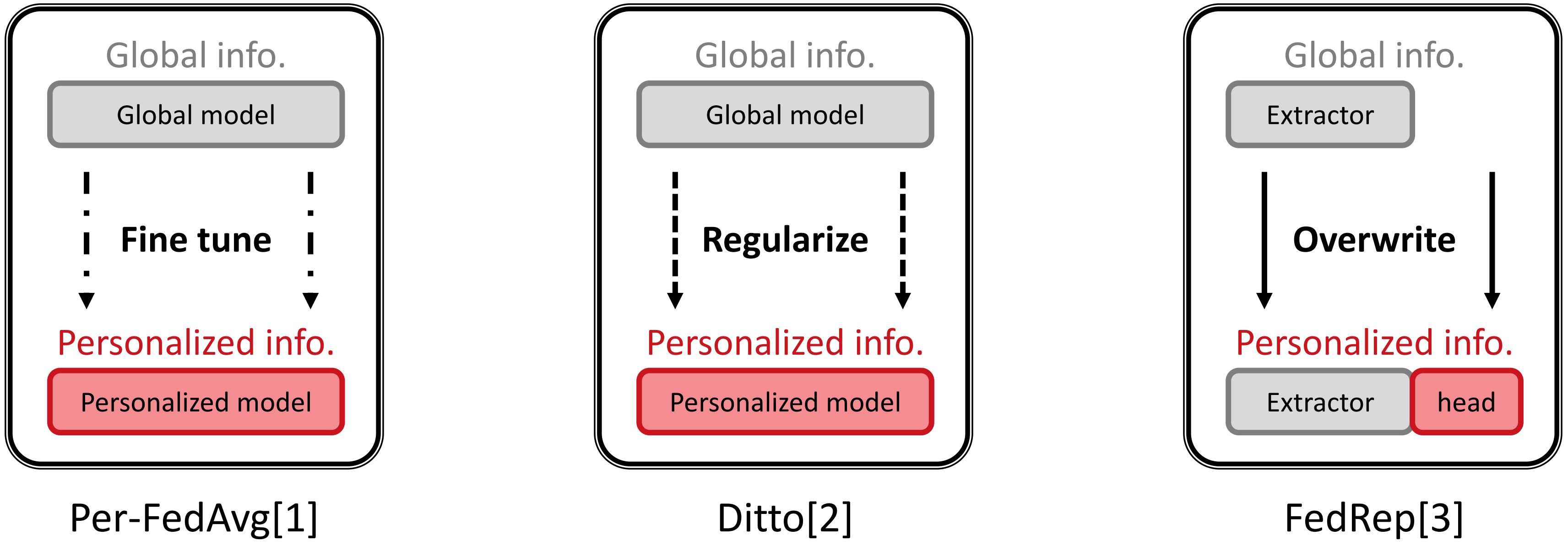
- **Consensus:** reasonably utilizing global and personalized information is the key for pFL.
  - E.g., meta-learning-based (Per-FedAvg), regularization-based (Ditto), and personalized-head-based (FedRep) pFL.



[1] Fallah A, Mokhtari A, Ozdaglar A. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. NeurIPS, 2020.
[2] Li T, Hu S, Beirami A, et al. Ditto: Fair and robust federated learning through personalization. ICML, 2021.
[3] Collins L, Hassani H, Mokhtari A, et al. utilizing shared representations for personalized federated learning. ICML, 2021.

Existing Personalized Federated Learning (pFL)

- **Consensus:** reasonably utilizing global and personalized information is the key for pFL.
  - E.g., meta-learning-based (Per-FedAvg), regularization-based (Ditto), and personalized-head-based (FedRep) pFL.



- They only focus on model parameters, but ignore ***the source of information: data.***

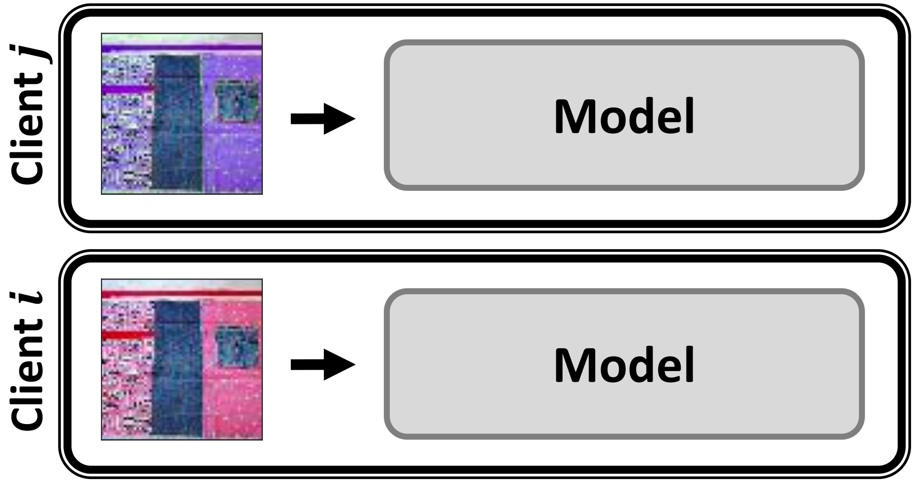
[1] Fallah A, Mokhtari A, Ozdaglar A. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. NeurIPS, 2020.
[2] Li T, Hu S, Beirami A, et al. Ditto: Fair and robust federated learning through personalization. ICML, 2021.
[3] Collins L, Hassani H, Mokhtari A, et al. utilizing shared representations for personalized federated learning. ICML, 2021.

Global and Personalized Information in Heterogeneous Data

- The ***heterogeneous data*** on clients contains both global and personalized information

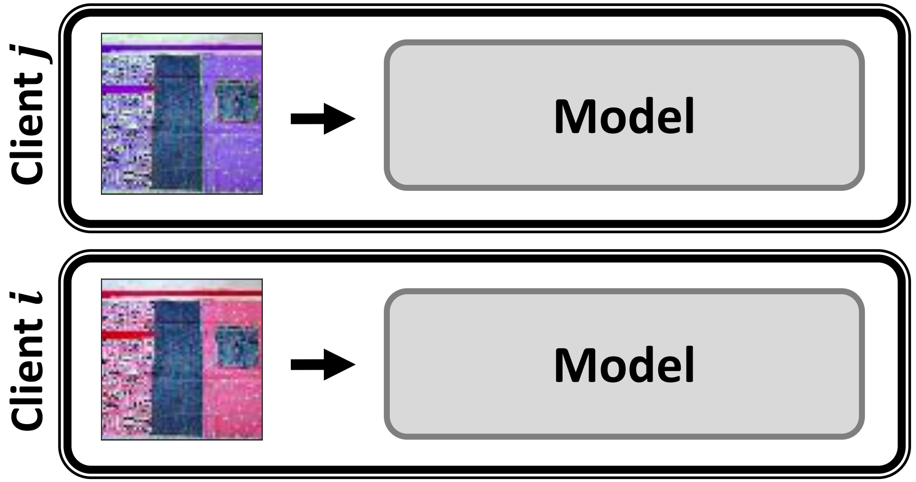
Global and Personalized Information in Heterogeneous Data

- The ***heterogeneous data*** on clients contains both global and personalized information
  - E.g., blue (widely-used) contains global information and purple/pink (rarely-used) contains personalized information.



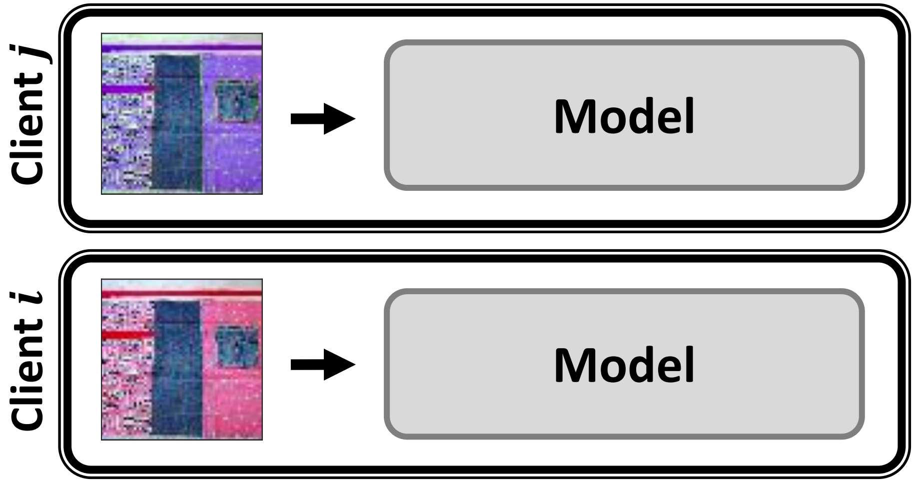
Global and Personalized Information in Heterogeneous Data

- The ***heterogeneous data*** on clients contains both global and personalized information
  - E.g., blue (widely-used) contains global information and purple/pink (rarely-used) contains personalized information.
  - Model learns this global and personalized information to its parameters, but



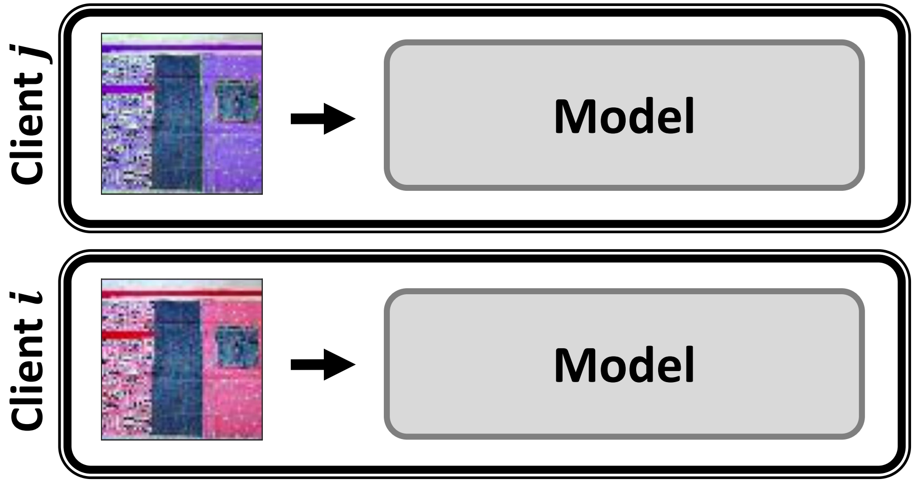
Global and Personalized Information in Heterogeneous Data

- The ***heterogeneous data*** on clients contains both global and personalized information
  - E.g., blue (widely-used) contains global information and purple/pink (rarely-used) contains personalized information.
  - Model learns this global and personalized information to its parameters, but
    - Parameters are second-hand information and only data is the first-hand information.



Global and Personalized Information in Heterogeneous Data

- The ***heterogeneous data*** on clients contains both global and personalized information
  - E.g., blue (widely-used) contains global information and purple/pink (rarely-used) contains personalized information.
  - Model learns this global and personalized information to its parameters, but
    - Parameters are second-hand information and only data is the first-hand information.
  - How to directly utilizing global and personalized information in data?



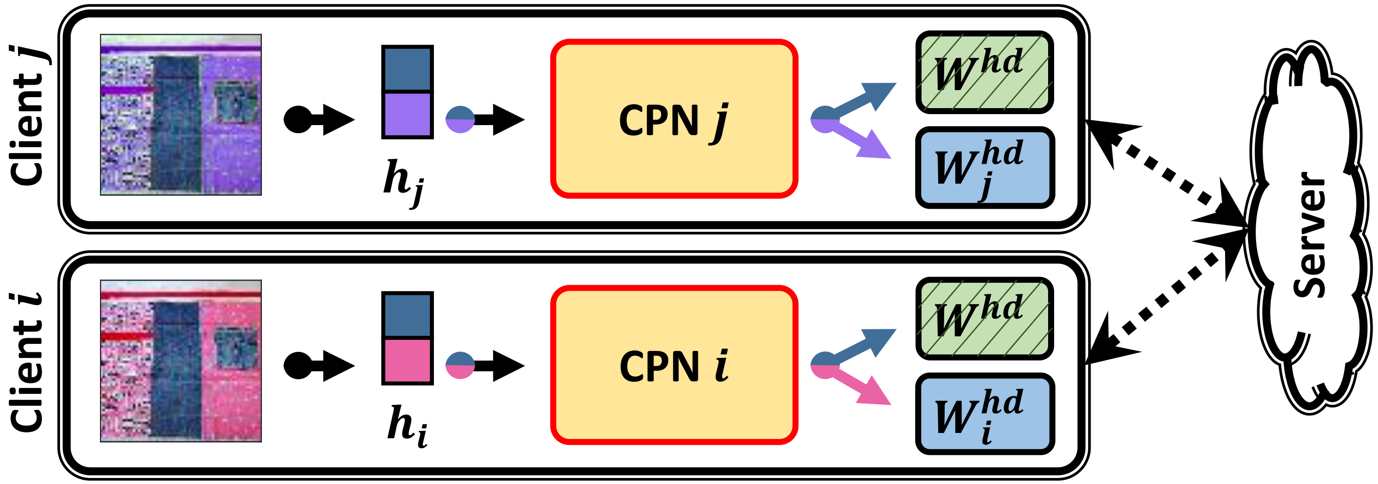
Global and Personalized Information in Heterogeneous Data

- The ***heterogeneous data*** on clients contains both global and personalized information
  - E.g., blue (widely-used) contains global information and purple/pink (rarely-used) contains personalized information.
  - Model learns this global and personalized information to its parameters, but
    - Parameters are second-hand information and only data is the first-hand information.
  - How to directly utilizing global and personalized information in data?

- Since the dimension of raw data is too large, we consider the extracted feature vector.

