THE UNIVERSITY OF ARIZONA

MLSys

FLoRIST: Singular Value Thresholding for Efficient and Accurate Federated Fine-Tuning of Large Language Models

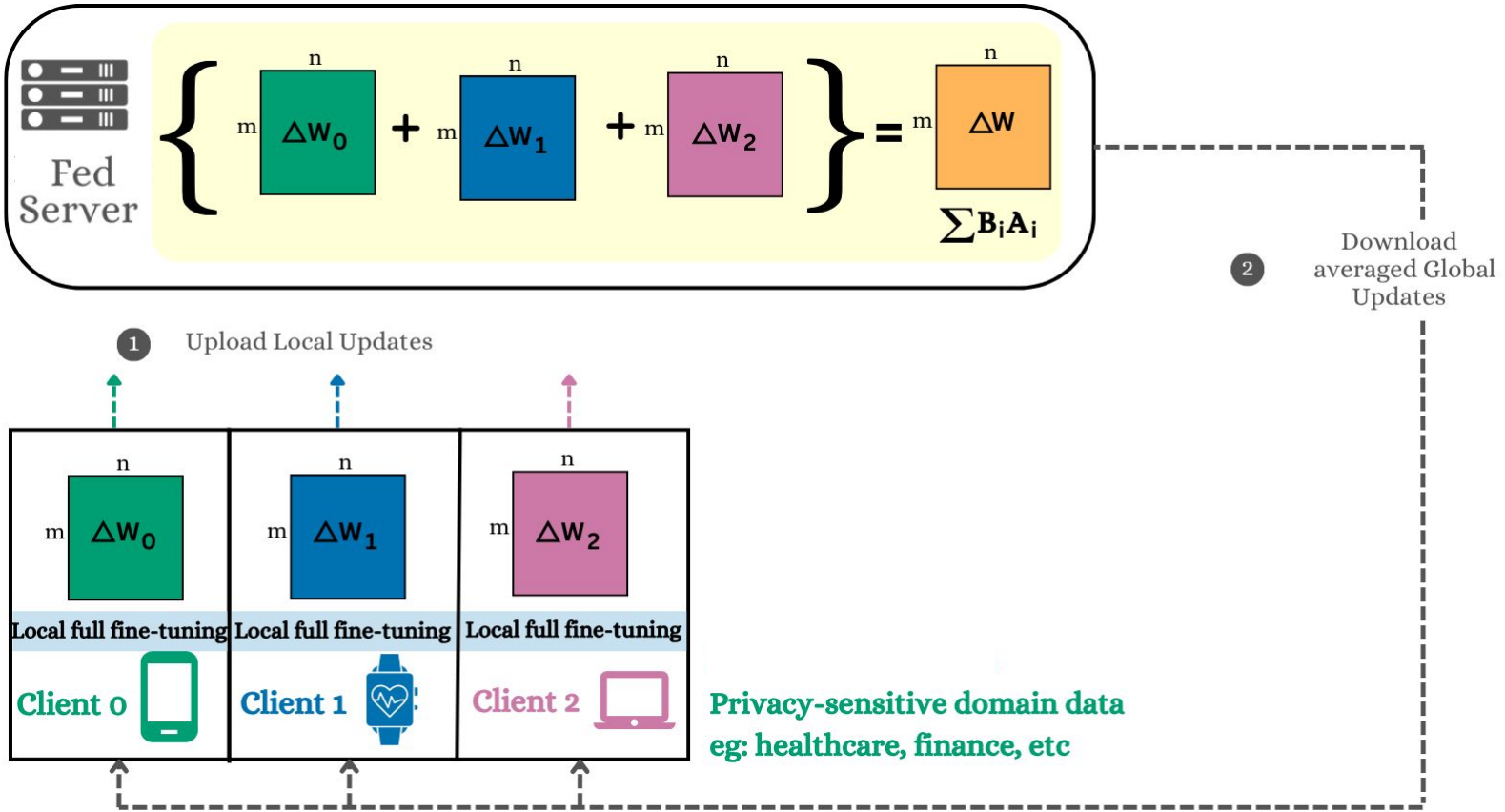
Hariharan Ramesh *Jyotikrishna Dass*

Distributed AI & Smart Systems (DASS) Lab
Electrical and Computer Engineering
University of Arizona

Paper

Code

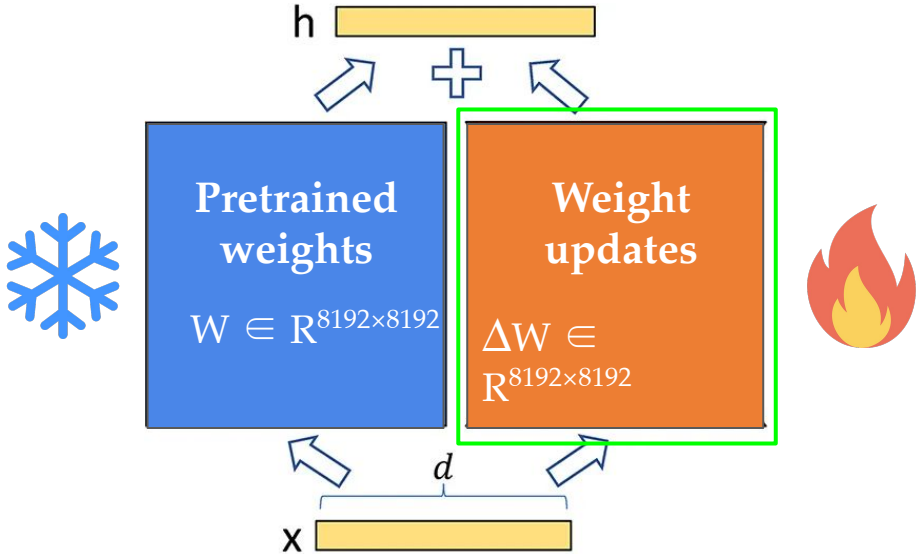
Background: Federated Full Fine-Tuning



THE UNIVERSITY OF ARIZONA

Background: Low-Rank Adaptation (LoRA)

Parameter Efficient Fine-Tuning



LlaMA-3.2-1B, $\mathbf{W} \in R^{8192\times8192}$

LoRA fine-tuning (rank, **r = 16**) creates efficient adapters, $\{\mathbf{B} \in R^{8192\times16}, \mathbf{A} \in R^{16\times8192}\}$

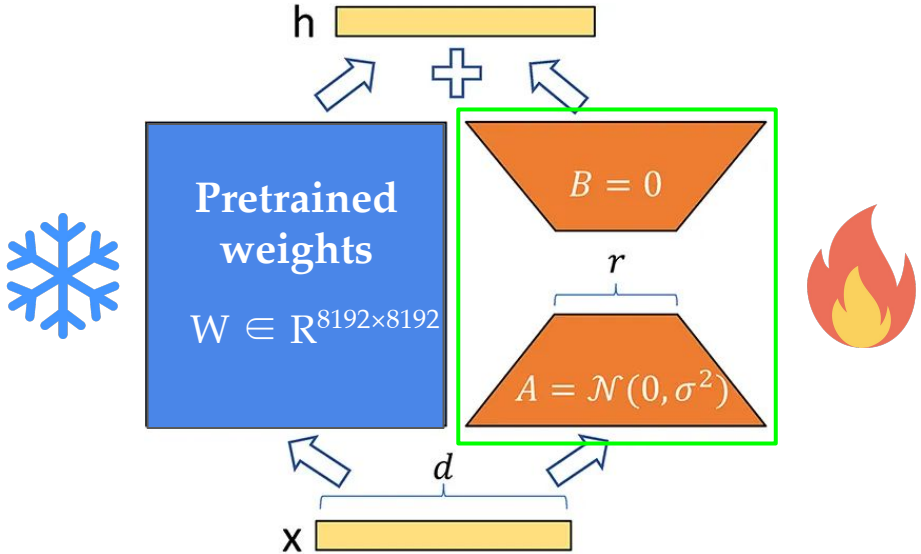
256x reduction in trainable parameters

Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," ICLR 2022.

