INFOCOM 2026 Workshop on Embodied Intelligence Networks (EIN)

pFedNavi: Structure-Aware Personalized Federated Vision-Language Navigation for Embodied AI

Qingqian Yang[1], Hao Wang[1], Sai Qian Zhang[2], Jian Li[3], Yang Hua[4], Miao Pan[5], Tao Song[6], Zhengwei Qi[6], and Haibin Guan[6]

IntelliSys Lab

Stevens Institute of Technology[1], New York University[2], Stony Brook University[3], Queen's University Belfast[4], University of Houston[5], Shanghai Jiao Tong University[6]

2026/05/18

Vision Language Navigation (VLN)

An embodied agent follows natural-language instructions to reach a target location.

Instruction

→ Observation

→ Action

→ Goal

Client *i*: “Fetch a bottle of drink from kitchen to bedroom”

Example of VLN task

Federated VLN: Privacy-Preserving Training

Client *i*'s VLN Model

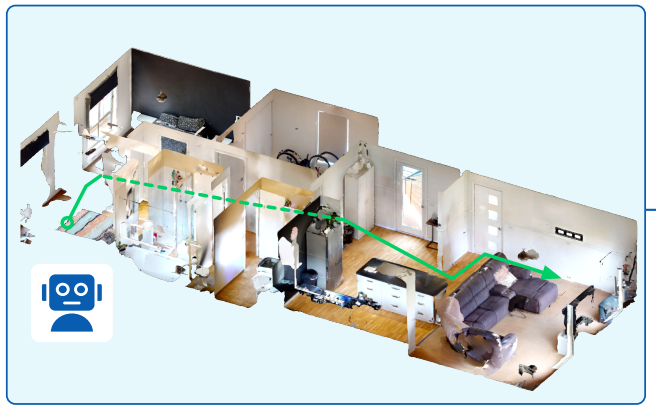
$\theta_{i,Enc}$ Language Encoder

$\theta_{i,Dec}$ Navigation Decoder

$\theta_{i,Cri}$ RL Critic Module

Local upload *Aggregate*

Client *i*: "Fetch a bottle of drink from kitchen to bedroom"

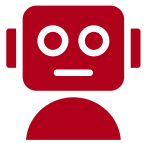


Client *j*'s VLN Model

Federated Server

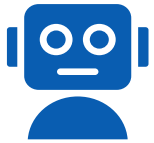
Existing Works on Federated VLN *(FedVLN)*

$$\Theta^{t+1} = \sum_{i \in \mathcal{S}_t} \frac{|\mathcal{D}_i|}{\sum_{j \in \mathcal{S}_t} |\mathcal{D}_j|} \Theta_i^{t+1}$$



Client *i*'s VLN Model

$$\Theta_i^t = \{\theta_{i,\text{Enc}}^t, \theta_{i,\text{Dec}}^t, \theta_{i,\text{Cri}}^t\}$$



Client *j*'s VLN Model

$$\Theta_i^t = \{\theta_{j,\text{Enc}}^t, \theta_{j,\text{Dec}}^t, \theta_{j,\text{Cri}}^t\}$$

. . .

Zhou, Kaiwen, and Xin Eric Wang. "Fedvln: Privacy-preserving federated vision-and-language navigation." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022.

Key Problems — Heterogeneity

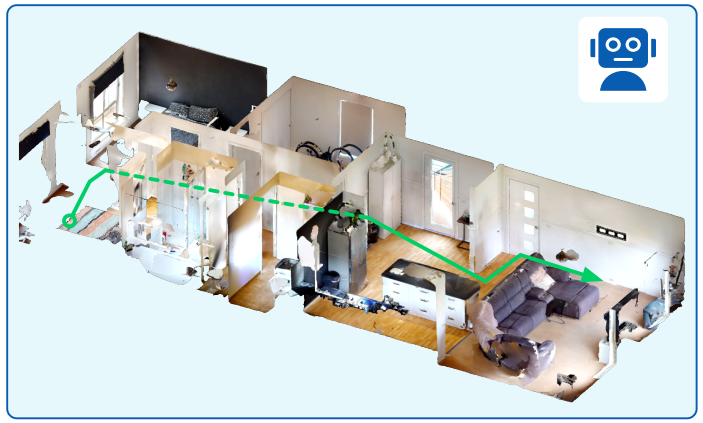
- **Personalized Instructions**:
  - *different vacabulary*
  - *detail levels*
  - *landmarks*