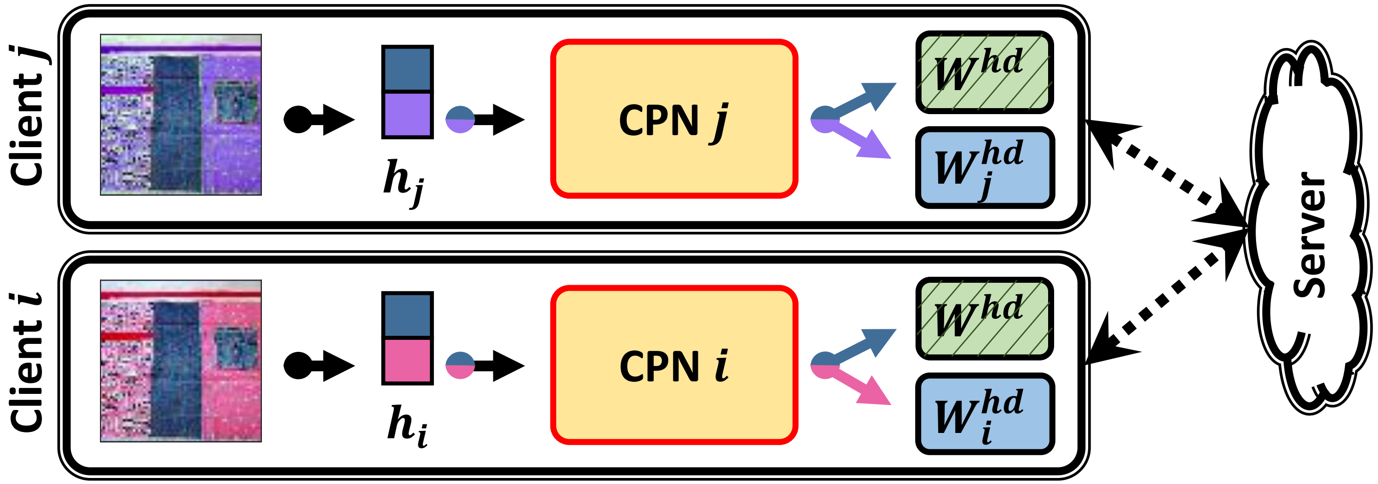
Separating Feature Information

- We propose to separate feature information via an *auxiliary* **Conditional Policy Network (CPN).**



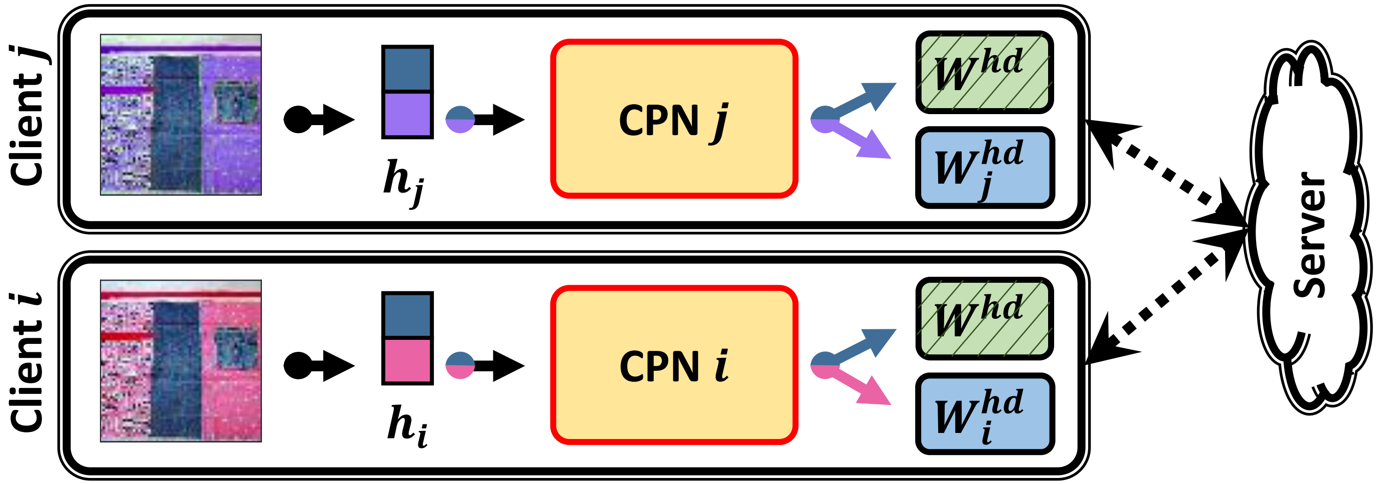
Separating Feature Information

- We propose to separate feature information via an *auxiliary* **Conditional Policy Network (CPN).**
  - Generate sample-specific policy
  - End-to-end training together with the client model
  - Lightweight (e.g., 4.67% parameters of ResNet-18)



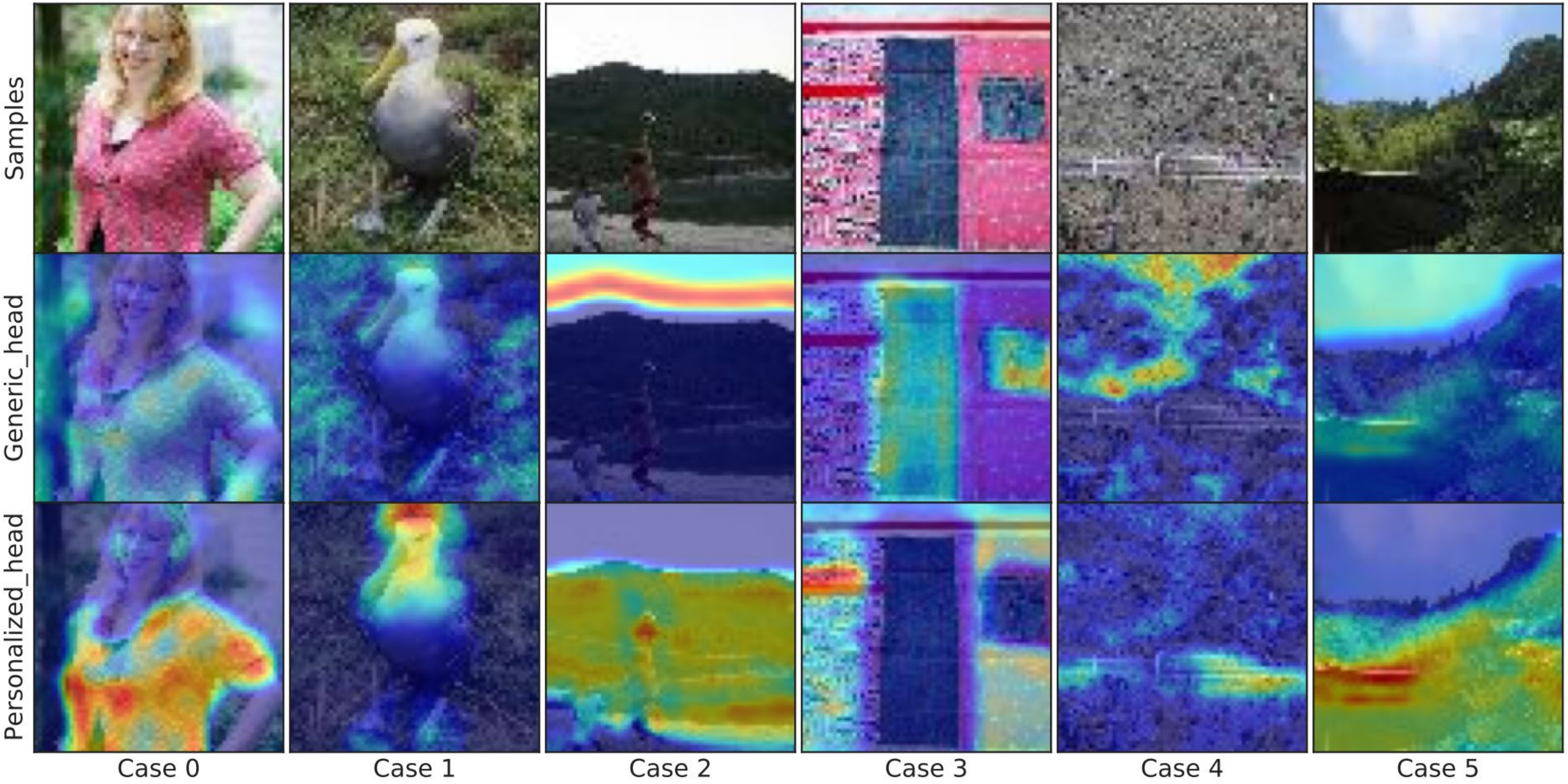
Separating Feature Information

- We propose to separate feature information via an *auxiliary* **Conditional Policy Network (CPN).**
  - Generate sample-specific policy
  - End-to-end training together with the client model
  - Lightweight (e.g., 4.67% parameters of ResNet-18)



- We utilize global and personalized information via global and personalized heads, respectively.

Separating Feature Information



Separating Feature Information

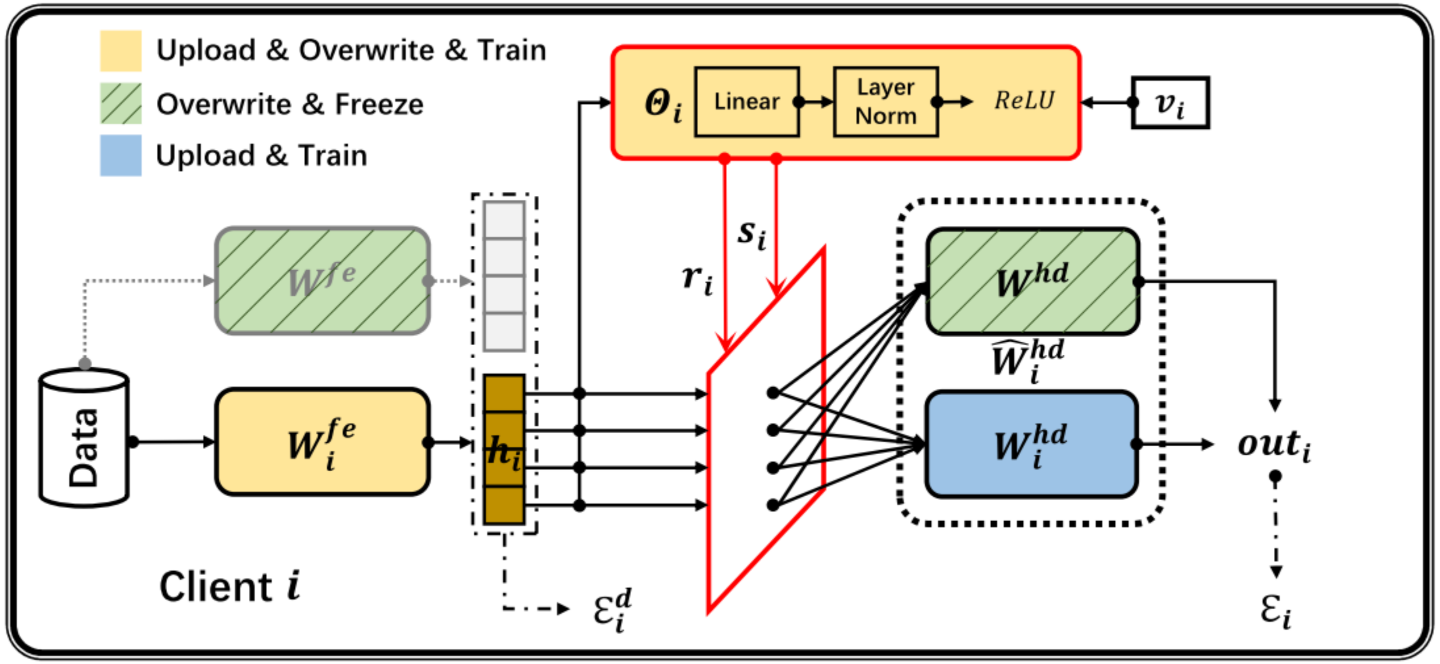
- How to realize?

Separating Feature Information

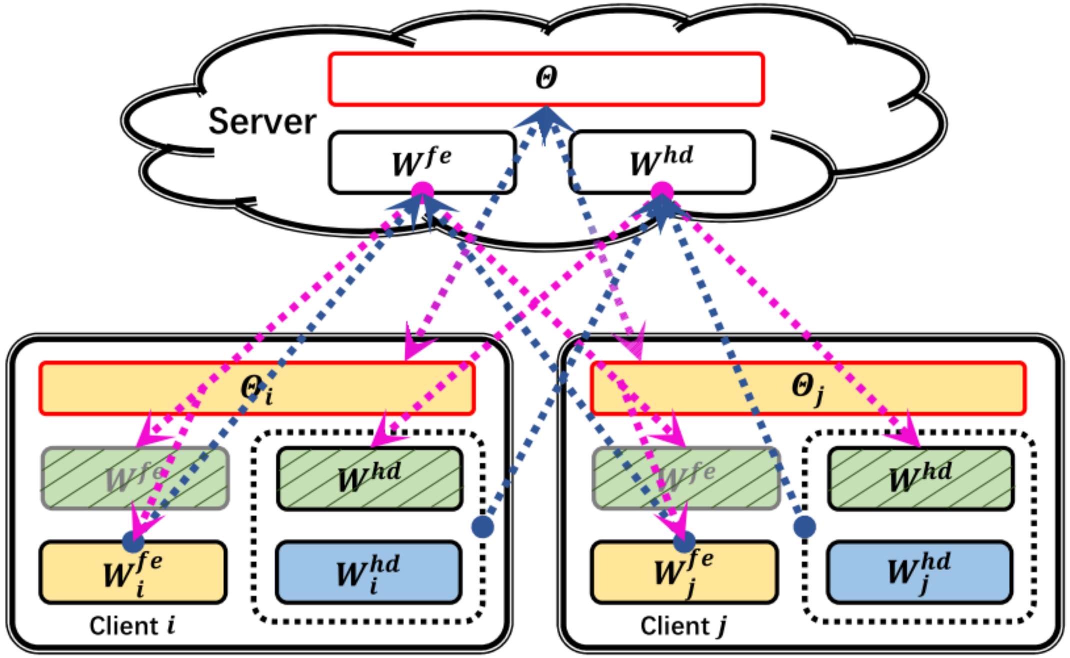
- How to realize?
- Use our **Federated Conditional Policy (FedCP)** framework.