THE UNIVERSITY OF ARIZONA

Background: Low-Rank Adaptation (LoRA)

Parameter Efficient Fine-Tuning



LlaMA-3.2-1B, $\mathbf{W} \in R^{8192\times8192}$

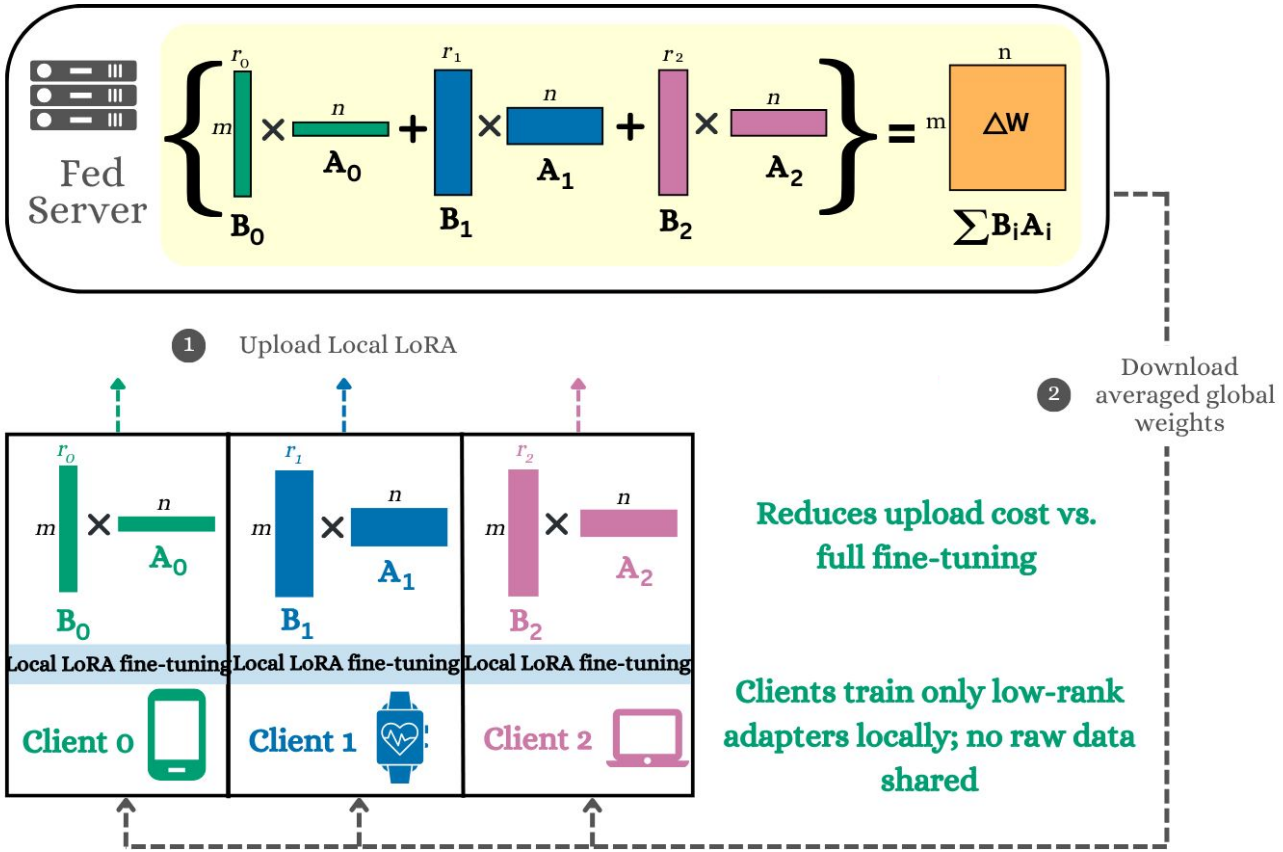
LoRA fine-tuning (rank, **r = 16**) creates efficient adapters, $\{\mathbf{B} \in R^{8192\times16}, \mathbf{A} \in R^{16\times8192}\}$

256x reduction in trainable parameters

Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," ICLR 2022.

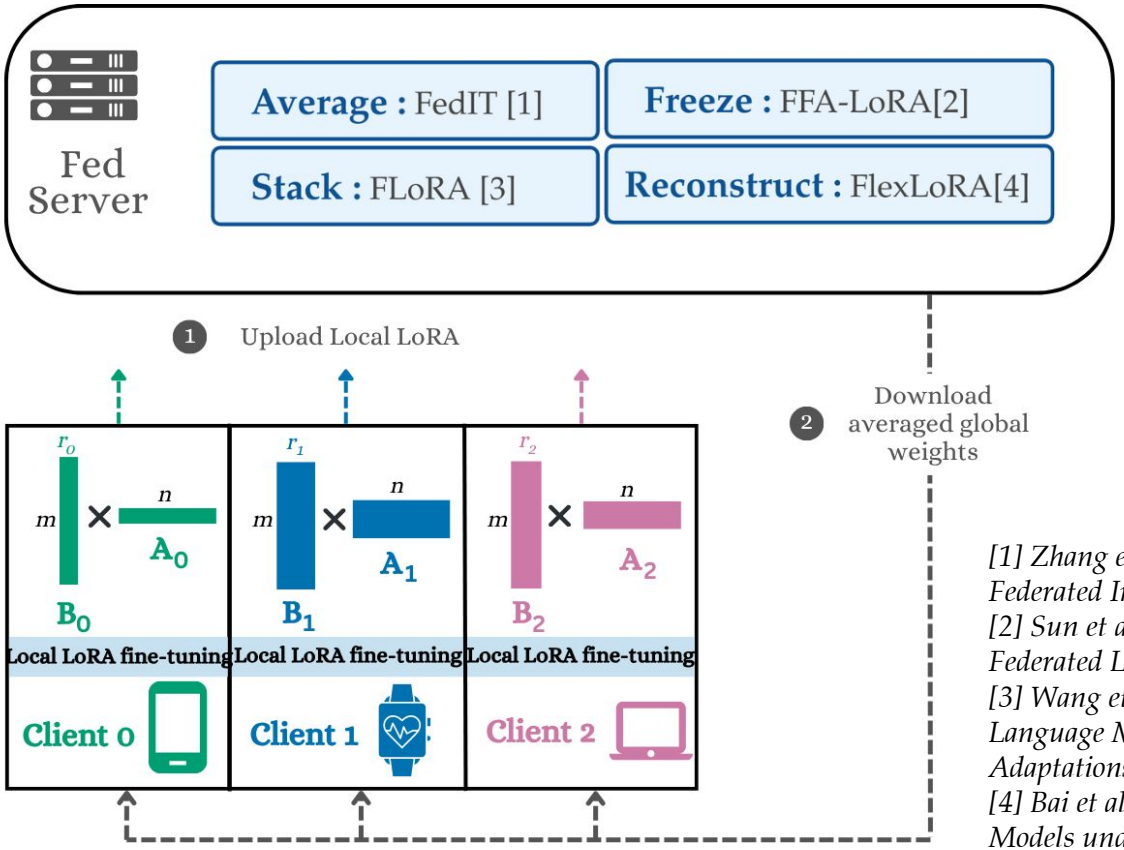
THE UNIVERSITY OF ARIZONA

Background: Federated LoRA Fine-Tuning



THE UNIVERSITY OF ARIZONA

Related Work



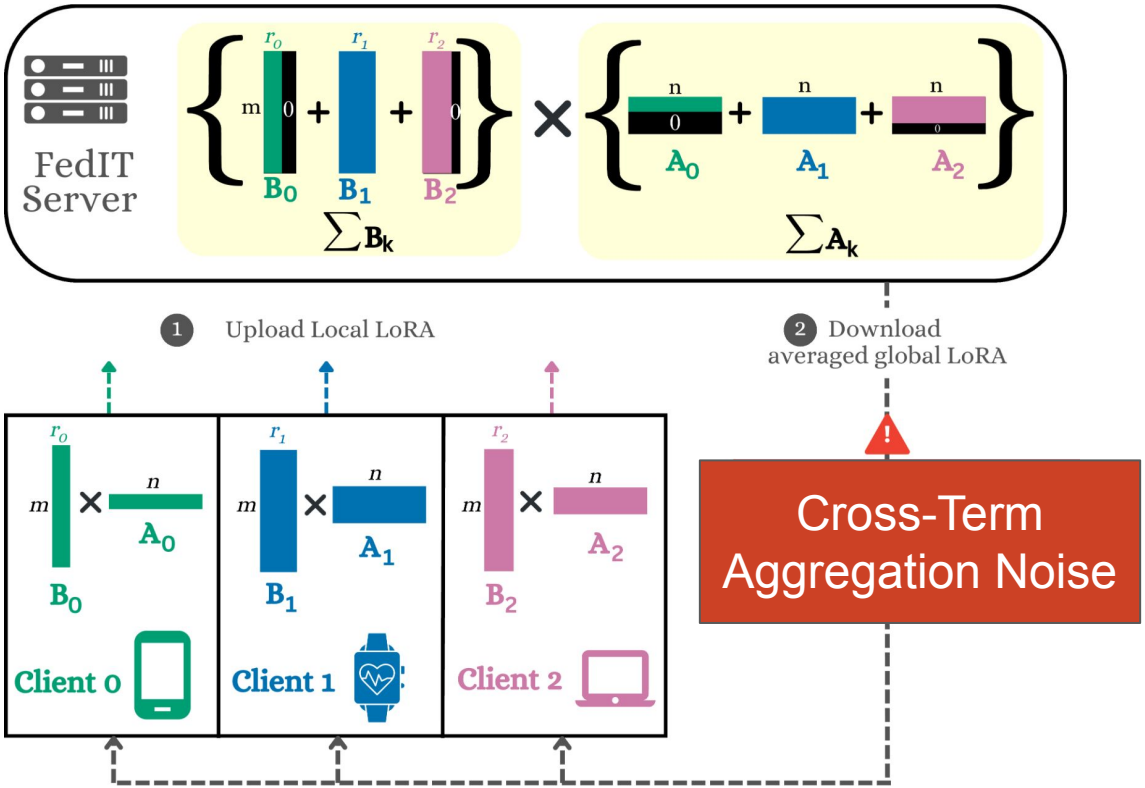
Categorize based on **server aggregation of Local LoRA adapters**