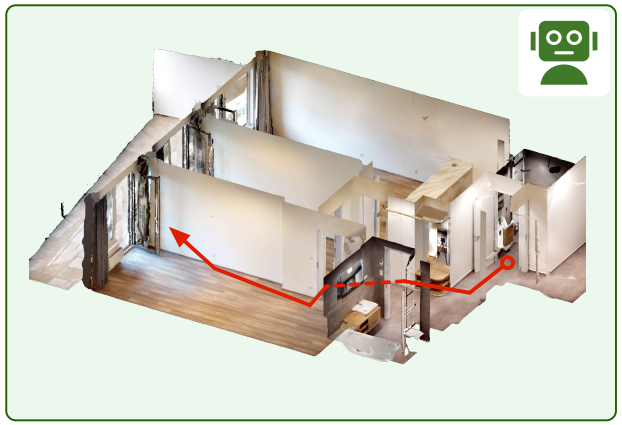
Client *i*: “Fetch a bottle of drink from kitchen to bedroom”

Client *j*: “Go through the hallway ... to the sofa”

Client *k*: “Go to the balcony via the living room”

...



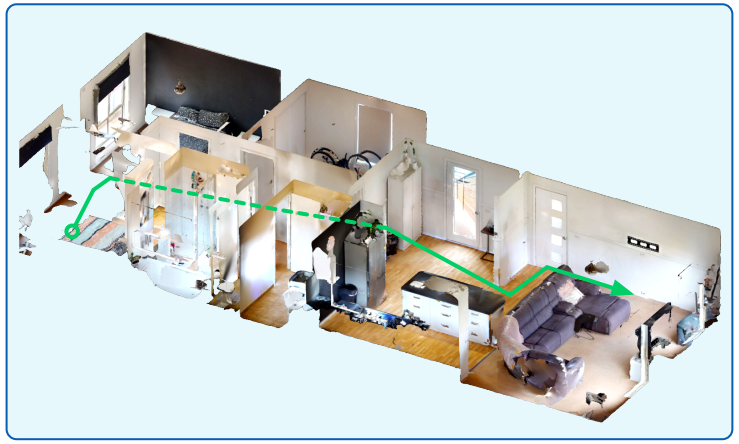


Key Problems — Heterogeneity

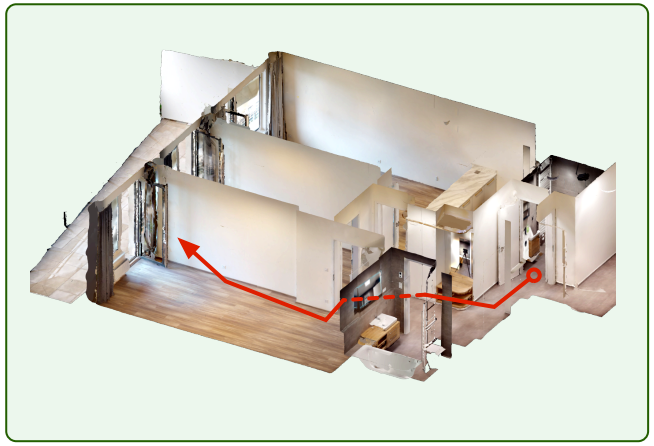
Client *i*: “Fetch a bottle of drink from kitchen to bedroom”

…

Client *j*: “Go through the hallway ... to the sofa”