Separating Feature Information

- How to realize?
- Use our **Federated Conditional Policy (FedCP)** framework.



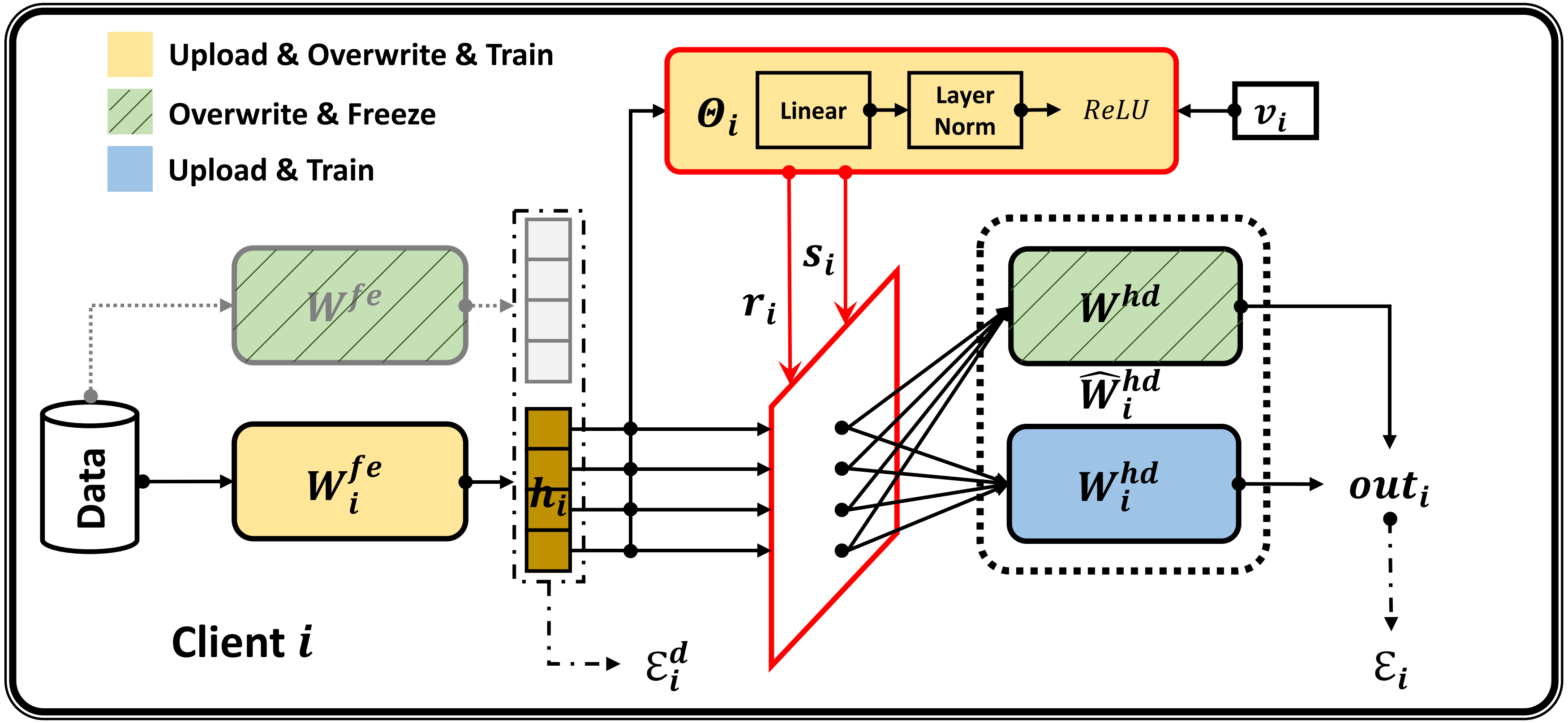
(a) Forward data flow corresponding to the local learning on client i.



(b) Upload and download streams in FedCP.

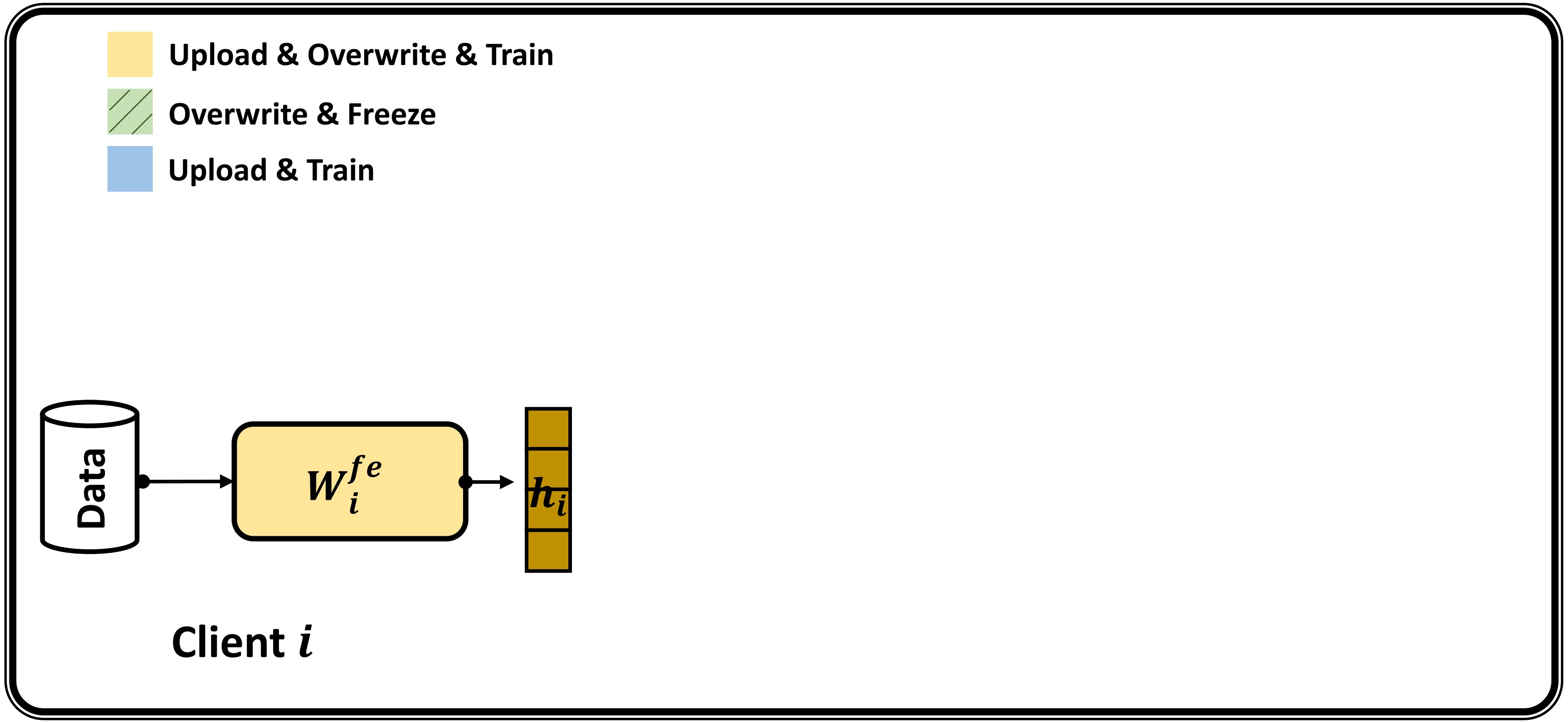
Separating Feature Information

- Key operations are done on the client side



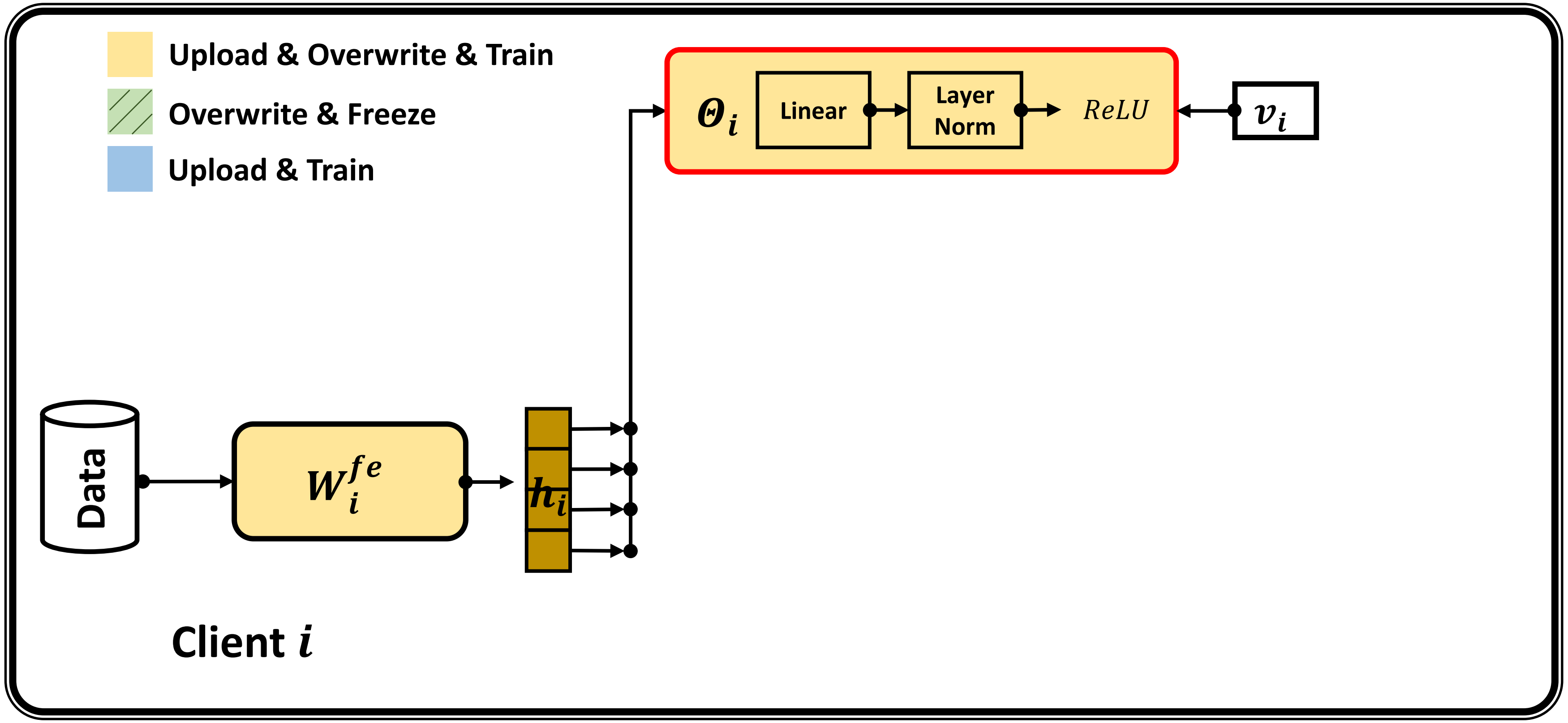
Separating Feature Information

- Obtain feature vector $\boldsymbol{h}_i$



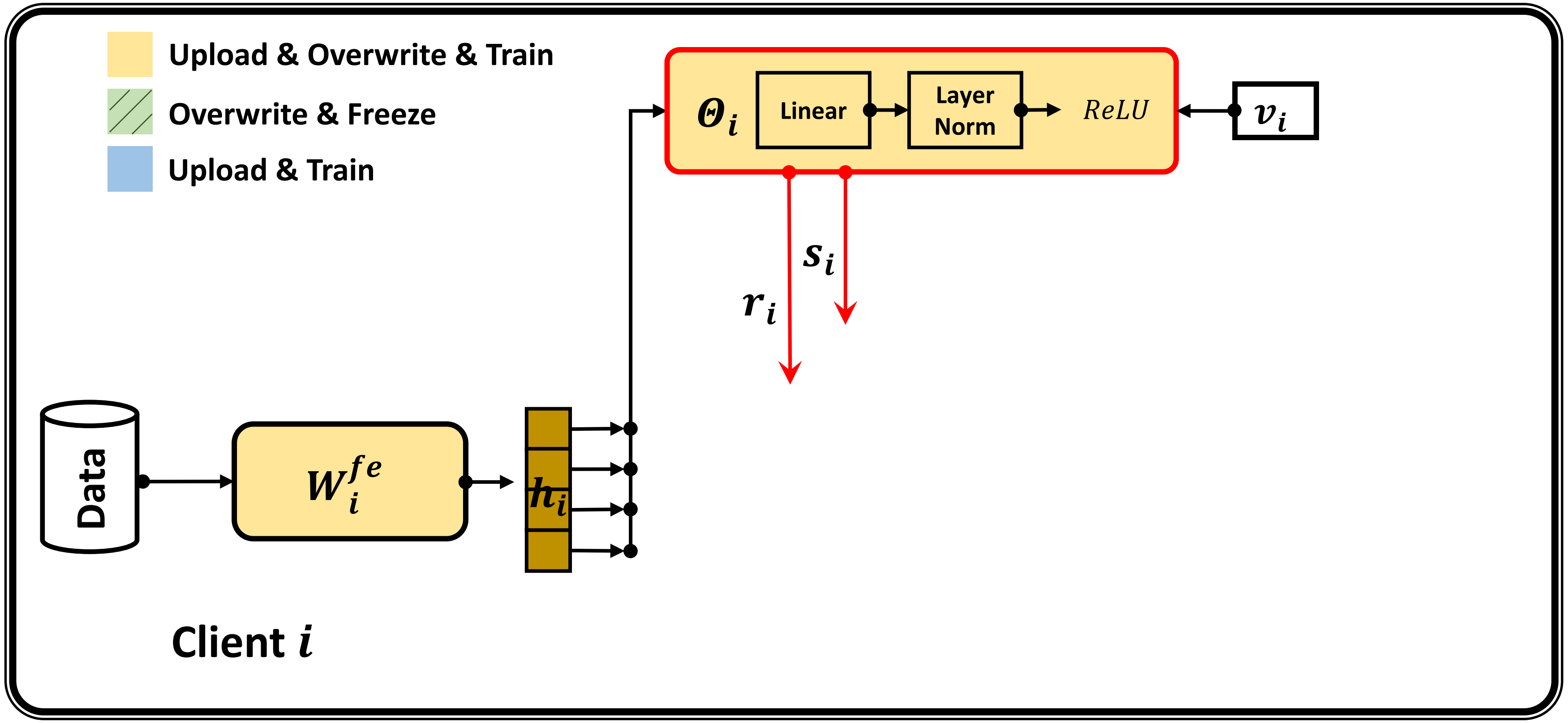
Separating Feature Information

- Consider sample-specific $\boldsymbol{h}_i$ and client-specific $\boldsymbol{v}_i$ as the conditional input $\mathcal{C}_i$ for CPN



