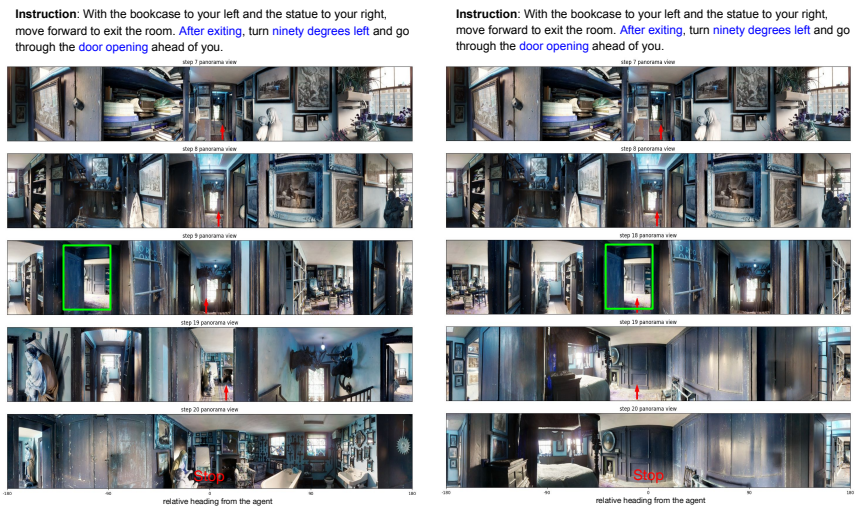
FedVLN: Privacy-preserving Federated Vision-and-Language Navigation – Supplement Materials

Kaiwen Zhou and Xin Eric Wang

University of California, Santa Cruz, CA 95064, USA
kzhou35@ucsc.edu,xwang366@ucsc.edu

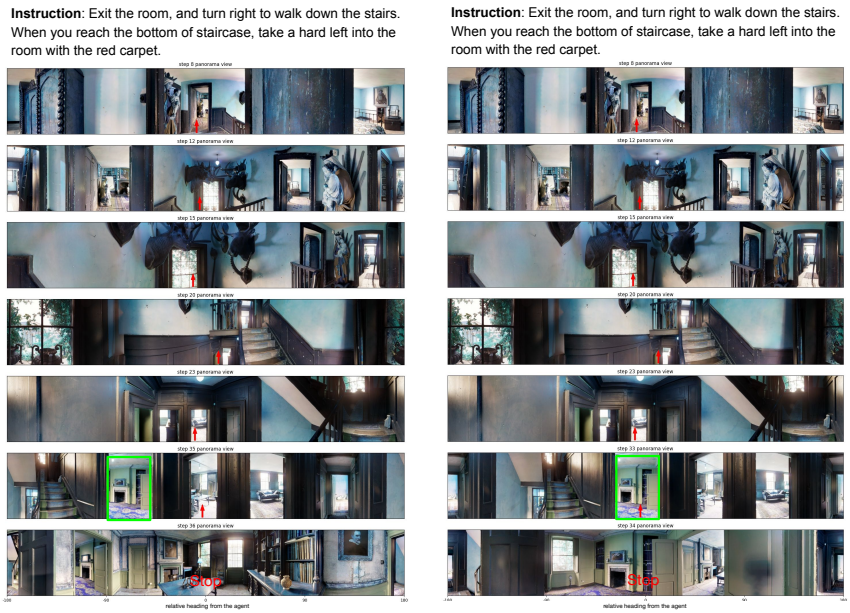
A Case analysis



(a) Environment-based (b) Fed-Enc

Fig. 1. Case study between environment-based pre-exploration and language encoder-sharing federated pre-exploration. In this example, the agent trained by environment-based pre-exploration can not fully understand that there will be an open door on the left as soon as it exits the room.

In this section, we further demonstrate the advantage of sharing language encoder and training with seen environments during federated pre-exploration by visualizing the navigation trajectory examples from environment-based pre-exploration [1], language encoder-sharing federated pre-exploration and federated pre-exploration sharing language encoder across seen and unseen environments. We also show how centralized training overfit better than federated training on seen environment.