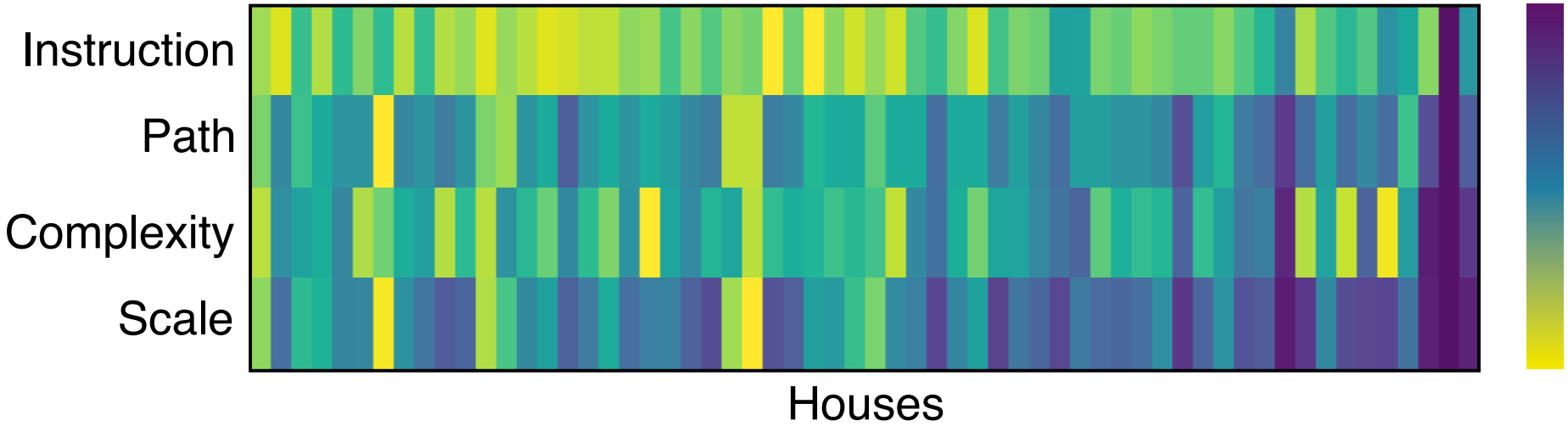
Client *k*: “Go to the balcony via the living room”



- **Heterogeneous Environments:**
  - *layout*
  - *scale*
  - *graph complexity*
  - *path distribution*