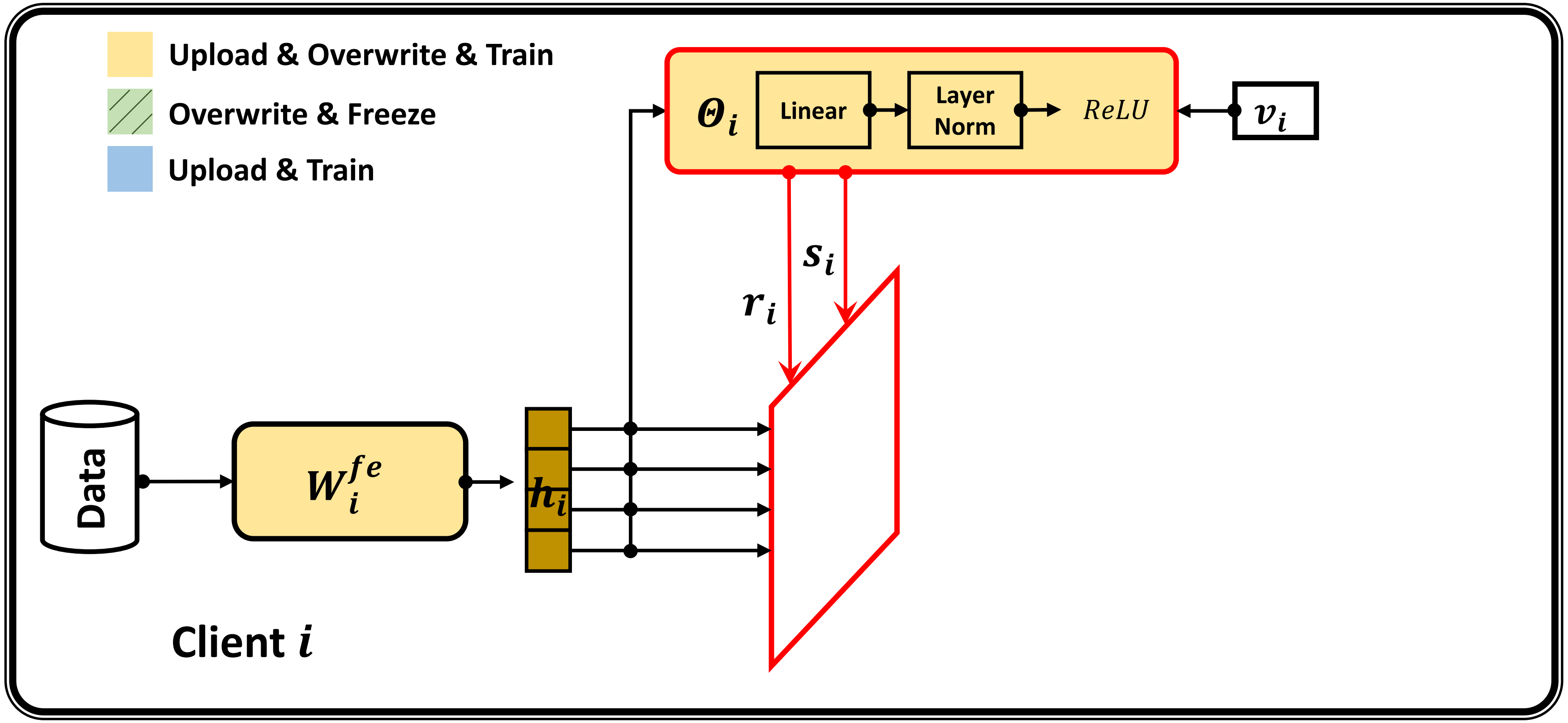
Separating Feature Information

- Generate conditional policy $\{\boldsymbol{r}_i, \boldsymbol{s}_i\}$ via CPN



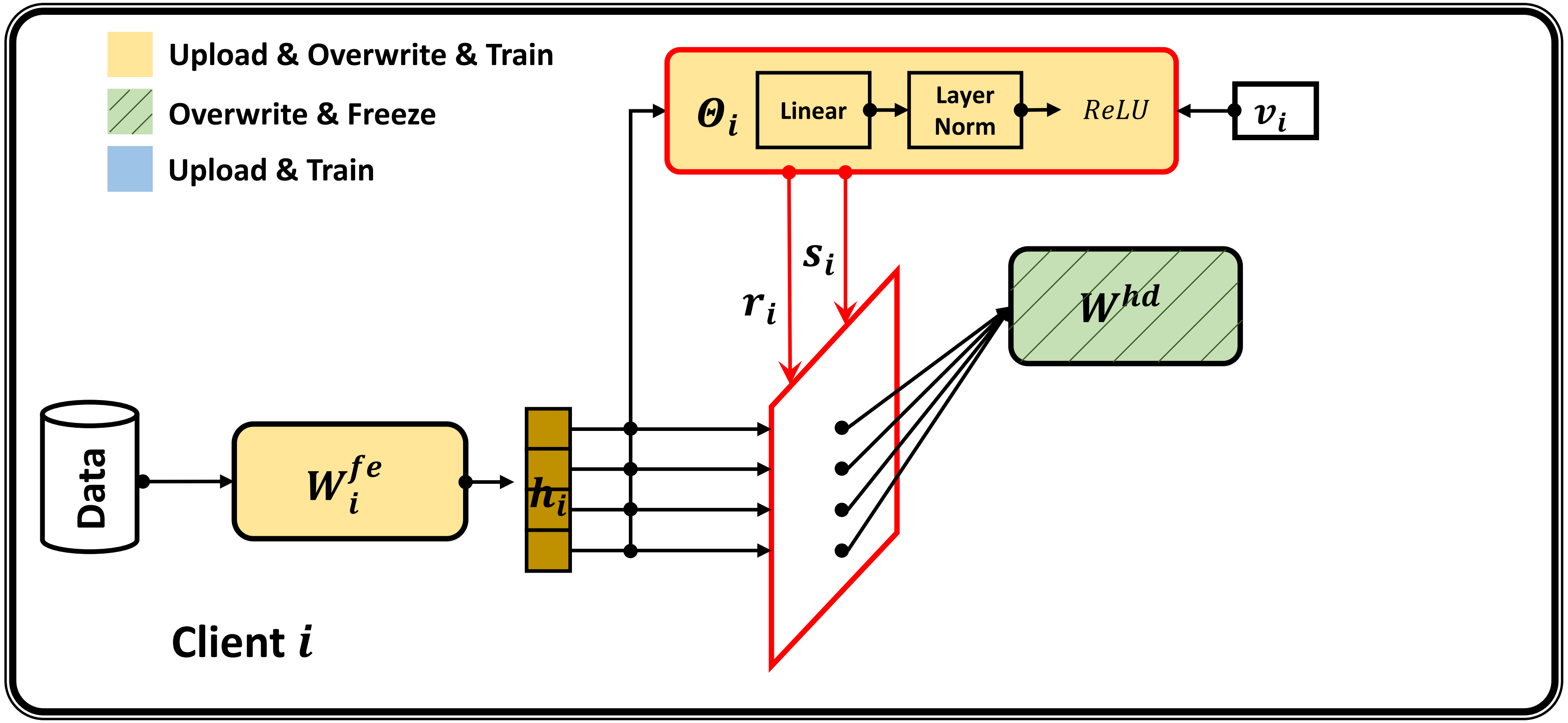
Separating Feature Information

- Multiply conditional policy $\{\boldsymbol{r}_i, \boldsymbol{s}_i\}$ to $\boldsymbol{h}_i$ to obtain $\boldsymbol{r}_i \odot \boldsymbol{h}_i$ and $\boldsymbol{s}_i \odot \boldsymbol{h}_i$



Separating Feature Information

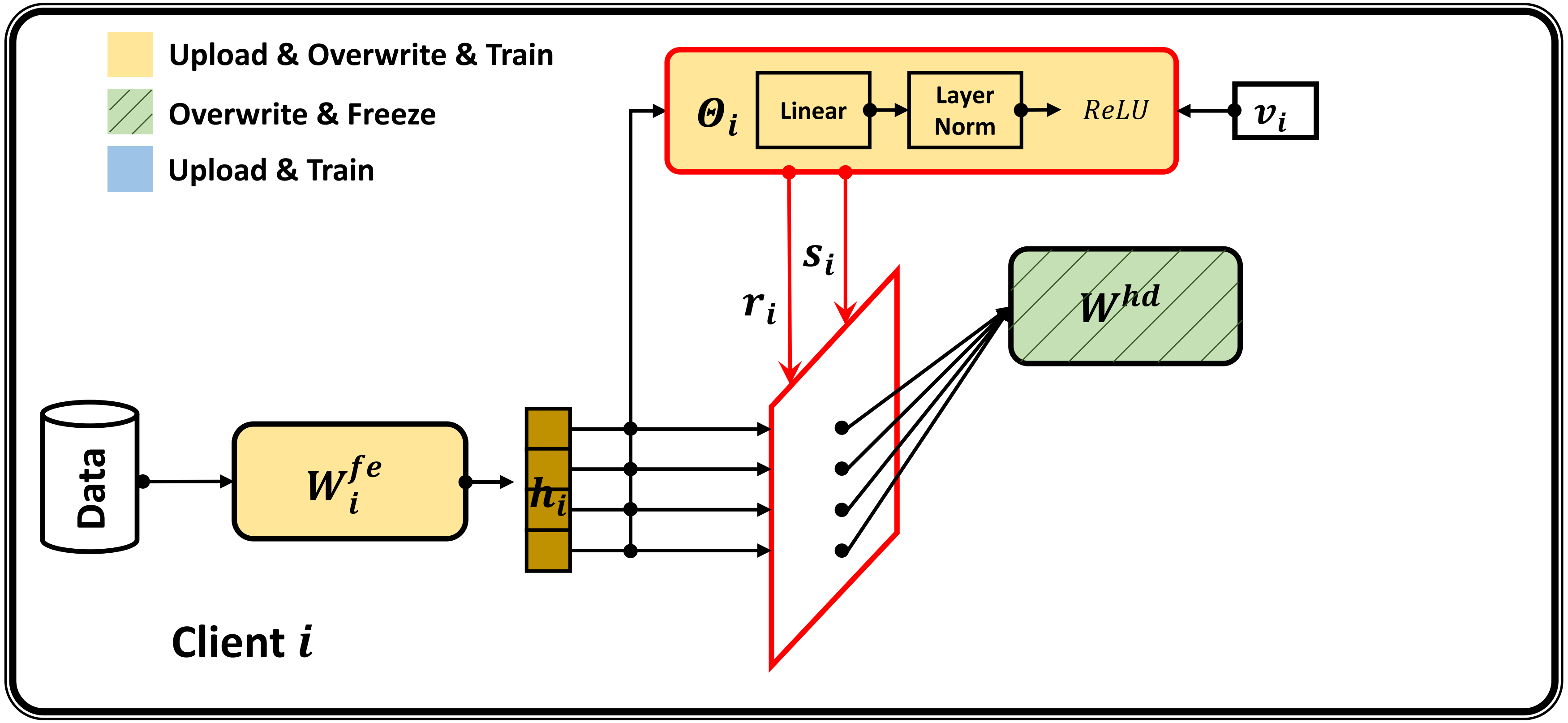
- Process global feature information $\boldsymbol{r}_i \odot \boldsymbol{h}_i$ via a ***frozen*** global head $\boldsymbol{W}^{hd}$



Separating Feature Information

- Process global feature information $\boldsymbol{r}_i \odot \boldsymbol{h}_i$ via a ***frozen*** global head $\boldsymbol{W}^{hd}$