[1] Zhang et al., "Towards Building the FederatedGPT: Federated Instruction Tuning," ICASSP 2024.
[2] Sun et al., "Improving LoRA in Privacy-Preserving Federated Learning," ICLR 2024.
[3] Wang et al., "FLoRA: Federated Fine-Tuning Large Language Models with Heterogeneous Low-Rank Adaptations," NeurIPS 2024.
[4] Bai et al., "Federated Fine-Tuning of Large Language Models under Heterogeneous Tasks and Client Resources," NeurIPS 2024.

THE UNIVERSITY OF ARIZONA

Related Work: FedIT *(Average)*

Standard FedAvg applied directly to LoRA adapters

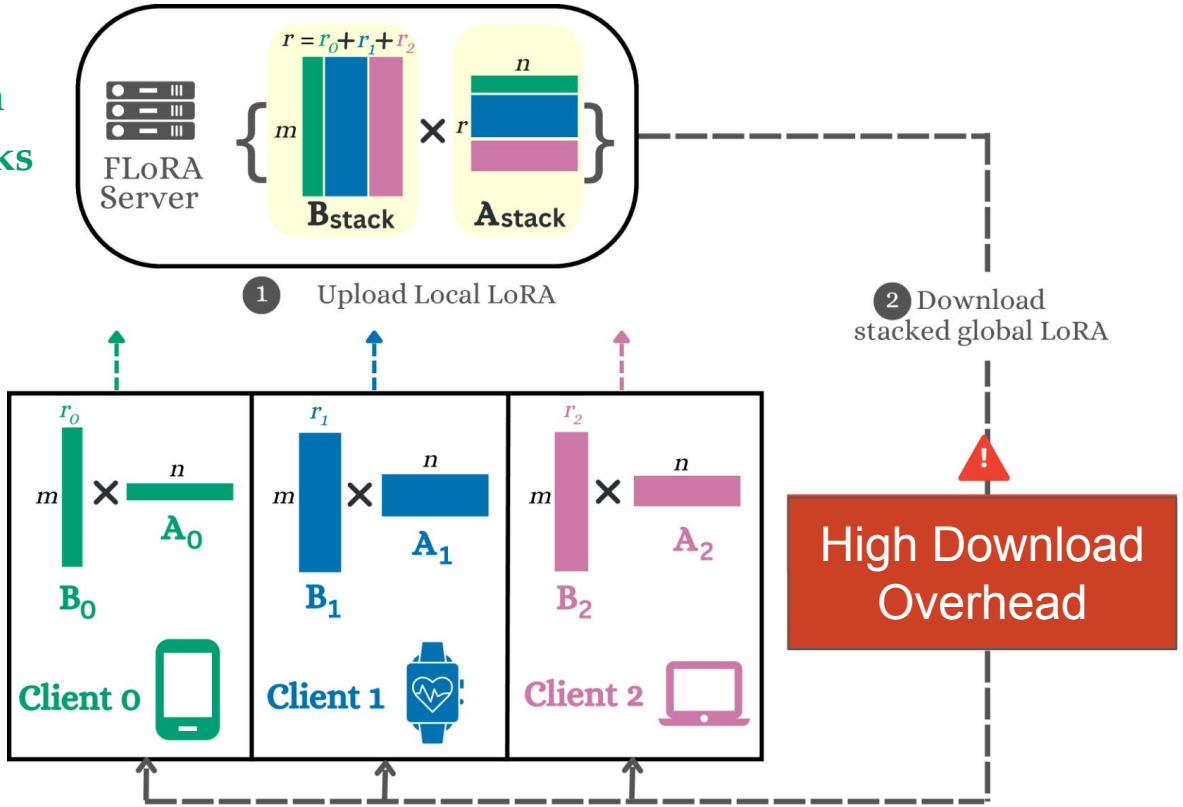


Zhang et al., "Towards Building the FederatedGPT: Federated Instruction Tuning," ICASSP 2024.

THE UNIVERSITY OF ARIZONA

Related Work: FLoRA *(Stack)*

Mathematically exact aggregation
Supports heterogeneous client ranks



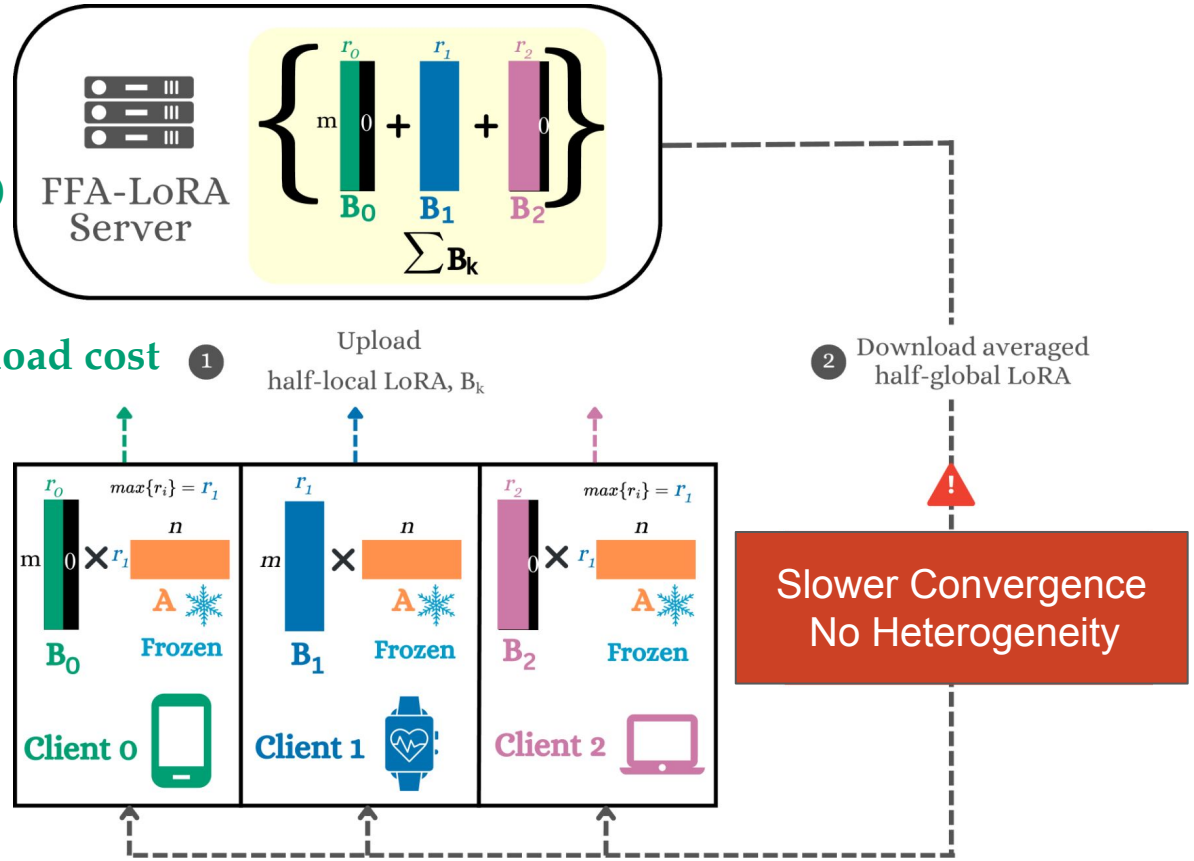
Wang et al., "FLoRA: Federated Fine-Tuning Large Language Models with Heterogeneous Low-Rank Adaptations," NeurIPS 2024.

THE UNIVERSITY OF ARIZONA

Related Work: FFA-LoRA *(Freeze)*

Mathematically exact aggregation (freezing A_k eliminates cross-terms)

Half the upload/download cost



Sun et al., "Improving LoRA in Privacy-Preserving Federated Learning," ICLR 2024.