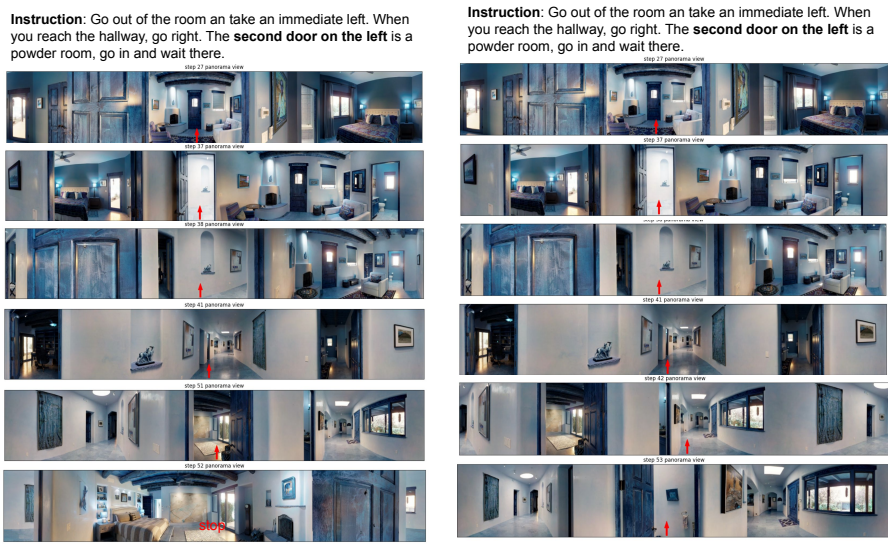
(a) Fed-Enc (b) Fed-Enc+seen

Fig. 2. Case study between language encoder-sharing federated pre-exploration with and without seen environments. In this example, the agent trained by federated pre-exploration without seen environments can not understand and notice the keyword 'red carpet'.

First, the agent trained by environment-based pre-exploration can not understand the language instruction and execute it precisely. Fig 1 demonstrates a qualitative example for this. The environment-based pre-explored agent successfully exits the room, but it does not notice the opened door at 90 degrees left and keeps going ahead, and finally enters the wrong room. The agent trained by sharing language encoder in federated pre-exploration can understand the words 'after exiting', 'ninety degrees left', and 'door opening' precisely and notice the open door right after it exits the room, thus navigating successfully.

Second, federated pre-exploration training with seen environments can further improve the language understanding ability of the agent, especially for words that the speaker generates rarely. In Fig 2, the agent trained by federated pre-exploration sharing encoder without seen environments navigates successfully at most part of the path. However, it can not understand 'red carpet' well and target the correct direction, thus entering the wrong room at the end, although the red carpet is within its vision all the time. The agent trained by federated pre-exploration with seen environments recognizes the red carpet and navigates successfully.

Third, the agent trained by centralized training memorized path appeared in the training data better than federated training. From Table. 1, we can see that



(a) Federated training (b) Centralized training

Fig. 3. Case study between decentralized federated training and centralized training on seen environment.

training	valid seen
“f4939bf6f00a4864832a358f1ea8394e”,	“fae83673fc694cd9a18c215ce6d92c58”,
“28c09c307b11487c999f88e1e9ec3231”,	“28c09c307b11487c999f88e1e9ec3231”,
“ecff9ecf0cfe4d8bb83260fc092f3b00”,	“ecff9ecf0cfe4d8bb83260fc092f3b00”,
“d1ffe5280fce4ac5a949cdc9ee8b6f7c”,	“d1ffe5280fce4ac5a949cdc9ee8b6f7c”,
“dbc0fd77d9384e14a6d0a19302b85a15”,	“dbc0fd77d9384e14a6d0a19302b85a15”,
“4ff62efbc0934e888120522e4c84e712”,	“4ff62efbc0934e888120522e4c84e712”,
“982920829a0b433880410222539f240e”	“982920829a0b433880410222539f240e”

Table 1. Comparison of two similar paths in training set and validation seen. The path in validation seen is used in Fig. 3.

the chosen example validation path is highly overlap with a training path with the same later part. In this example, as shown in Fig. 3, centralized training successfully navigate to the location, while federated trained agent fail at the later part of the path and choose the wrong room in the hallway.

B Discussion of convergence speed–performance trade-off

In federated learning, the convergence speed is also an important indicator to evaluate the model except for model performance. In federated vision-and-language navigation, faster convergence speed means fewer communication rounds

η	40	41	42	43	44	45	46
1	164	192	266	312	–	–	–
2	89	121	163	186	268	–	–
6	31	57	62	84	109	151	197
12	**29**	**29**	**42**	**47**	**73**	**89**	**89**

Table 2. Comparison of communication rounds needed to achieve target success rates(%) between different server learning rates η (10^{-4}), based on Envdrop model on R2R unseen validation. Entries with '–' means the model can not achieve a certain SR within 365 communication rounds.