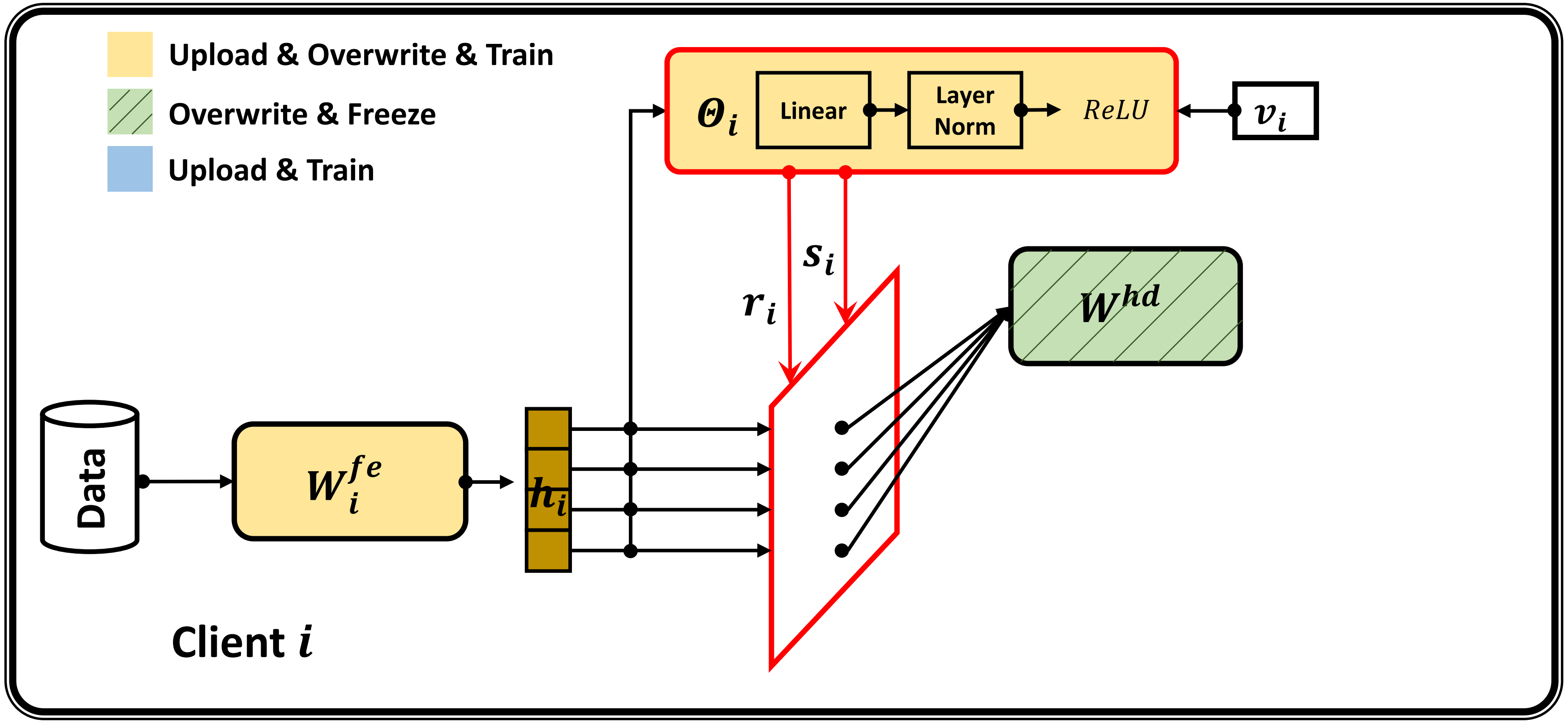
Why?



Separating Feature Information

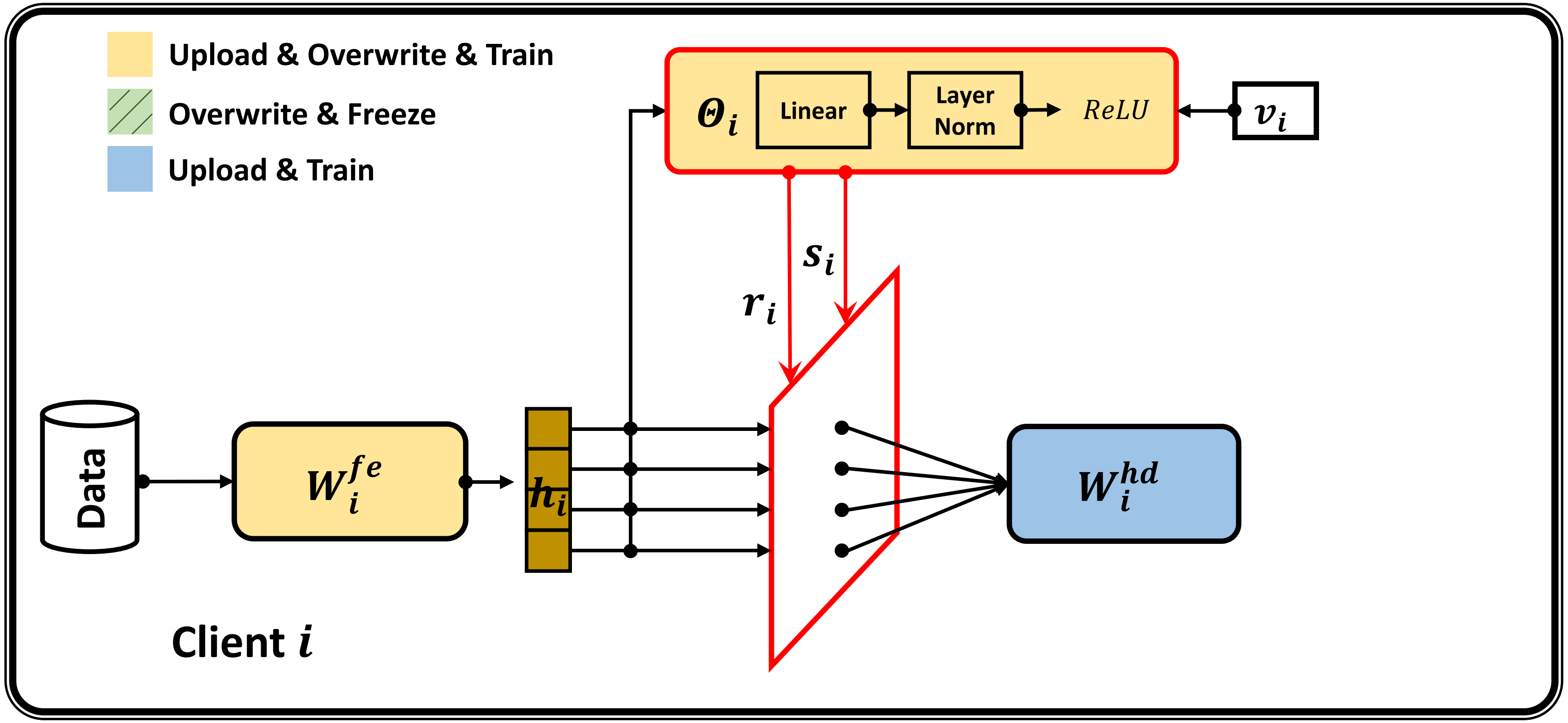
- Process global feature information $\boldsymbol{r}_i \odot \boldsymbol{h}_i$ via a ***frozen*** global head $\boldsymbol{W}^{hd}$

Retain global information to guide CPN training during *backward*



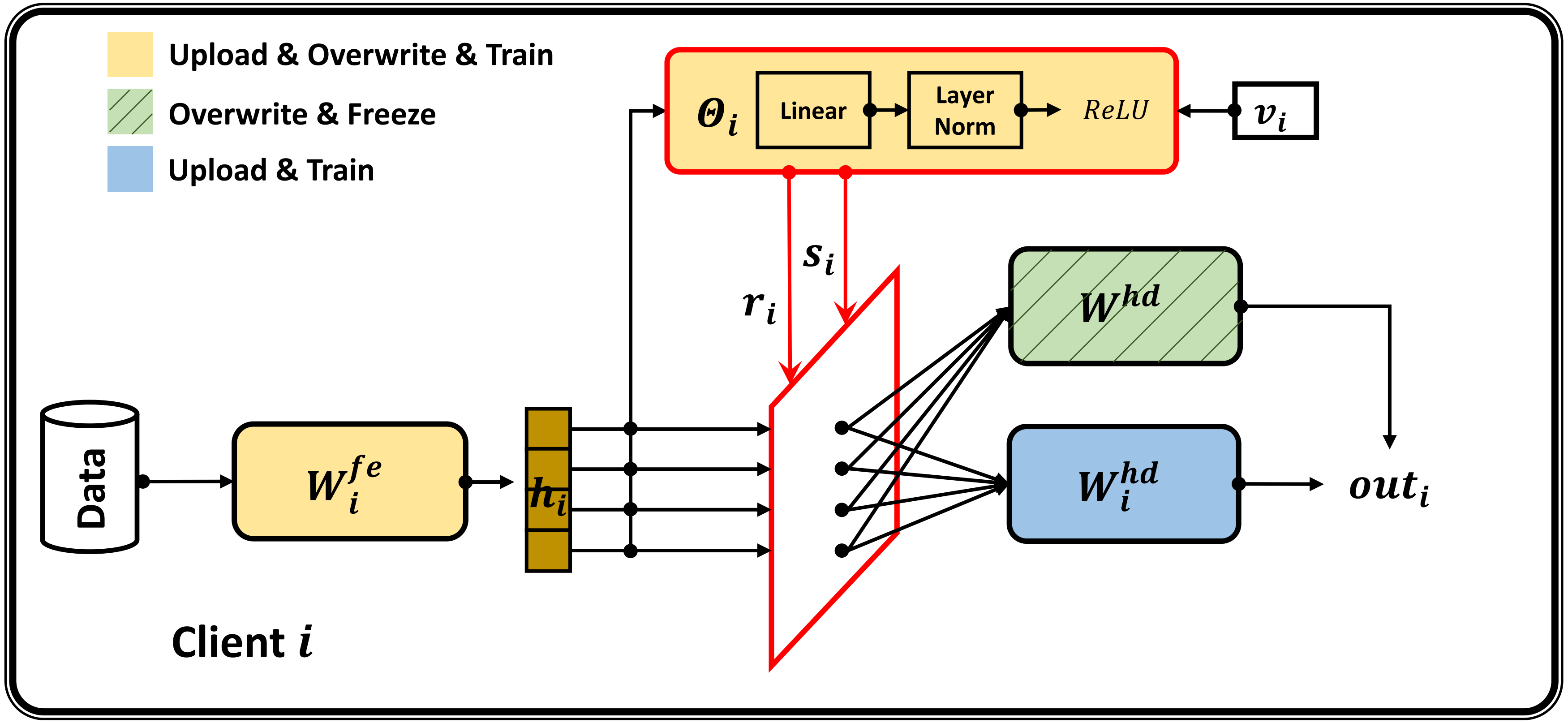
Separating Feature Information

- Process personalized feature information $s_i \odot h_i$ via a personalized head W_i^{hd}