THE UNIVERSITY OF ARIZONA

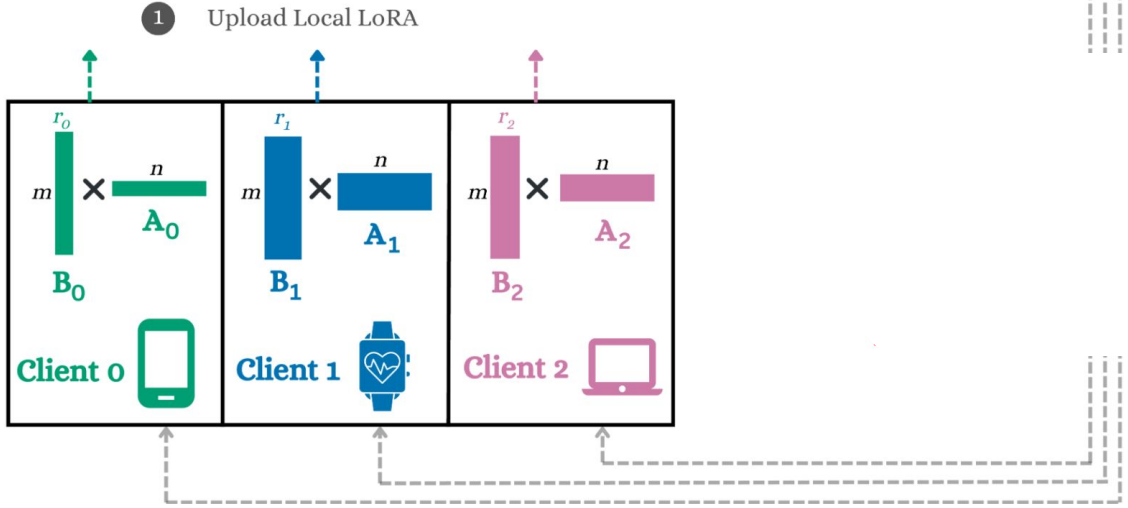
Related Work: FlexLoRA *(Reconstruct)*

Large Server Computational Cost

Exact aggregation

Heterogeneous ranks

$$\left\{ B_0 A_0 + B_1 A_1 + B_2 A_2 \right\} = \Delta W = \sum B_i A_i$$

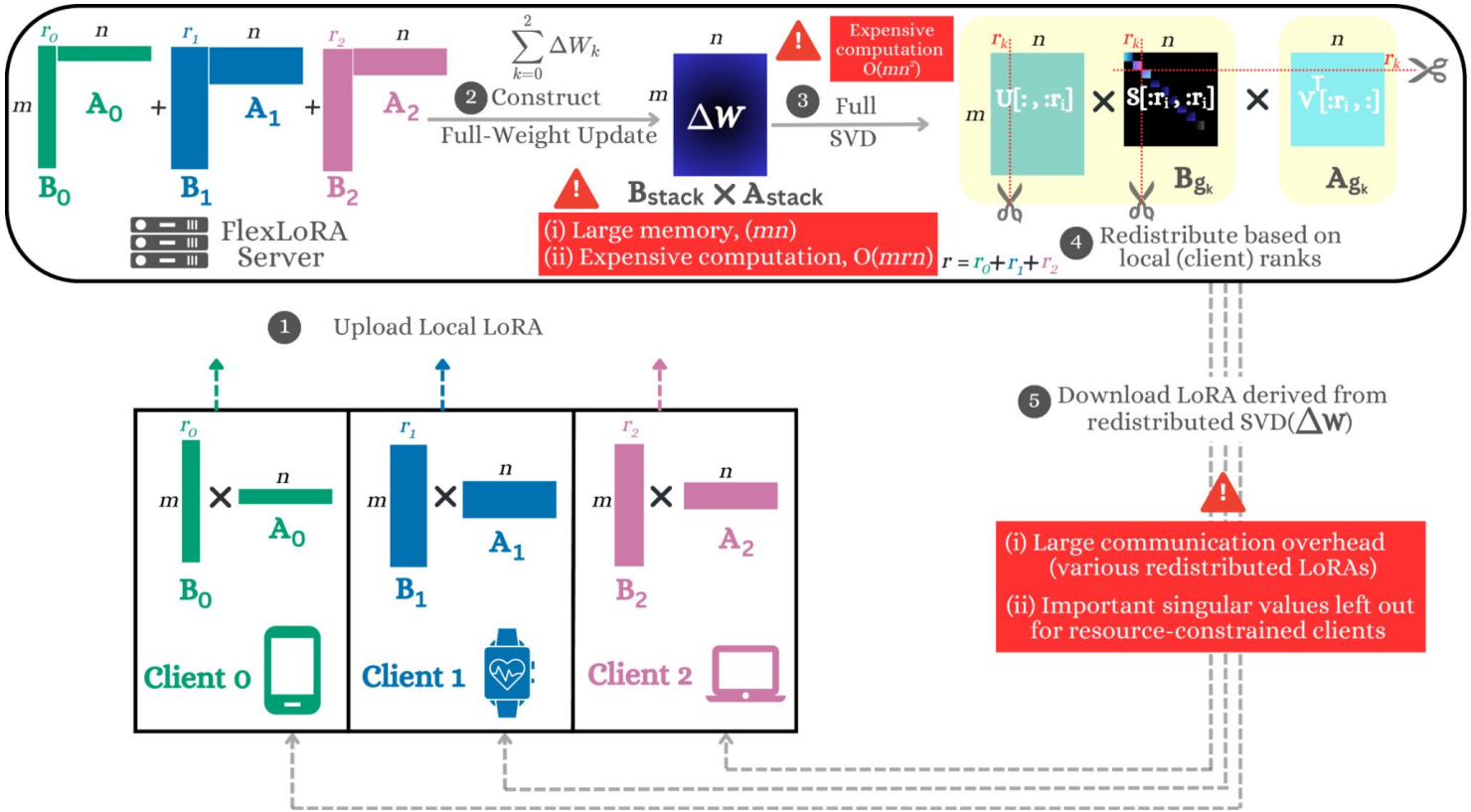


Bai et al., "Federated Fine-Tuning of Large Language Models under Heterogeneous Tasks and Client Resources," NeurIPS 2024.

THE UNIVERSITY OF ARIZONA

Related Work: FlexLoRA *(Reconstruct)*

Large Server Computational Cost



Bai et al., "Federated Fine-Tuning of Large Language Models under Heterogeneous Tasks and Client Resources," NeurIPS 2024.

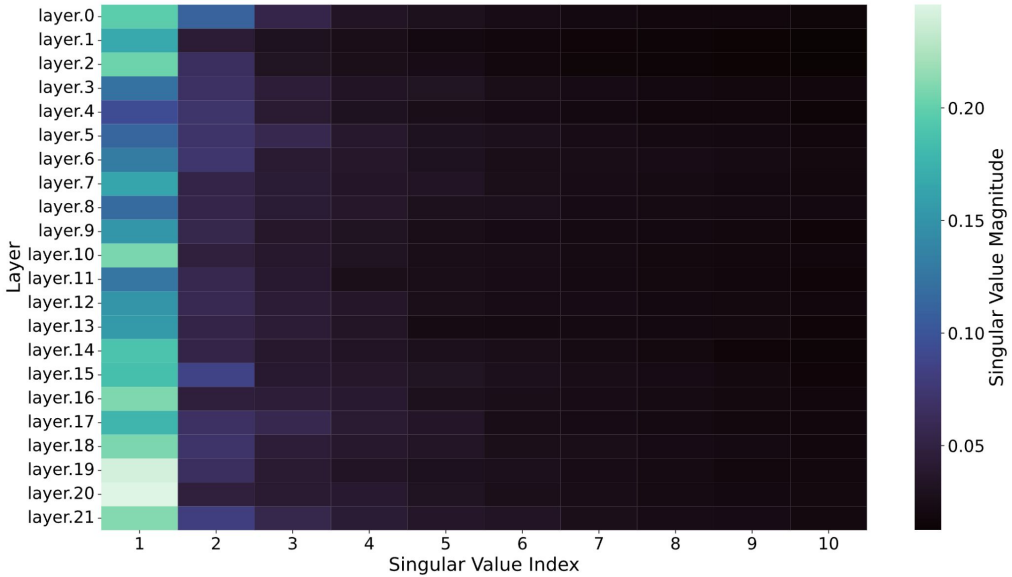
THE UNIVERSITY OF ARIZONA

Challenges in Existing Methods

Method	Aggregation	Heterogeneity	Key Challenge
FedIT *(Average)*	✗ Noisy	✗ Homo only	Cross-term Noise
FFA-LoRA *(Freeze)*	✓ Exact	✗ Homo only	Slower Convergence
FLoRA *(Stack)*	✓ Exact	✓ Native	High Download Overhead
FlexLoRA *(Reconstruct)*	✓ Exact	✓ Native	Large Server Computational Cost (ΔW)

THE UNIVERSITY OF ARIZONA

What is the intrinsic dimensionality of the Global Adapters?