Key Problems — Heterogeneity

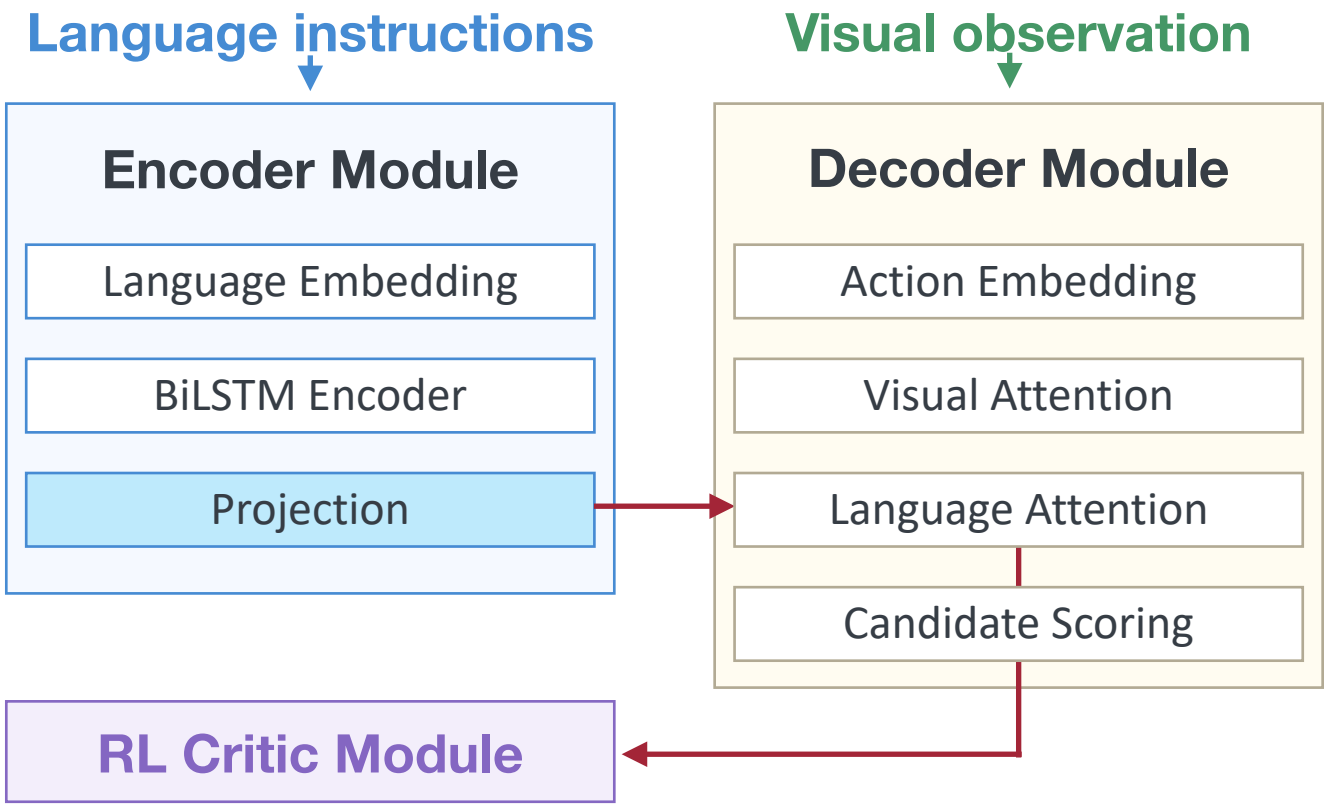
- **Data heterogeneity analysis on RxR dataset [1]:**
  - *layout*
  - *scale*
  - *graph complexity*
  - *path distribution*