E	41	43	45	47
1	180	264	421	798
2	76	123	168	717
5	57	84	151	–
10	24	34	**47**	**127**
20	**18**	**27**	–	–

Table 3. Comparison of communication rounds needed to achieve target success rates(%) on R2R unseen validation between different local epochs(E) based on Envdrop model, with fix server learning rate of 6×10^{-4}.

E	48	50	52	54
1	258	476	775	–
2	201	274	443	**573**
5	88	136	269	–
10	**75**	**98**	**156**	–
20	**75**	99	–	–

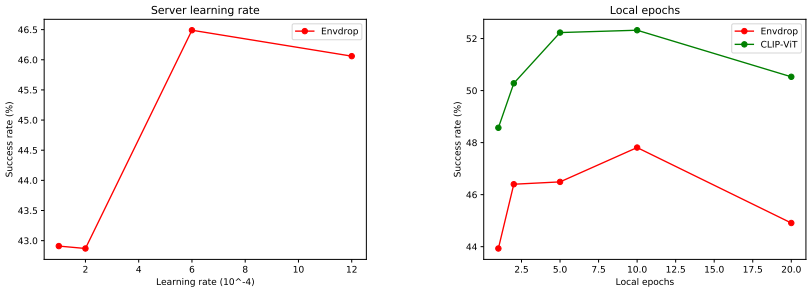
Table 4. Comparison of communication rounds needed to achieve target success rates(%) on R2R unseen validation between different local epochs(E) based on CLIP-ViT model, with fix server learning rate of 12×10^{-4}.

towards a target success rate. The local models in the environments can thus achieve satisfactory navigation performance sooner. Also, fewer communication rounds lead to less communication overhead [2] and privacy leakage. Thus, in this section, we discuss the influence of two hyper-parameters, server learning rate and communication frequency on convergence speed and model performance by ablation studies.

Server learning rate As shown in Table 2, generally, the model converges faster with a larger server learning rate, and can achieve good performance within 365 communication rounds. When the server learning rate is low, the scale of the global model update is only about $\frac{\eta}{rn}$ of the update scale with centralized training, where r is participation rate and n is the number of clients, which will lead to slower convergence. Also, as in Fig 4(a), using server learning rates of 6×10^{-4} and 12×10^{-4} achieves significantly higher performance within 300 communication rounds. Thus, it is better to scale up the server learning rate with more clients.

Communication frequency As shown in Table 3 and Table 4[1], overall, using small local epochs can achieve better performance at the end. However, the communication rounds it takes may be unacceptable and impractical in application.

[1] The models are trained till convergence here.



(a) Server learning rate (b) Communication frequency

Fig. 4. The best SR(%) on R2R validation for different server learning rates or local epochs within 300 communication rounds.

Metrics	Cent	Fed	Cent	Env	Fed-Full	FedLan+seen
Step (10^2)	1,064	1,109	235	110	189	140

Table 5. The first two columns are for seen environment training, last four columns are for pre-exploration.

Using larger local epochs leads to faster convergence, while it suffers from local over-fitting and can not achieve better navigation performance. In Fig 4(a), with local training epochs of 5 or 10 in each communication round, the model achieves better performance within 300 communication rounds. Thus, it's more practical to use a local training epoch of medium size (5-10 in our case), which can achieve faster convergence and will not lead to local over-fitting.

C Comparison of training speed

We here evaluate the training efficiency of federated learning. On seen environment training, the average training steps towards best SR are 1.06×10^5 for centralized training, and 1.11×10^5 for federated training. On pre-exploration, our final federated pre-exploration method uses only 1.40×10^4 steps to achieve best performance, which improves over full model sharing federated learning and centralized training. The training steps of 'FedLan+seen' is larger than environment-based pre-exploration but not by a large margin. More importantly, our federated learning framework does not bring any extra cost during inference and maintains high inference efficiency.

References

1. Fu, T.J., Wang, X.E., Peterson, M.F., Grafton, S.T., Eckstein, M.P., Wang, W.Y.: Counterfactual vision-and-language navigation via adversarial path sampler. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 71–86 (2020)
2. Vanhaesebrouck, P., Bellet, A., Tommasi, M.: Decentralized collaborative learning of personalized models over networks. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA. pp. 509–517. Proceedings of Machine Learning Research (2017)