Singular value spectrum (q_proj), TinyLlama (Wizard dataset)

Our Motivating Observation

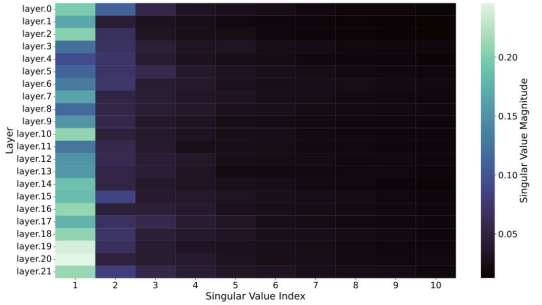
Singular values of aggregated ΔW decay rapidly

Only 6–10 components are significant (max client rank = 64)

THE UNIVERSITY OF ARIZONA

Motivating Research Question

Method	Aggregation	Heterogeneity	Key Challenge
FedIT *(Average)*	✗ Noisy	✗ Homo only	Cross-term Noise
FFA-LoRA *(Freeze)*	✓ Exact	✗ Homo only	Slower Convergence
FLoRA *(Stack)*	✓ Exact	✓ Native	High Download Overhead
FlexLoRA *(Reconstruct)*	✓ Exact	✓ Native	Large Server Computational Cost (ΔW)



Singular value spectrum (q_proj), TinyLlama (Wizard dataset)

Our Motivating Observation

Singular values of aggregated ΔW decay rapidly

Only 6–10 components are significant; max rank = 64

*How to identify the **most informative components** of global aggregated update (**ΔW**) **without reconstructing it** for faster aggregation & efficient download ?*

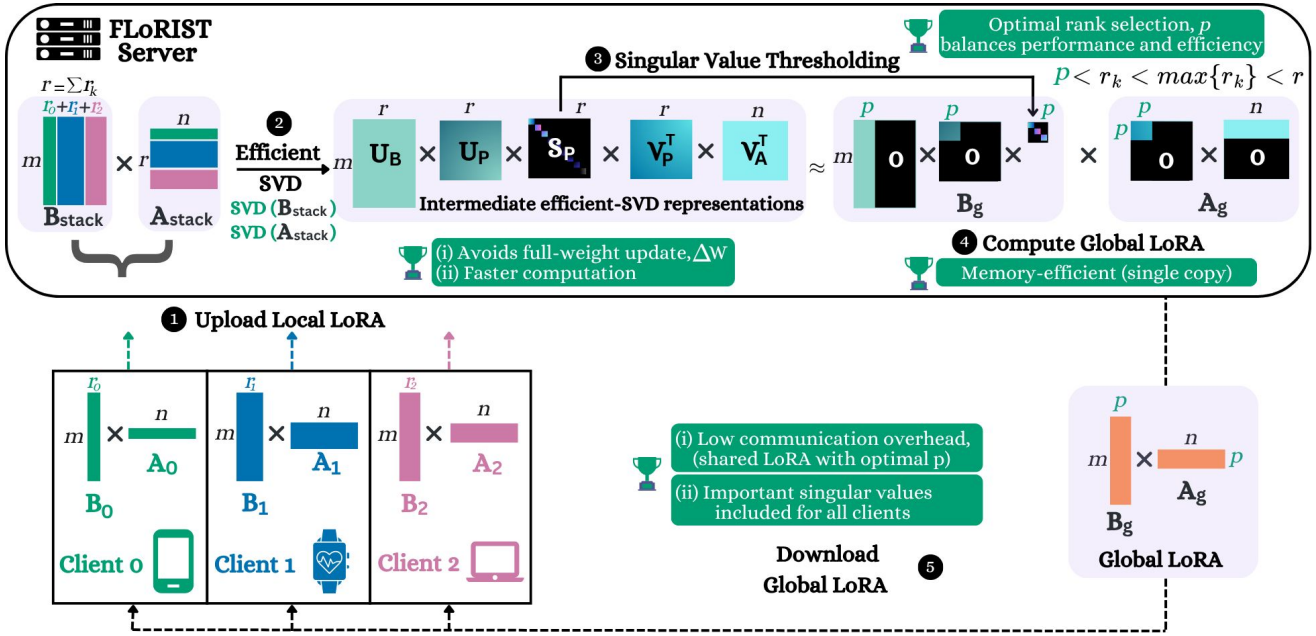
THE UNIVERSITY OF ARIZONA

Proposed Method

FLoRIST: Singular Value Thresholding for Efficient and Accurate Federated Fine-Tuning of Large Language Models

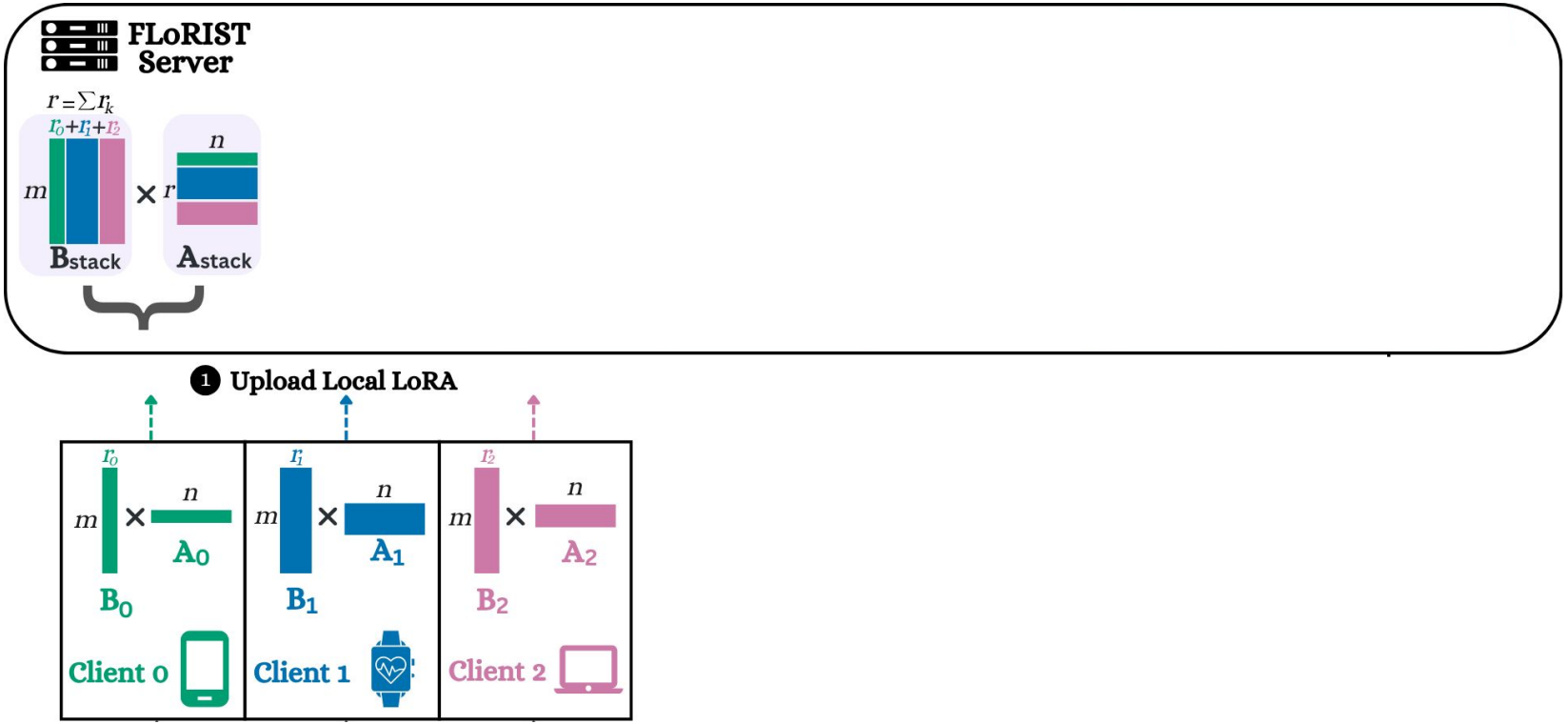
Avoids ΔW reconstruction;
Lowers server compute cost for faster aggregation

Identifies most informative components;
Reduces download overhead for high communication efficiency