[1] Ku, Alexander, et al. "Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

Key Problems — Multimodal Models

Traditional Vision Model



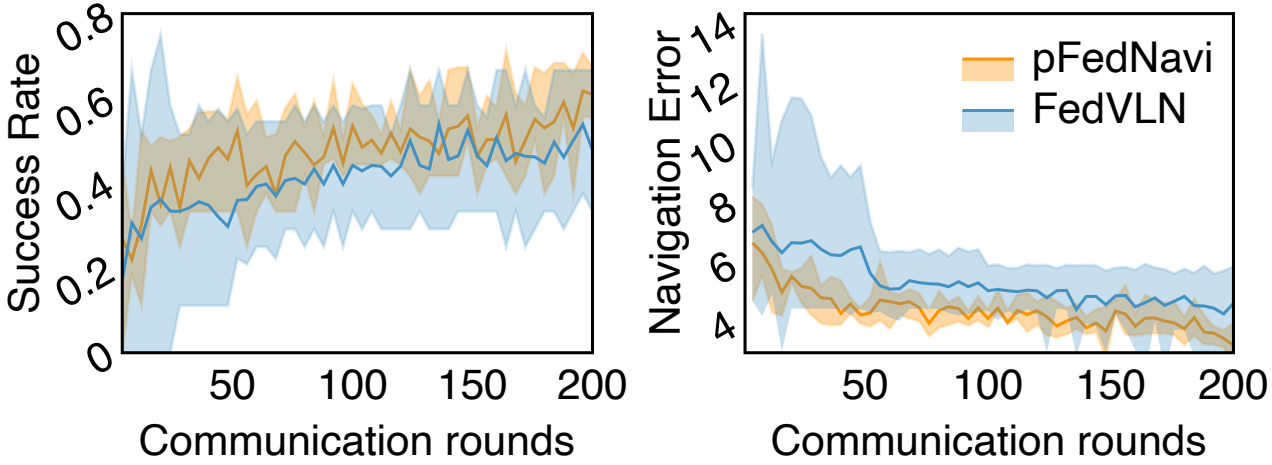
VLN Navigation model

Personalized Federated VLN

Heterogeneous Environments

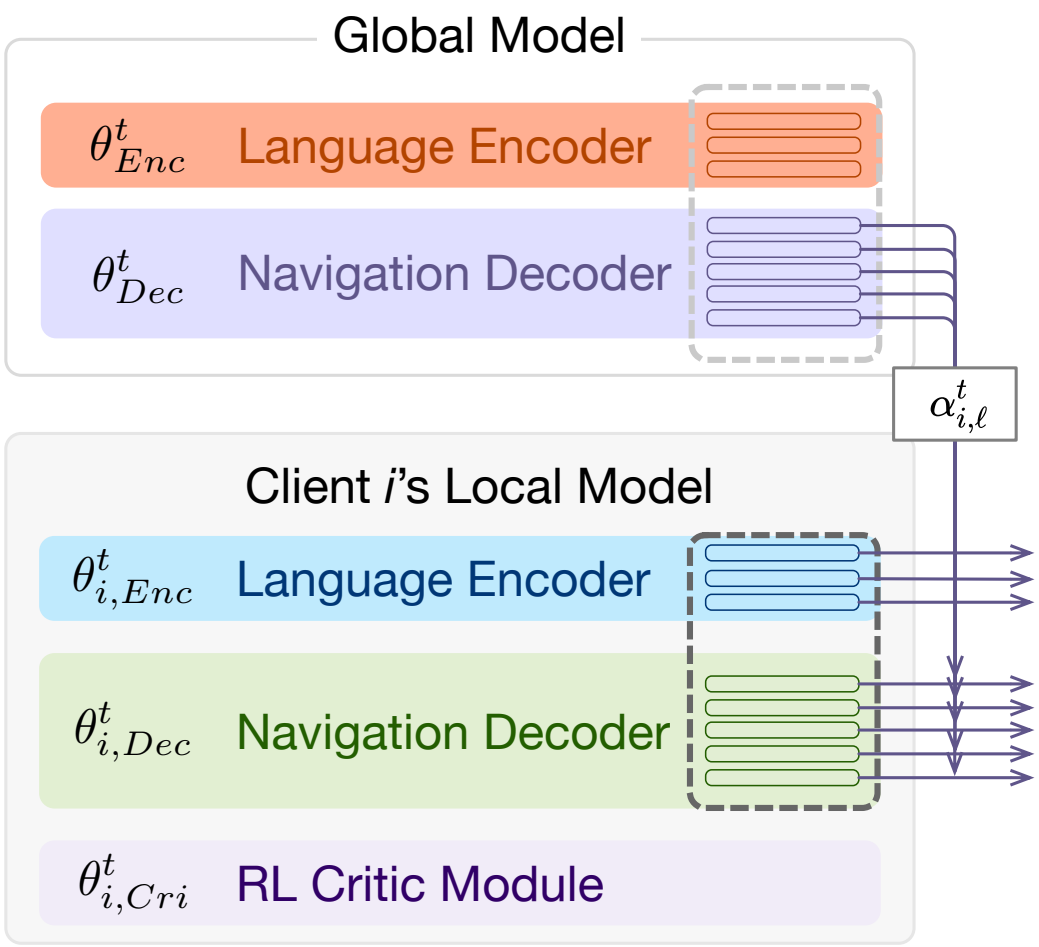
Personalized Instructions

Complex Multimodal Models



Non-personalized FL (FedVLN) vs. personalized FL (pFedNavi).

Adaptive Layer Selection for Personalization



$$\text{Personalized\_layer}_i^t = \{\ell \in \text{L}_{\text{dec}} \mid \alpha_{i,\ell}^t \geq \delta\}.$$

Layers	Coefficient	Decision
0	0.31	Shared
1	0.84	Personalized
	...	

Select layers

Training

$$\arg\min_{\boldsymbol{\alpha}} \mathcal{L}_{\text{IL}}\left(\mathcal{B}_i;\ \hat{\theta}_i^{t+1,0}(\boldsymbol{\alpha})\right).$$