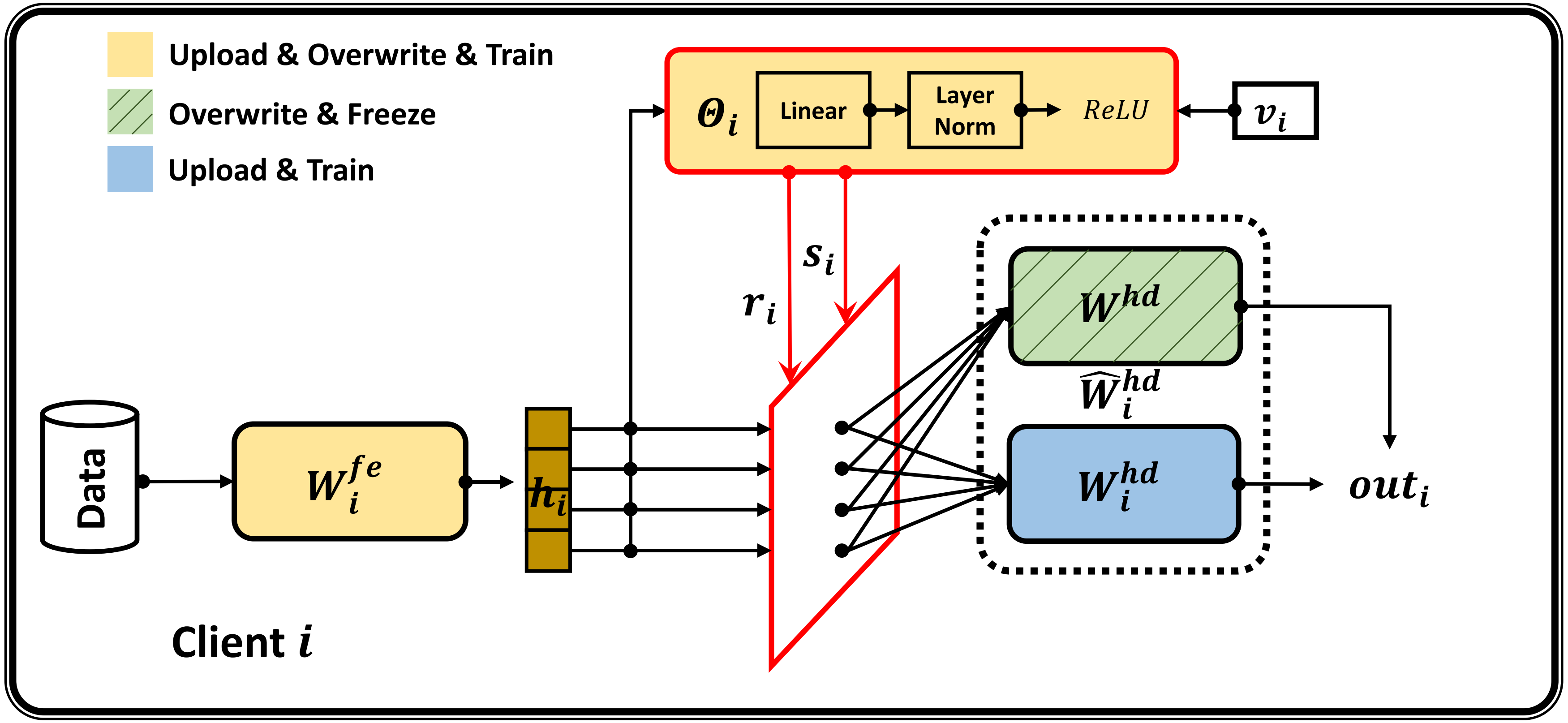
Separating Feature Information

- Combine the outputs of two heads to form final output $\boldsymbol{out_i}$



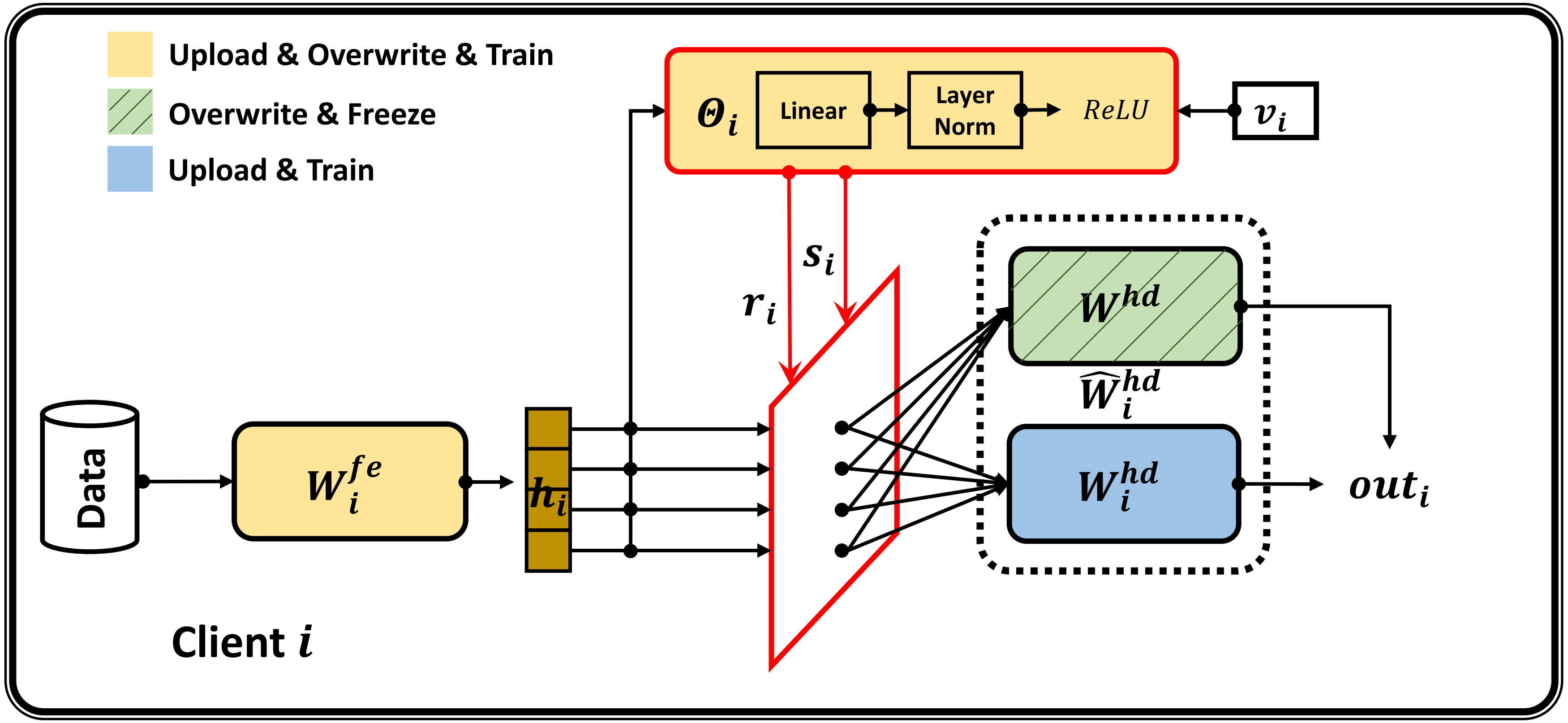
Separating Feature Information

- From the view of each sample, its features are processed by an unified head $\widehat{W}_i^{hd}$



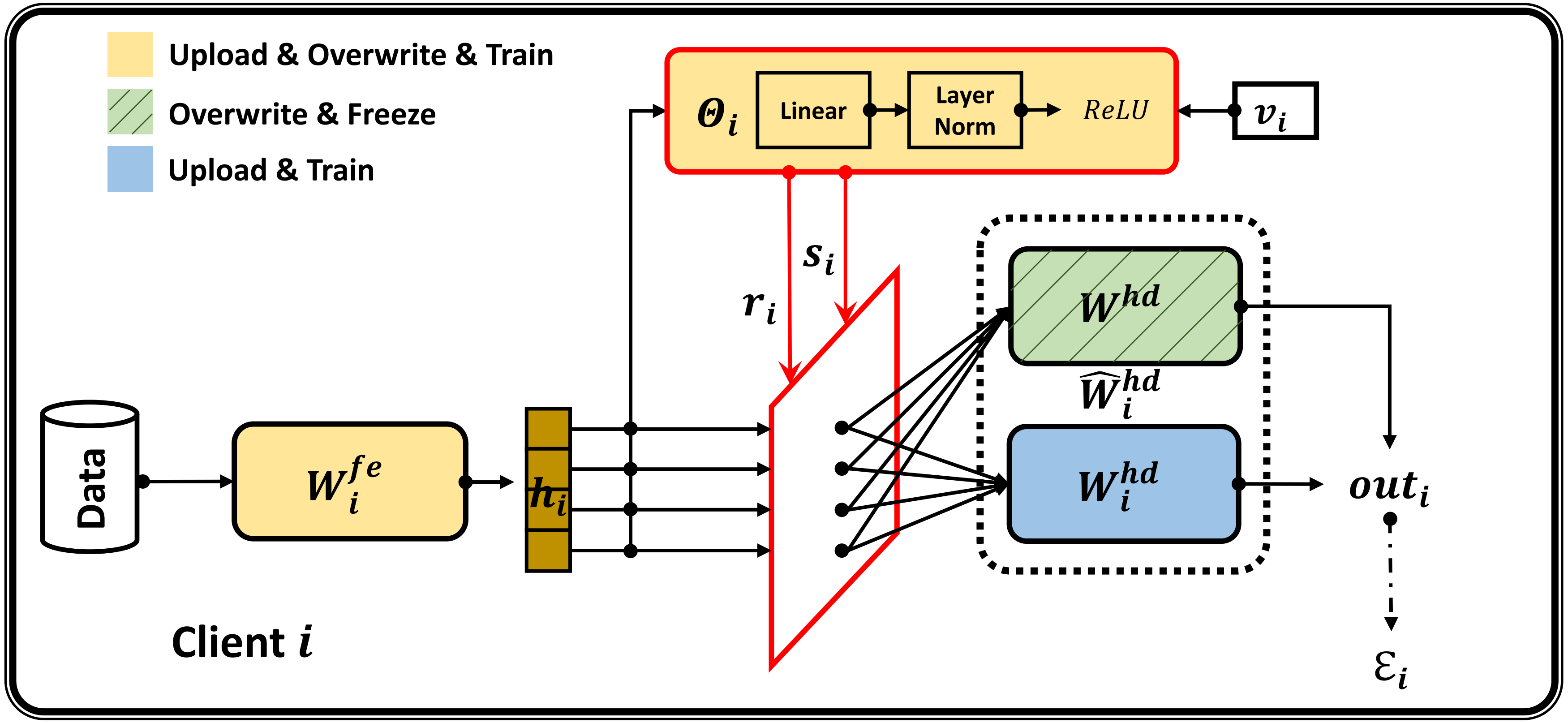
Separating Feature Information

- Personalized model for inference



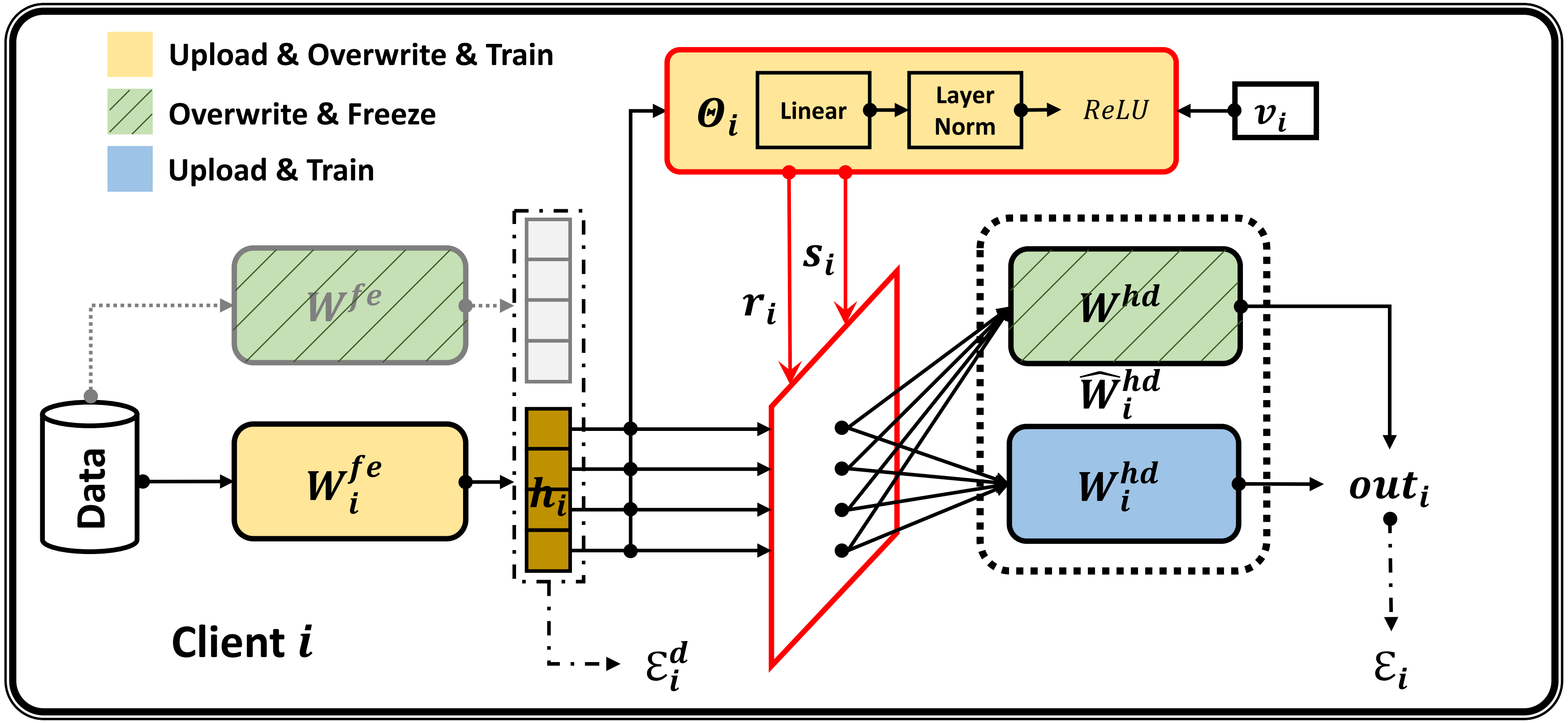
Separating Feature Information

- Personalized model for training: classification error, local cross entropy loss $\mathcal{E}_i$