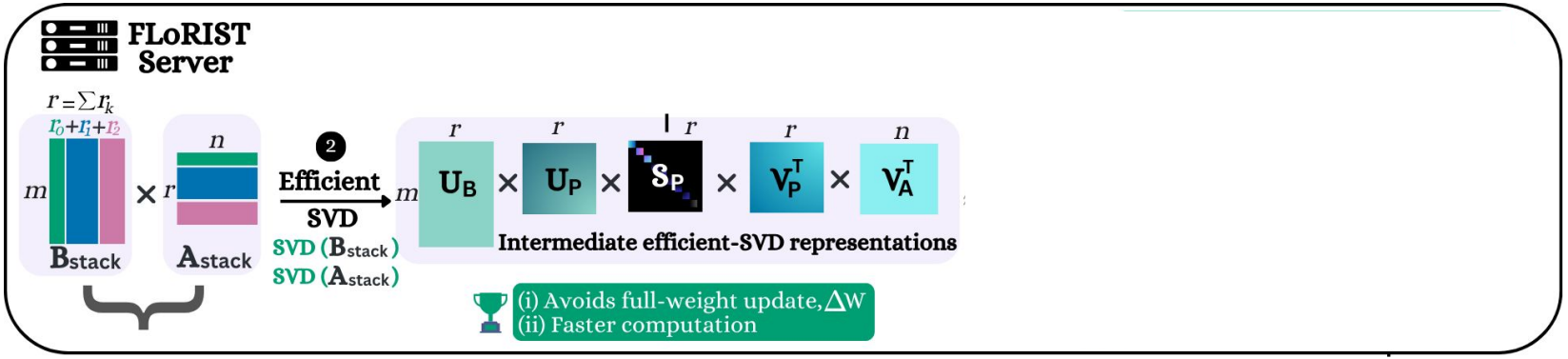
THE UNIVERSITY OF ARIZONA

Proposed FLoRIST



THE UNIVERSITY OF ARIZONA

Proposed FLoRIST



$$SVD(B_{\text{stack}}) = U_B S_B V_B^T$$

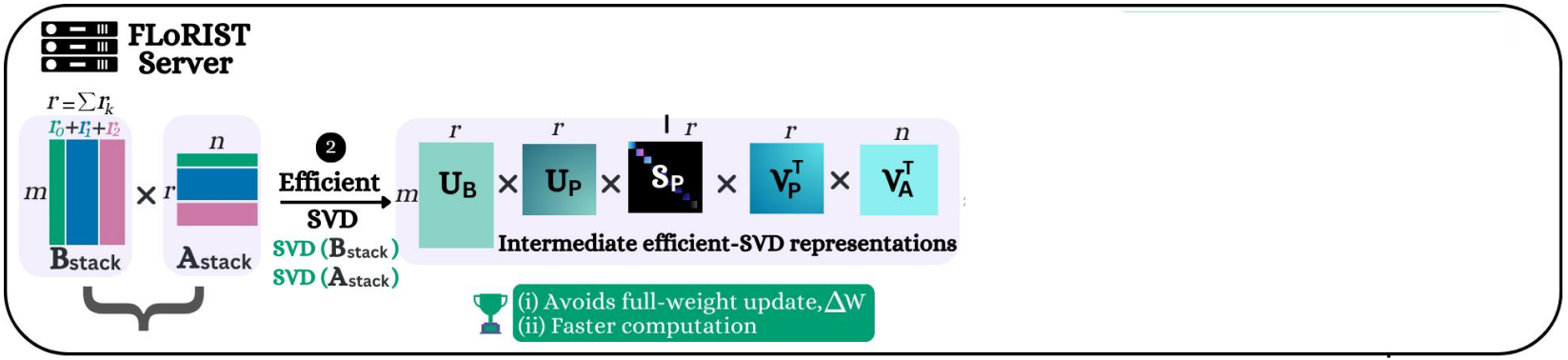
$$SVD(A_{\text{stack}}) = U_A S_A V_A^T$$

$$\Delta W = U_B \underbrace{S_B V_B^T U_A S_A}_{P \in \mathbb{R}^{r \times r}} V_A^T$$

P captures the **cross-adapter interaction** between Local LoRA while maintaining low dimension, r < {m, n}

THE UNIVERSITY OF ARIZONA

Proposed FLoRIST



$$SVD(B_{\text{stack}}) = U_B S_B V_B^T$$

$$SVD(A_{\text{stack}}) = U_A S_A V_A^T$$

$$\Delta W = U_B \underbrace{S_B V_B^T U_A S_A}_{P} V_A^T$$

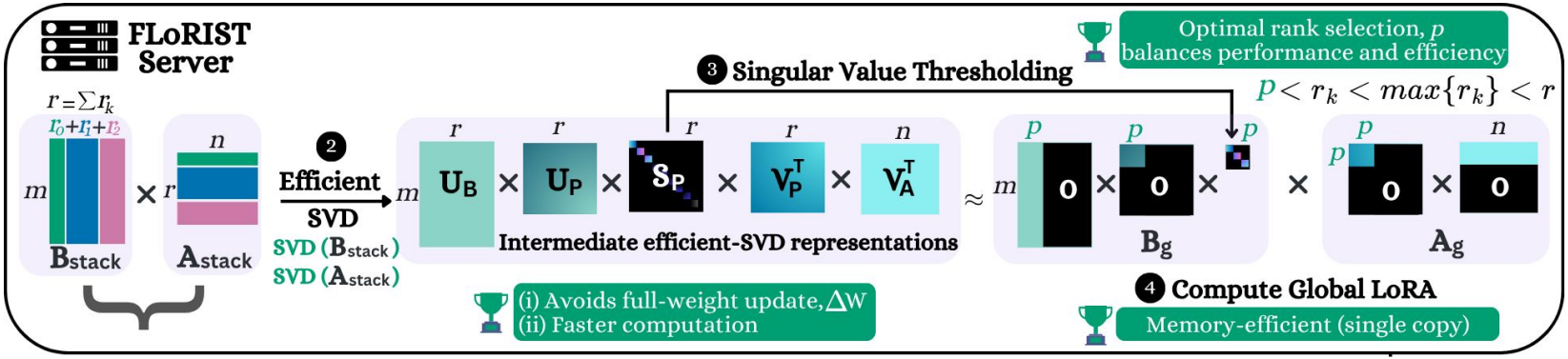
$$SVD(P) = U_P S_P V_P^T$$

$$\Delta W = (U_B U_P) S_P (V_P^T V_A^T)$$

$$SVD(\Delta W) = U S V^T$$

Singular values (P) == Singular values (ΔW) without global update reconstruction

Proposed FLoRIST



Energy-based Rank Selection

We select the smallest rank p that ensures at least a τ -fraction of the total variance is preserved,