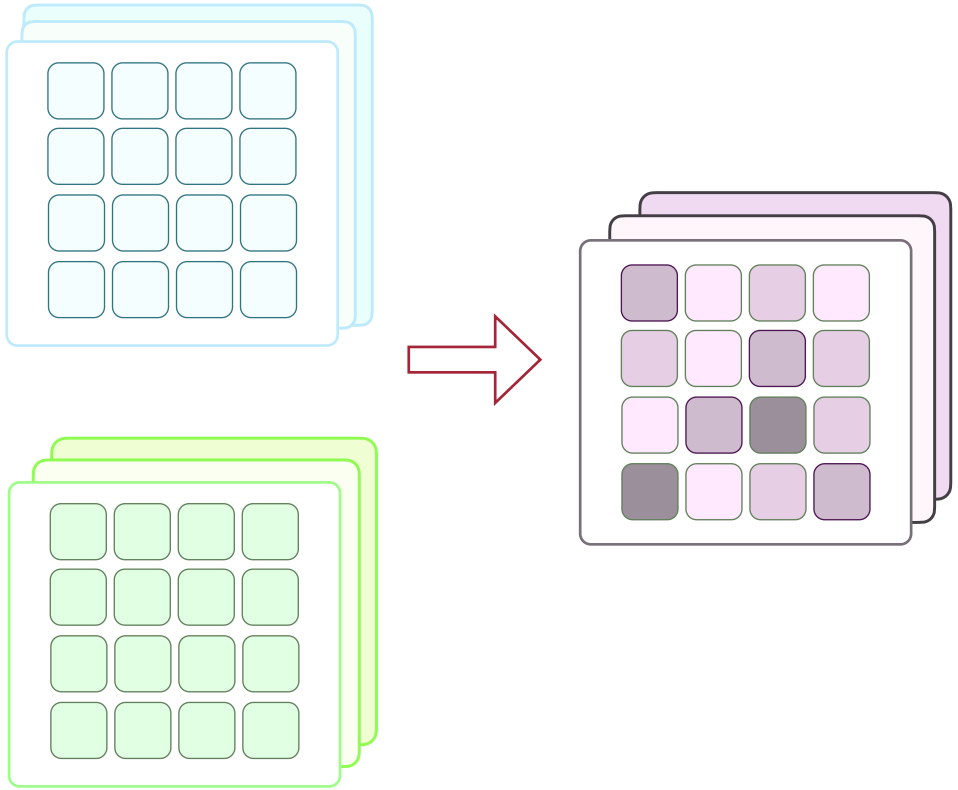
Learning mixing coefficients $\alpha^t_{i,\ell}$

Fine-grained Parameter Fusion

For each selected personalized layers:

Global Layer

Previous Local Layer



Fusion formulation

$$\theta_{i,l}^{t+1,0} = \theta_{i,l}^{t} + W_{i,l}^{t} \odot \left(\theta_{l}^{t} - \theta_{i,l}^{t}\right)$$

Optimized by:

$$\min_{\mathbf{W}_i^t} \mathcal{L}_{\mathrm{IL}}\left(\mathcal{D}_i;\ \mathbf{\Theta}_i^{t+1,0}(\mathbf{W}_i^t)\right)$$

Evaluation

- Datasets:
  - R2R
  - RxR
- Models:
  - Seq2seq model
- Baselines:
  - EnvDrop (centralized setting)
  - FedVLN (federated setting)
- Pretrained Image features:
  - ResNet-152 representations
  - CLIP representations
- Metrics:
  - Success Rate (SR)
  - Success weighted by Path Length (SPL)
  - Oracle Success Rate (OSR)
  - Navigation Error (NE)
  - Coverage weighted by Length Score (CLS)
  - Normalized Dynamic Time Warping (nDTW)