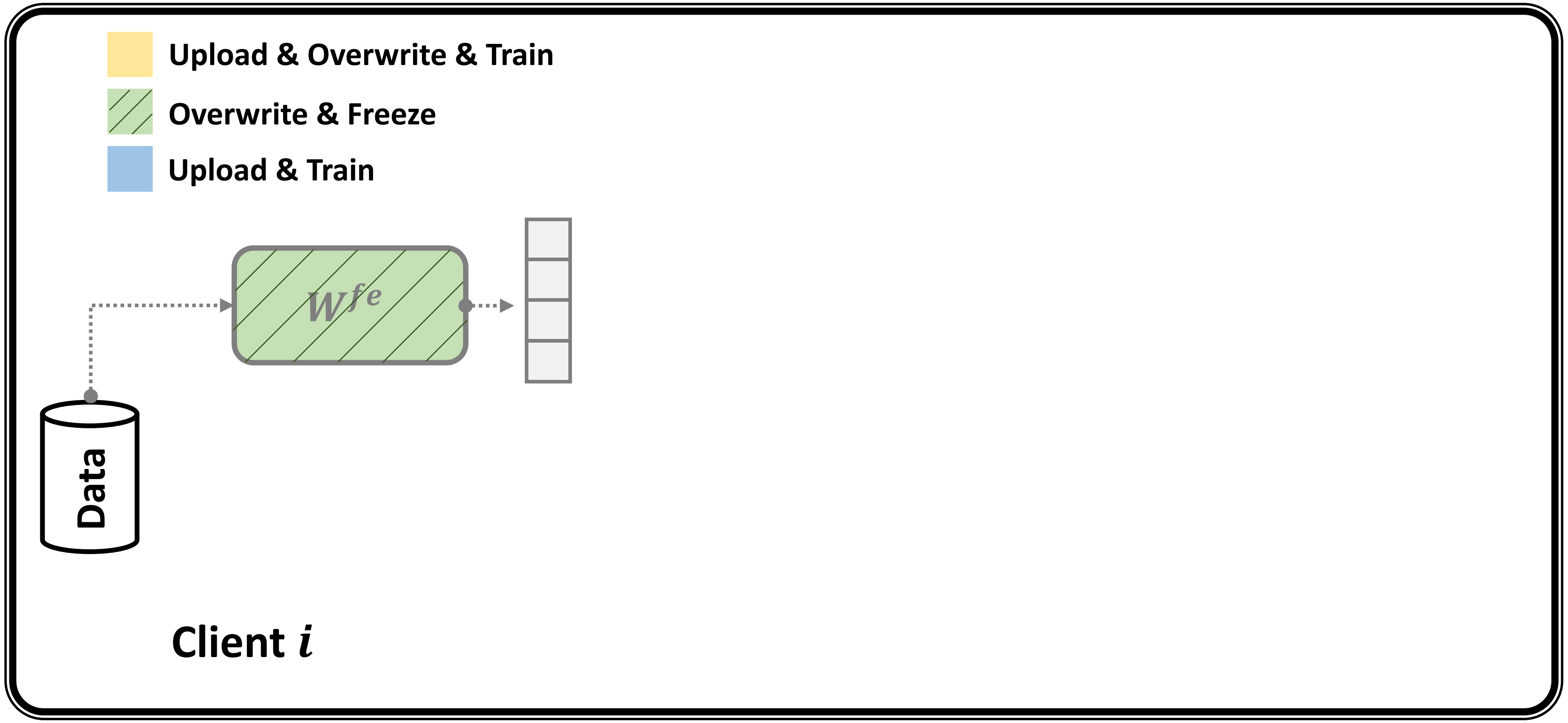
Separating Feature Information

- Personalized model for training: aligning features, MMD loss $\mathcal{E}_i^d$



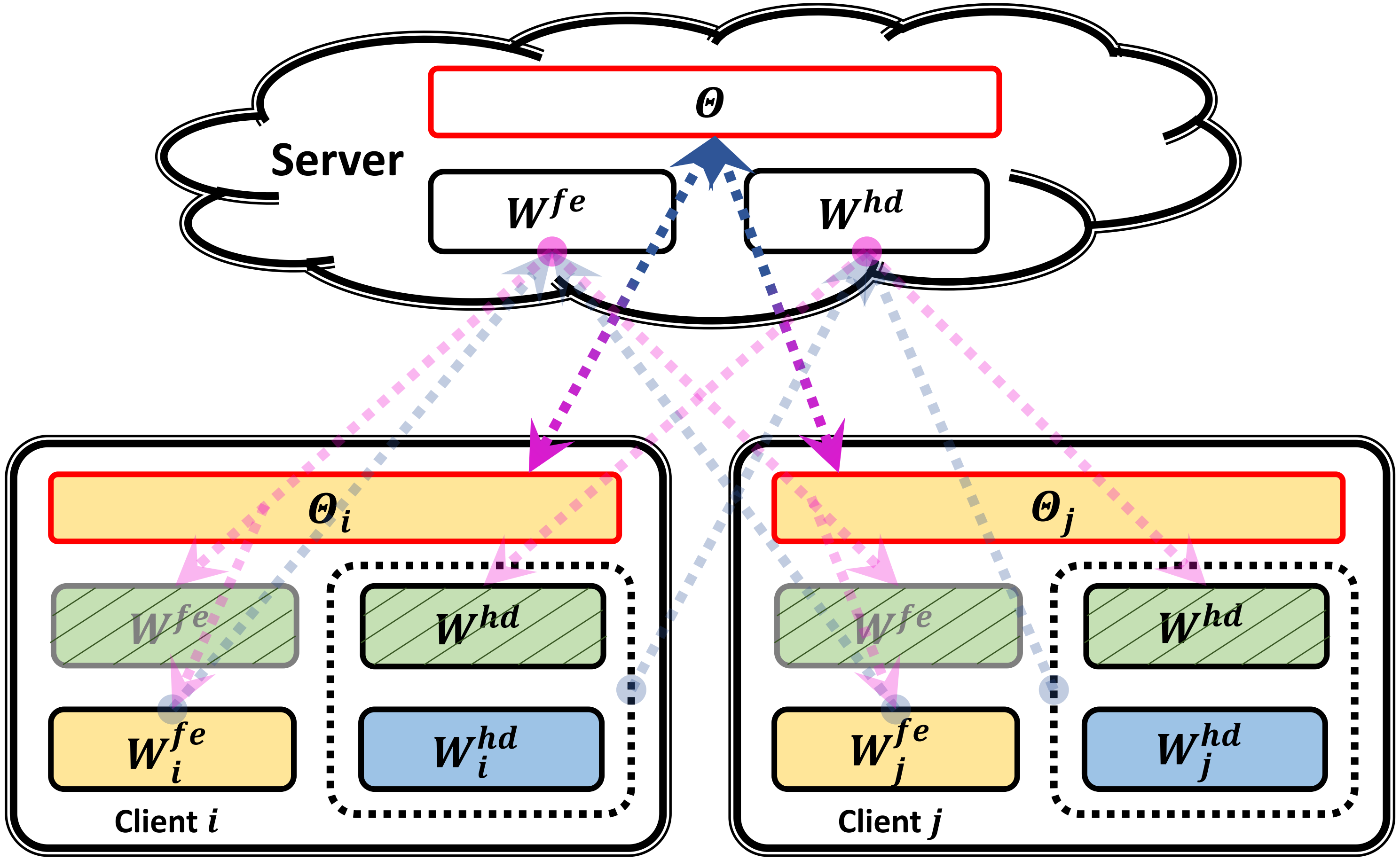
Separating Feature Information

- Personalized model for training: gray-colored components are **only** used for training



Separating Feature Information

- **Only $\boldsymbol{\Theta}$** introduces additional communication overhead per iteration (e.g., 4.67% for ResNet-18)



Extensive Experiments

- FedCP outperforms **11** SOTA traditional FL and pFL methods by up to **6.69%**

The accuracy (%) of the image/text classification tasks in the main experiments.

Settings	Pathological setting			Default practical setting (β = 0.1)					
	MNIST	Cifar10	Cifar100	MNIST	Cifar10	Cifar100	TINY	TINY*	AG News
FedAvg [32]	97.93±0.05	55.09±0.83	25.98±0.13	98.81±0.01	59.16±0.47	31.89±0.47	19.46±0.20	19.45±0.13	79.57±0.17
FedProx [26]	98.01±0.09	55.06±0.75	25.94±0.16	98.82±0.01	59.21±0.40	31.99±0.41	19.37±0.22	19.27±0.23	79.35±0.23
Per-FedAvg [8]	99.63±0.02	89.63±0.23	56.80±0.26	98.90±0.05	87.74±0.19	44.28±0.33	25.07±0.07	21.81±0.54	93.27±0.25
pFedMe [42]	99.75±0.02	90.11±0.10	58.20±0.14	99.52±0.02	88.09±0.32	47.34±0.46	26.93±0.19	33.44±0.33	91.41±0.22
FedAMP [15]	99.76±0.02	90.79±0.16	64.34±0.37	99.47±0.02	88.70±0.18	47.69±0.49	27.99±0.11	29.11±0.15	94.18±0.09
Ditto [24]	99.81±0.00	92.39±0.06	67.23±0.07	99.64±0.00	90.59±0.01	52.87±0.64	32.15±0.04	35.92±0.43	95.45±0.17
FedPer [2]	99.70±0.02	91.15±0.21	63.53±0.21	99.47±0.04	89.22±0.33	49.63±0.54	33.84±0.34	38.45±0.85	95.54±0.32
FedRep [6]	99.77±0.03	91.93±0.14	67.56±0.31	99.48±0.02	90.40±0.24	52.39±0.35	37.27±0.20	39.95±0.61	96.28±0.14
FedRoD [4]	99.90±0.00	91.98±0.03	62.30±0.02	99.66±0.00	89.93±0.01	50.94±0.11	36.43±0.05	37.99±0.26	95.99±0.08
FedFomo [55]	99.83±0.00	91.85±0.02	62.49±0.22	99.33±0.04	88.06±0.02	45.39±0.45	26.33±0.22	26.84±0.11	95.84±0.15
FedPHP [27]	99.73±0.00	90.01±0.00	63.09±0.04	99.58±0.00	88.92±0.02	50.52±0.16	35.69±3.26	29.90±0.51	94.38±0.12
FedCP	**99.91±0.01**	**92.67±0.09**	**71.80±0.16**	**99.71±0.00**	**91.30±0.17**	**59.56±0.08**	**43.49±0.04**	**44.18±0.21**	**96.78±0.09**

Extensive Experiments

- FedCP outperforms 11 SOTA traditional FL and pFL methods in various settings and datasets

The accuracy (%) of the image/text classification tasks in the main experiments.

Settings	Pathological setting			Default practical setting (β = 0.1)					
	MNIST	Cifar10	Cifar100	MNIST	Cifar10	Cifar100	TINY	TINY*	AG News
FedAvg [32]	97.93±0.05	55.09±0.83	25.98±0.13	98.81±0.01	59.16±0.47	31.89±0.47	19.46±0.20	19.45±0.13	79.57±0.17
FedProx [26]	98.01±0.09	55.06±0.75	25.94±0.16	98.82±0.01	59.21±0.40	31.99±0.41	19.37±0.22	19.27±0.23	79.35±0.23
Per-FedAvg [8]	99.63±0.02	89.63±0.23	56.80±0.26	98.90±0.05	87.74±0.19	44.28±0.33	25.07±0.07	21.81±0.54	93.27±0.25
pFedMe [42]	99.75±0.02	90.11±0.10	58.20±0.14	99.52±0.02	88.09±0.32	47.34±0.46	26.93±0.19	33.44±0.33	91.41±0.22
FedAMP [15]	99.76±0.02	90.79±0.16	64.34±0.37	99.47±0.02	88.70±0.18	47.69±0.49	27.99±0.11	29.11±0.15	94.18±0.09
Ditto [24]	99.81±0.00	92.39±0.06	67.23±0.07	99.64±0.00	90.59±0.01	52.87±0.64	32.15±0.04	35.92±0.43	95.45±0.17
FedPer [2]	99.70±0.02	91.15±0.21	63.53±0.21	99.47±0.04	89.22±0.33	49.63±0.54	33.84±0.34	38.45±0.85	95.54±0.32
FedRep [6]	99.77±0.03	91.93±0.14	67.56±0.31	99.48±0.02	90.40±0.24	52.39±0.35	37.27±0.20	39.95±0.61	96.28±0.14
FedRoD [4]	99.90±0.00	91.98±0.03	62.30±0.02	99.66±0.00	89.93±0.01	50.94±0.11	36.43±0.05	37.99±0.26	95.99±0.08
FedFomo [55]	99.83±0.00	91.85±0.02	62.49±0.22	99.33±0.04	88.06±0.02	45.39±0.45	26.33±0.22	26.84±0.11	95.84±0.15
FedPHP [27]	99.73±0.00	90.01±0.00	63.09±0.04	99.58±0.00	88.92±0.02	50.52±0.16	35.69±3.26	29.90±0.51	94.38±0.12
FedCP	**99.91±0.01**	**92.67±0.09**	**71.80±0.16**	**99.71±0.00**	**91.30±0.17**	**59.56±0.08**	**43.49±0.04**	**44.18±0.21**	**96.78±0.09**

Extensive Experiments

- FedCP outperforms 11 SOTA methods on scalability

The accuracy (%) on Cifar100 for scalability.

	N = 10	N = 30	N = 50	N = 100	N = 200	N = 500
FedAvg	31.47±0.01	31.15±0.05	31.90±0.27	31.95±0.37	31.20±0.58	29.51±0.73
FedProx	31.24±0.08	31.21±0.08	31.94±0.30	31.97±0.24	31.22±0.62	29.84±0.81
Per-FedAvg	37.24±0.12	41.57±0.21	44.31±0.20	36.07±0.24	—	—
pFedMe	44.06±0.29	47.04±0.28	48.36±0.64	46.45±0.18	39.55±0.61	31.30±0.89
FedAMP	49.23±0.18	45.33±0.04	44.39±0.35	40.43±0.17	35.40±0.70	*diverged*
Ditto	52.32±0.19	52.53±0.42	54.22±0.04	52.89±0.22	35.18±0.53	30.24±0.72
FedPer	50.31±0.19	44.98±0.20	44.22±0.18	40.37±0.41	34.99±0.48	30.56±0.59
FedRep	52.89±0.10	50.24±0.01	47.41±0.18	44.61±0.20	36.79±0.60	31.92±0.71
FedRoD	49.83±0.07	50.11±0.03	49.38±0.01	46.65±0.22	43.53±0.86	34.61±0.98
FedFomo	46.71±0.23	43.20±0.05	42.56±0.33	38.91±0.08	34.79±0.71	29.24±1.28
FedPHP	49.32±0.19	49.28±0.06	52.44±0.16	49.70±0.31	34.48±0.33	30.26±0.84
FedCP	**58.36±0.02**	**56.93±0.19**	**55.43±0.21**	**53.81±0.32**	**44.86±0.87**	**35.87±0.52**

Extensive Experiments

- FedCP outperforms 11 SOTA methods on scalability in real-world scenarios

The accuracy (%) on Cifar100 for scalability in real-world scenarios.