$$\frac{\sum_{i=1}^{p} \sigma_i^2}{\sum_{i=1}^{r^*} \sigma_i^2} \geq \tau$$

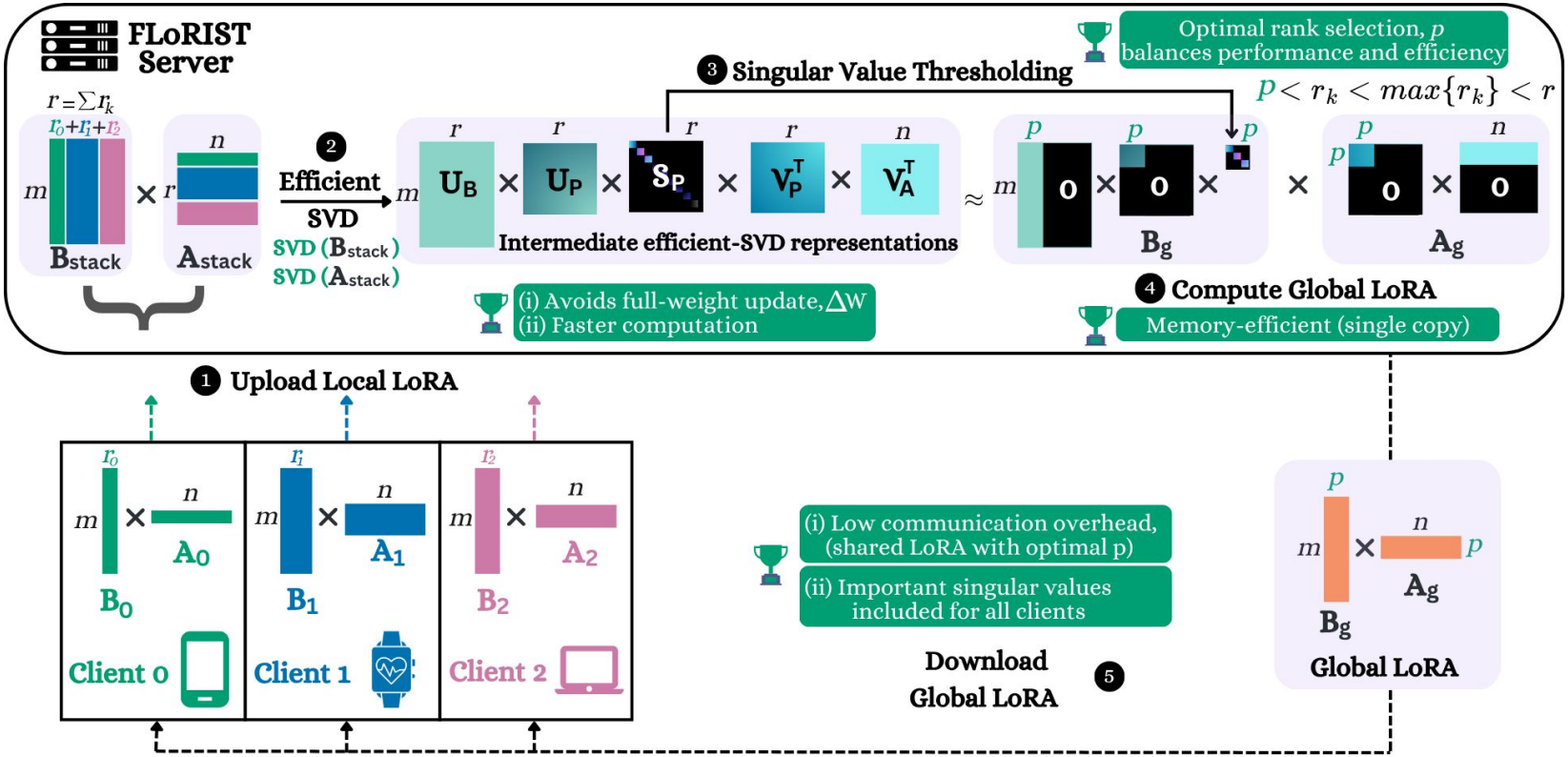
Approximation Error Bound

By the Eckart–Young–Mirsky theorem, the approximation error is bounded by the tail energy of the discarded singular values:

$$\|\Delta W - B_g A_g\|_F \leq \left(\sum_{i=p+1}^{r^*} \sigma_i^2 \right)^{1/2}.$$

THE UNIVERSITY OF ARIZONA

Proposed FLoRIST



THE UNIVERSITY OF ARIZONA

Experimental Setup

Models & Datasets

LLMs: TinyLlama-1.1B, LLaMA-3.2-1B
Datasets: Dolly-15k, Alpaca-52k, WizardLM-70k
Eval: MMLU (1,444 questions, 57 subjects)

FL Setup

Hardware: NVIDIA H100 and A100 GPUs
Clients: 100 clients, 10 sampled/round, 75 rounds, non-IID (Dirichlet α=0.5)

Homogeneous ranks: 100 clients = 100 x **r16**
Heterogeneous ranks:
100 clients = {40×**r4**, 20×**r8**, 20×r**16**, 10×**r32**, 10×**r64**}

FLoRIST: Energy Threshold Variants

FLoRIST [τ*]

- Diagnostic upper bound on the accuracy–efficiency
- Smallest τ that matches or beats all baseline accuracy
- Binary search over [0.80, 0.99]

FLoRIST [τ=0.9]

- Single fixed threshold for practical deployment

Performance Metrics for baseline comparison

Accuracy (MMLU)
Communication Efficiency (1/download parameters)
Convergence
Computational FLOPs (Server)

THE UNIVERSITY OF ARIZONA

Results: Accuracy vs Efficiency

MMLU performance across models, client configurations and 3 datasets

Model	Client	Method	Dolly		Alpaca		Wizard	
			Acc. (%)	Eff. ($\times 10^{-4}$)	Acc. (%)	Eff. ($\times 10^{-4}$)	Acc. (%)	Eff. ($\times 10^{-4}$)
TinyLlama	Homo	FedIT	27.46	14.20	28.35	14.20	36.61	14.20
		FLoRA	28.99	1.78	28.99	1.78	34.20	1.78
		FlexLoRA	28.06	14.20	29.22	14.20	39.75	14.20
		FFA-LoRA	27.79	28.40	31.58	28.40	36.01	28.40
		FLoRIST [τ^*]	29.16	**53.54**	**32.26**	35.08	**40.95**	54.38
		FLoRIST [τ=0.9]	**30.94**	23.48	31.68	**60.06**	38.92	**63.09**
	Heter	FedIT (zero-pad)	25.73	3.55	31.04	3.55	**44.19**	3.55
		FLoRA	27.20	0.50	28.58	0.50	33.74	0.50
		FlexLoRA	28.49	11.96	29.61	11.96	36.39	11.96
		FFA-LoRA	18.54	7.10	23.19	7.10	23.75	7.10
		FLoRIST [τ^*]	**28.87**	**37.87**	**31.33**	**45.09**	38.20	13.60
		FLoRIST [τ=0.9]	27.69	34.93	29.69	34.90	41.51	**36.16**
Llama-3.2-1B	Homo	FedIT	25.00	19.50	29.44	19.50	**30.48**	19.50
		FLoRA	22.48	2.44	29.16	2.44	28.57	2.44
		FlexLoRA	27.39	19.50	29.24	19.50	30.03	19.50
		FFA-LoRA	26.15	39.06	29.18	**39.06**	28.40	39.06
		FLoRIST [τ^*]	28.28	**51.71**	**30.46**	33.30	29.31	**51.19**
		FLoRIST [τ=0.9]	**29.48**	33.64	29.73	33.42	29.62	**51.19**
	Heter	FedIT (zero-pad)	21.41	4.88	28.37	4.88	28.42	4.88
		FLoRA	23.85	2.06	30.15	2.06	27.73	2.06
		FlexLoRA	**26.74**	16.44	29.95	16.44	29.13	16.44
		FFA-LoRA	22.45	9.77	22.68	9.77	28.78	9.77
		FLoRIST [τ^*]	24.01	46.35	**30.53**	45.22	**29.79**	47.82
		FLoRIST [τ=0.9]	24.10	**49.38**	30.24	**48.24**	29.45	**50.22**

FLoRIST achieves highest Accuracy-Efficiency tradeoff for both homogeneous and heterogeneous ranks

With [τ=0.9], accuracy stays within ±1% of τ*

THE UNIVERSITY OF ARIZONA

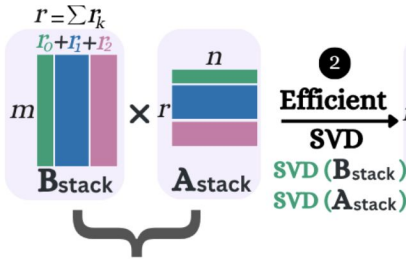
Results: Server Computation FLOPs

METHOD	SERVER FLOPS
FEDIT	4.76M
FFA-LORA	0.52M
FLORA	0B
FLEXLORA	3516.01M
FLORIST (OURS)	466.95M

TinyLlama-Alpaca-homo

FLoRIST: 467M FLOPs
7.5× faster than FlexLoRA (3516 M)

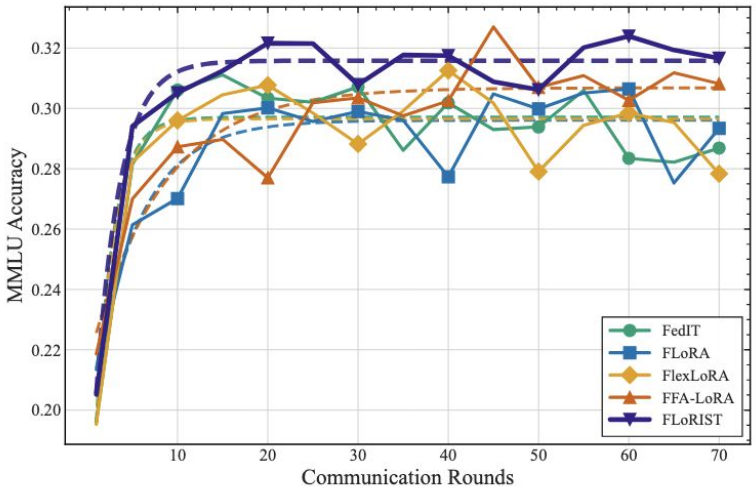
Avoids reconstructing global weight
by using **Efficient SVD** (operate in compact r×r space)



THE UNIVERSITY OF ARIZONA

Results: Convergence Comparison

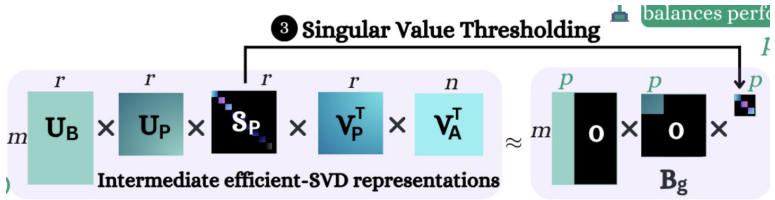
MMLU accuracy over rounds for TinyLlama-Alpaca-homo