Main Results

Improves Goal-Reaching Ability

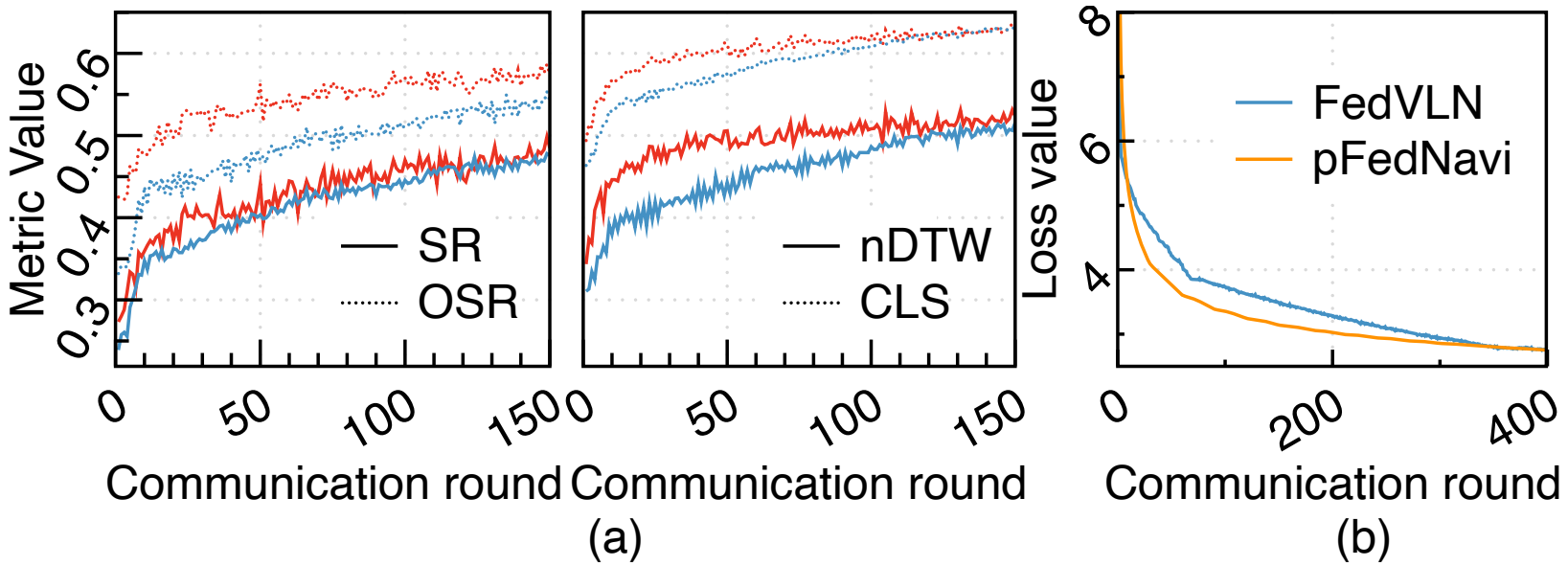
up to 7.5%

Improves Trajectory Fidelity

up to 7.8%

	Method on R2R	SR ↑	SPL ↑	OSR ↑	CLS ↑	nDTW ↑	NE ↓
ResNet	EnvDrop [9]	**56.7**	**54.3**	63.9	**67.2**	**56.0**	**4.55**
	FedVLN [2]	50.7	47.3	60.3	63.4	50.8	5.37
	pFedNavi	54.5	51.7	**65.2**	66.4	54.8	4.86
CLIP	EnvDrop [9]	**61.5**	56.5	**69.8**	67.3	55.4	3.94
	FedVLN [2]	59.5	56.3	65.1	67.5	55.1	4.24
	pFedNavi	60.7	**57.4**	67.3	**68.6**	**57.1**	**3.65**
	Method on RxR	**SR ↑**	**SPL ↑**	**OSR ↑**	**CLS ↑**	**nDTW ↑**	**NE ↓**
ResNet	EnvDrop [9]	**41.3**	36.6	50.4	57.2	**53.6**	**8.03**
	FedVLN [2]	39.9	33.1	47.8	55.5	50.7	8.62
	pFedNavi	40.1	**36.7**	**51.2**	**57.3**	52.9	8.10
CLIP	EnvDrop [9]	**47.7**	**43.4**	56.1	**61.2**	56.8	**6.53**
	FedVLN [2]	43.0	39.4	51.5	58.7	54.8	7.37
	pFedNavi	46.1	41.2	**56.3**	59.8	**56.9**	6.91

Main Results



Higher efficiency: 1.38x faster convergence

Various personalized layer selection strategies

pFedNavi

All Layers -18.25% on SPL

No Layers -9.80% on SPL

IntelliSys Lab

STEVENS
INSTITUTE OF TECHNOLOGY
1870

Code

Paper

Stevens Institute of Technology
1 Castle Point Terrace, Hoboken, NJ 07030