	N = 10|50	N = 30|50	N = 50
FedAvg	25.28±0.32	29.04±0.21	31.90±0.27
FedProx	25.65±0.34	29.04±0.36	31.94±0.30
Per-FedAvg	40.20±0.21	42.96±0.42	44.31±0.20
pFedMe	40.27±0.54	42.19±0.38	48.36±0.64
FedAMP	43.57±0.30	43.18±0.31	44.39±0.35
Ditto	48.23±0.35	50.98±0.29	54.22±0.04
FedPer	43.64±0.42	43.54±0.43	44.22±0.18
FedRep	46.85±0.12	47.63±0.26	47.41±0.18
FedRoD	46.32±0.02	49.15±0.12	49.38±0.01
FedFomo	41.53±0.45	40.69±0.41	42.56±0.33
FedPHP	45.71±0.21	48.65±0.24	52.44±0.16
FedCP	**50.93±0.34**	**54.31±0.25**	**55.43±0.21**

Extensive Experiments

- FedCP keeps superiority with large local epochs

The accuracy (%) on Cifar10 in the default practical setting with large local epochs.

Local epochs	5	10	20	40
FedAvg	57.51±0.35	57.55±0.32	57.28±0.23	56.27±0.29
FedProx	57.48±0.28	57.69±0.31	57.53±0.33	56.18±0.24
Per-FedAvg	86.13±0.12	86.09±0.19	85.57±0.15	85.45±0.16
pFedMe	88.72±0.02	88.58±0.17	88.37±0.14	88.16±0.20
FedAMP	88.72±0.21	88.77±0.27	88.76±0.30	88.70±0.26
Ditto	90.79±0.21	90.59±0.06	90.34±0.23	90.02±0.38
FedPer	89.62±0.12	89.73±0.31	89.79±0.35	89.49±0.55
FedRep	90.20±0.41	90.08±0.26	89.46±0.13	89.22±0.25
FedRoD	89.71±0.32	89.11±0.33	88.13±0.21	87.55±0.28
FedFomo	88.39±0.15	88.43±0.16	88.41±0.13	88.13±0.32
FedPHP	90.29±0.37	90.03±0.23	89.92±0.27	89.87±0.26
FedCP	**91.13±0.34**	**91.24±0.31**	**91.02±0.28**	**90.86±0.37**

Extensive Experiments

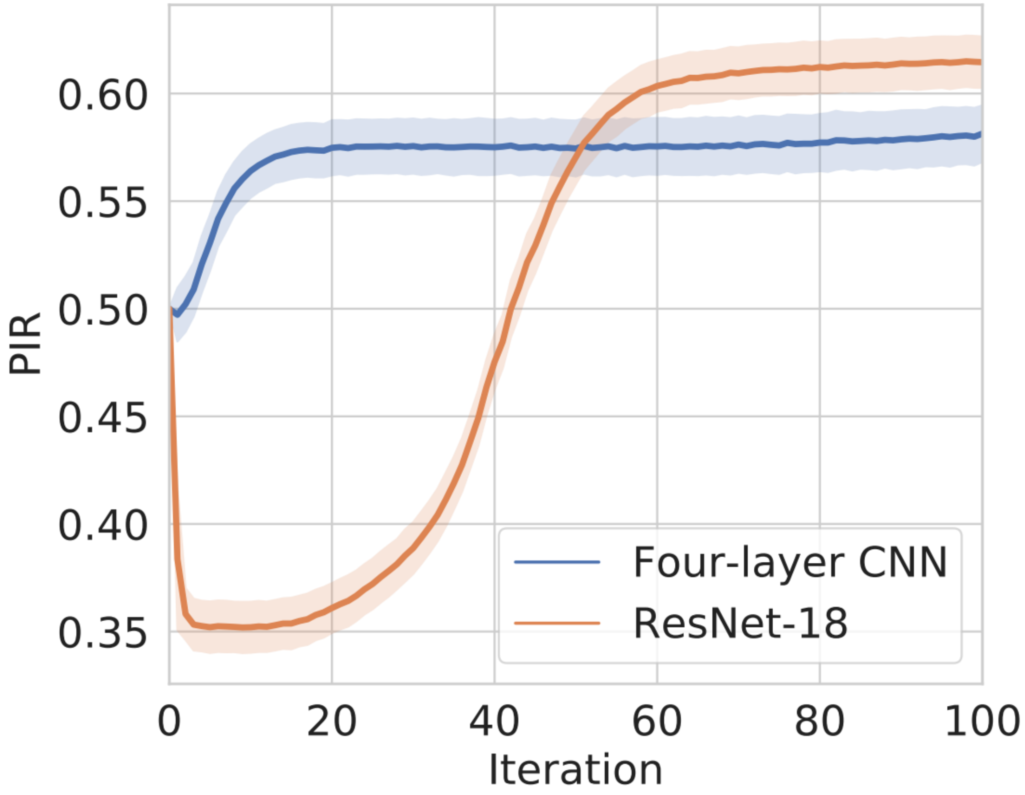
- FedCP keeps superiority in unstable settings with clients randomly drop out

The accuracy (%) on Cifar100 (N= 50, β= 0.1) when clients accidentally drop out.

	$\rho = 1$	$\rho \in [0.5, 1]$	$\rho \in [0.1, 1]$
Per-FedAvg	44.31±0.20	43.66±1.38	43.63±1.07
pFedMe	48.36±0.64	43.28±0.85	41.71±1.02
FedAMP	44.39±0.35	42.91±0.08	42.92±0.14
Ditto	50.59±0.22	49.78±0.36	48.33±3.27
FedPer	44.22±0.18	44.12±0.21	44.07±0.27
FedRep	47.41±0.18	46.93±0.21	46.61±0.22
FedRoD	49.38±0.01	49.07±0.43	47.80±1.35
FedFomo	42.56±0.33	40.96±0.02	40.93±0.07
FedPHP	50.23±0.12	45.19±0.07	44.43±0.12
FedCP	**54.81±0.20**	**54.68±0.35**	**54.20±0.21**

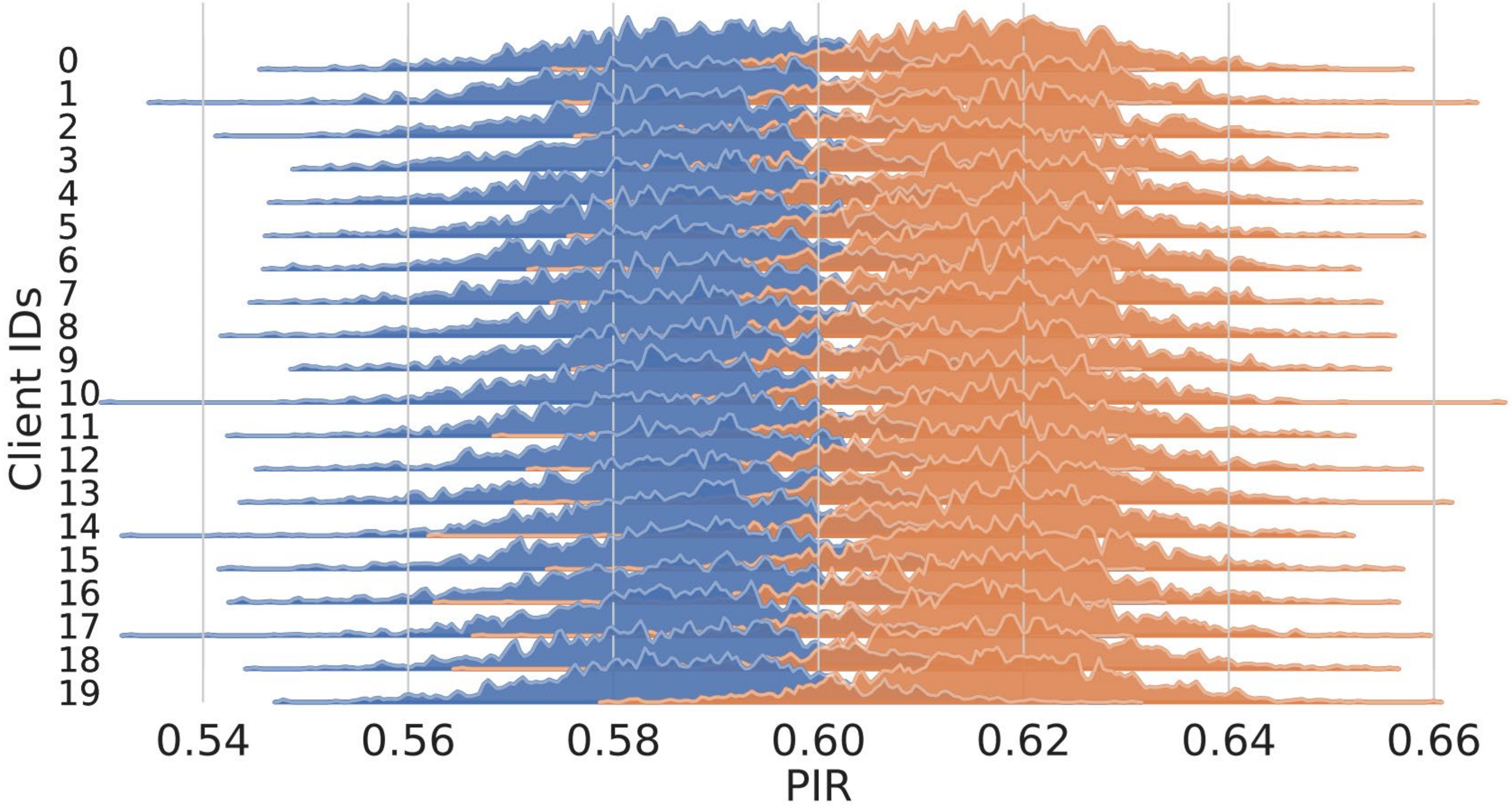
Policy Study

- Personalization Identification Ratio (PIR) change on client #0 in FedCP

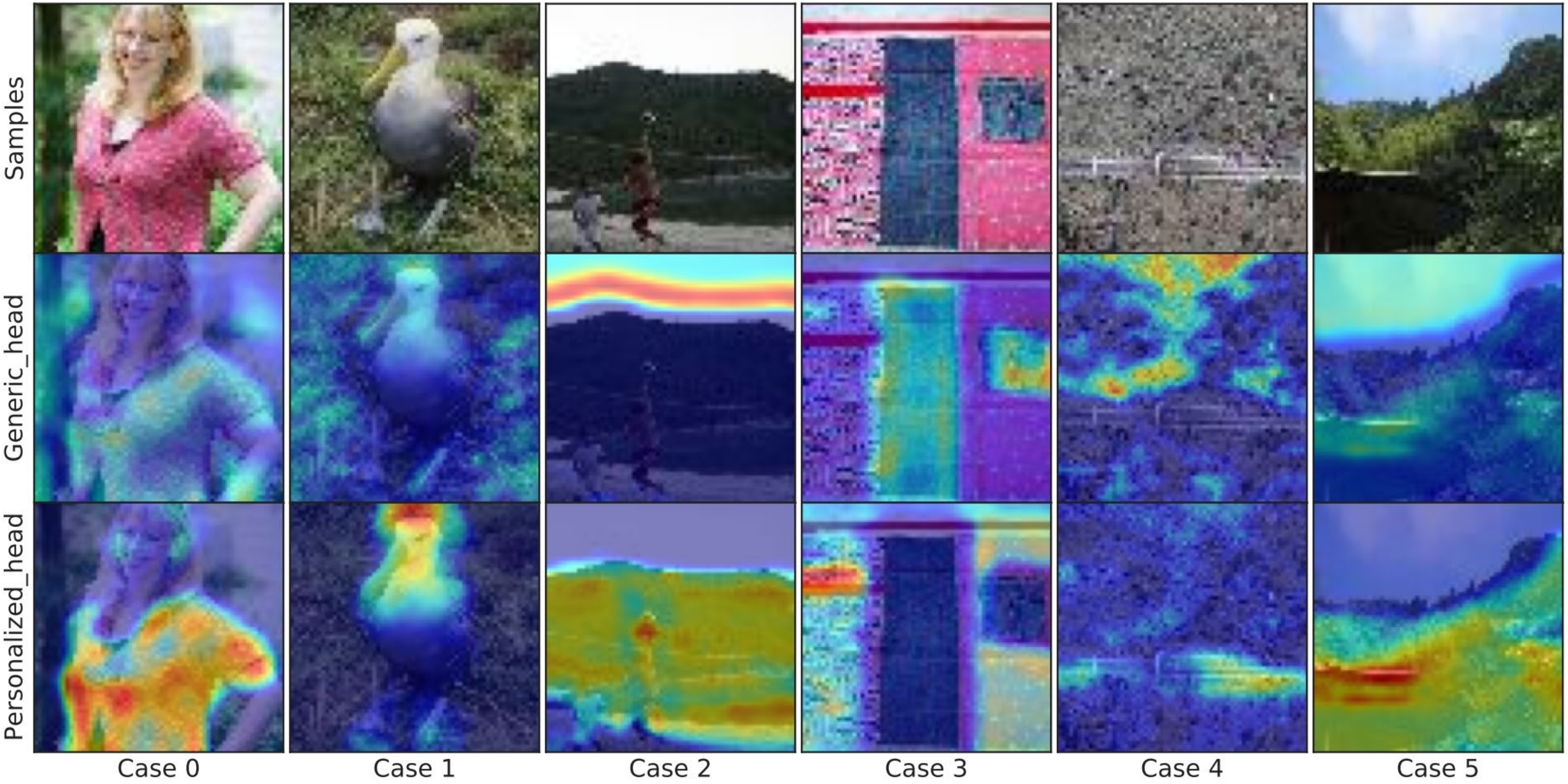


Policy Study

- $\boldsymbol{s}_i$ distribution of test samples on all clients



Use FedCP to Separate Feature Information Now



FedCP: Separating Feature Information for Personalized Federated Learning via Conditional Policy

Paper: https://arxiv.org/abs/2307.01217

Code: https://github.com/TsingZ0/FedCP

E-mail: tsingz@sjtu.edu.cn

Paper

Code

Thanks!