FLoRIST:

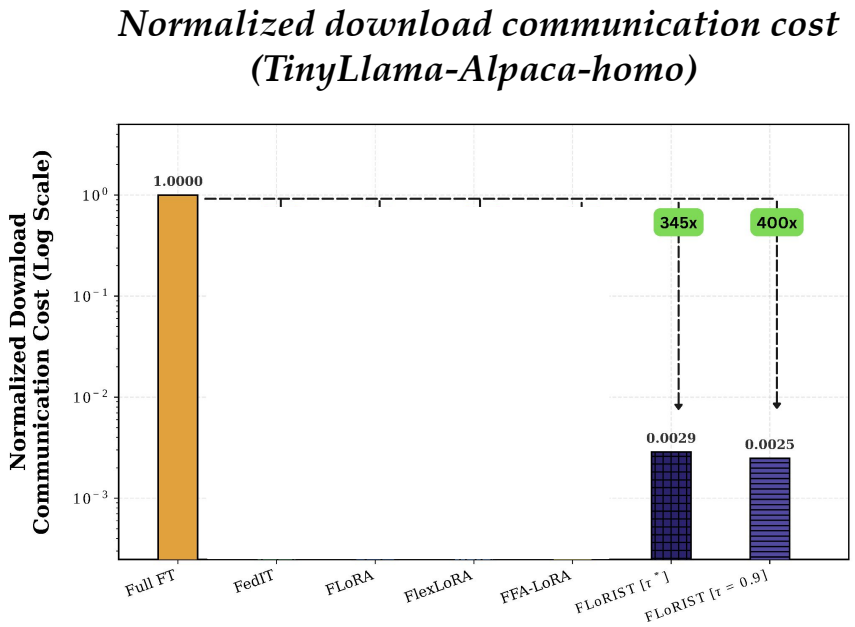
- **Fastest Convergence**
- **Highest MMLU accuracy (Generalizes better)**

Singular Value Thresholding: Higher signal to noise ratio (identifying most critical components of the aggregated global update ΔW)



THE UNIVERSITY OF ARIZONA

Results: Communication Efficiency

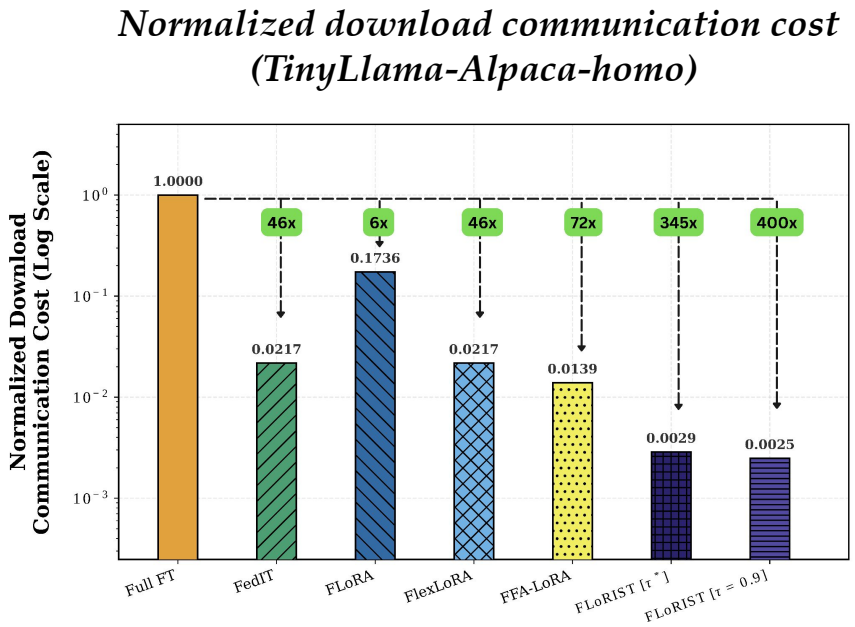


Download overhead: FLoRIST[τ =0.9]

- 400× reduction vs Full Fine Tuning
- 70× reduction vs FLoRA (stack)
- 5× reduction vs FFA-LoRA (freeze)

THE UNIVERSITY OF ARIZONA

Results: Communication Efficiency

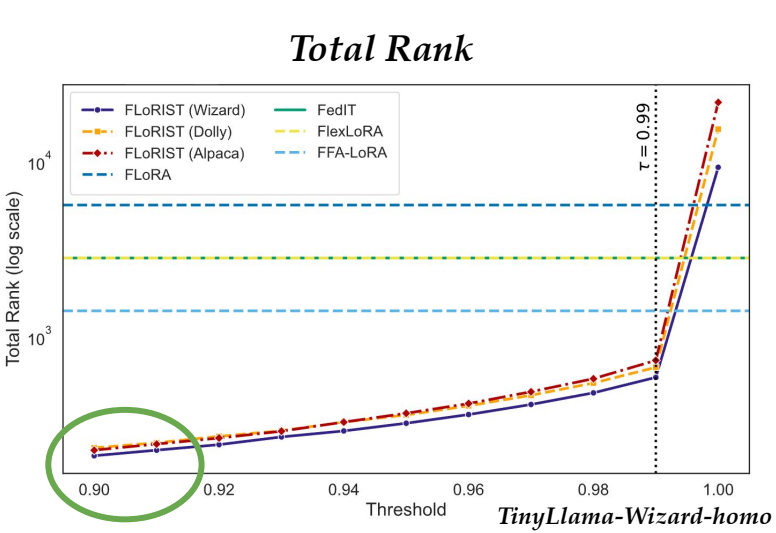


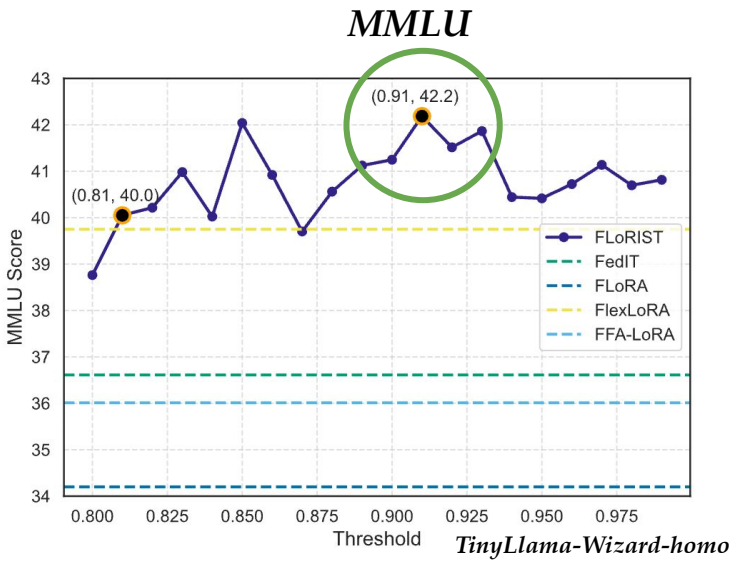
Download overhead: FLoRIST[τ =0.9]

- 400× reduction vs Full Fine Tuning
- 70× reduction vs FLoRA (stack)
- 5× reduction vs FFA-LoRA (freeze)

Ablation study: Energy Threshold

Identify optimal energy threshold (τ) for accuracy-efficiency tradeoff





For $\tau \leq 0.99$, FLoRIST achieves **lowest global rank** (higher efficiency)

$\tau = 0.9$ is a **robust practical threshold** with strong accuracy-efficiency tradeoff

THE UNIVERSITY OF ARIZONA

Conclusion

1. FLoRIST aggregates directly in **low-rank latent space**, **avoiding reconstruction** of full ΔW using **Efficient SVD**
2. **Singular value thresholding** identifies intrinsic dimensionality and selects optimal global rank
3. FLoRIST supports **heterogeneous client ranks** natively via a single shared global adapter
4. FLoRIST achieves **best accuracy-efficiency tradeoff** compared to baselines with $\tau = 0.9$ as practical threshold

METHOD	HETEROGENEITY	PERFORMANCE ↑	COMM. EFF. ↑	COMP. COST ↓
FEDIT	✗	LOW HIGH	LOW HIGH	LOW HIGH
FFA-LORA	✗	LOW HIGH	LOW HIGH	LOW HIGH
FLORA	✓	LOW HIGH	LOW HIGH	LOW HIGH
FLEXLORA	✓	LOW HIGH	LOW HIGH	LOW HIGH
FLORIST (OURS)	✓	LOW HIGH	LOW HIGH	LOW HIGH

Poster 29

Paper

Code

Artifacts Available V1.1

Artifacts Evaluated Functional V1.1

Results Reproduced V1.1

Acknowledgments

NSF ACCESS NVIDIA

THE UNIVERSITY OF ARIZONA

Thank you!