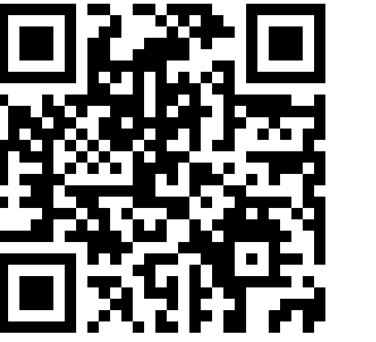




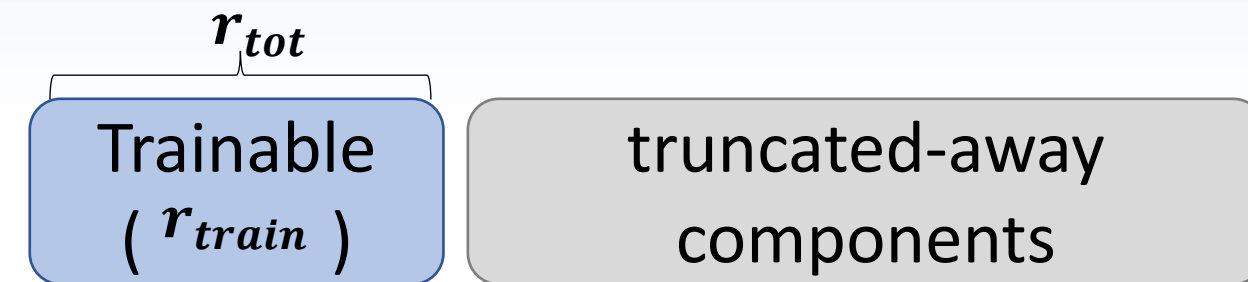
TL;DR

Main message: Communication rank and trainable rank should not be the same knob.

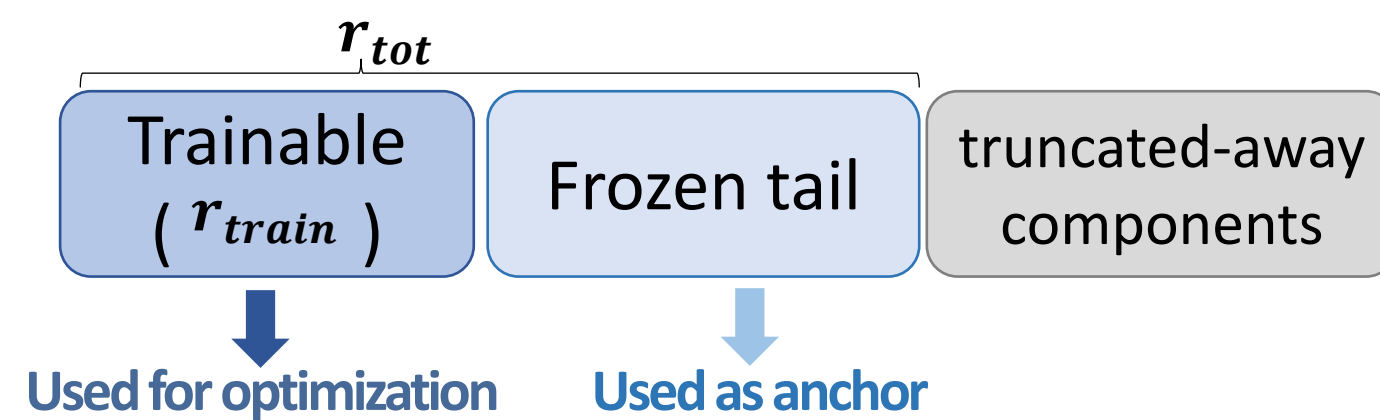
FedHera enables clients to download a richer global LoRA basis but trains only a budget-feasible prefix, utilizing a frozen tail to suppress heterogeneous truncation drift.

Problem: Coupling wastes bandwidth and increases drift

Coupled method: $r_{tot} = r_{train}$



FedHera: $r_{tot} > r_{train}$

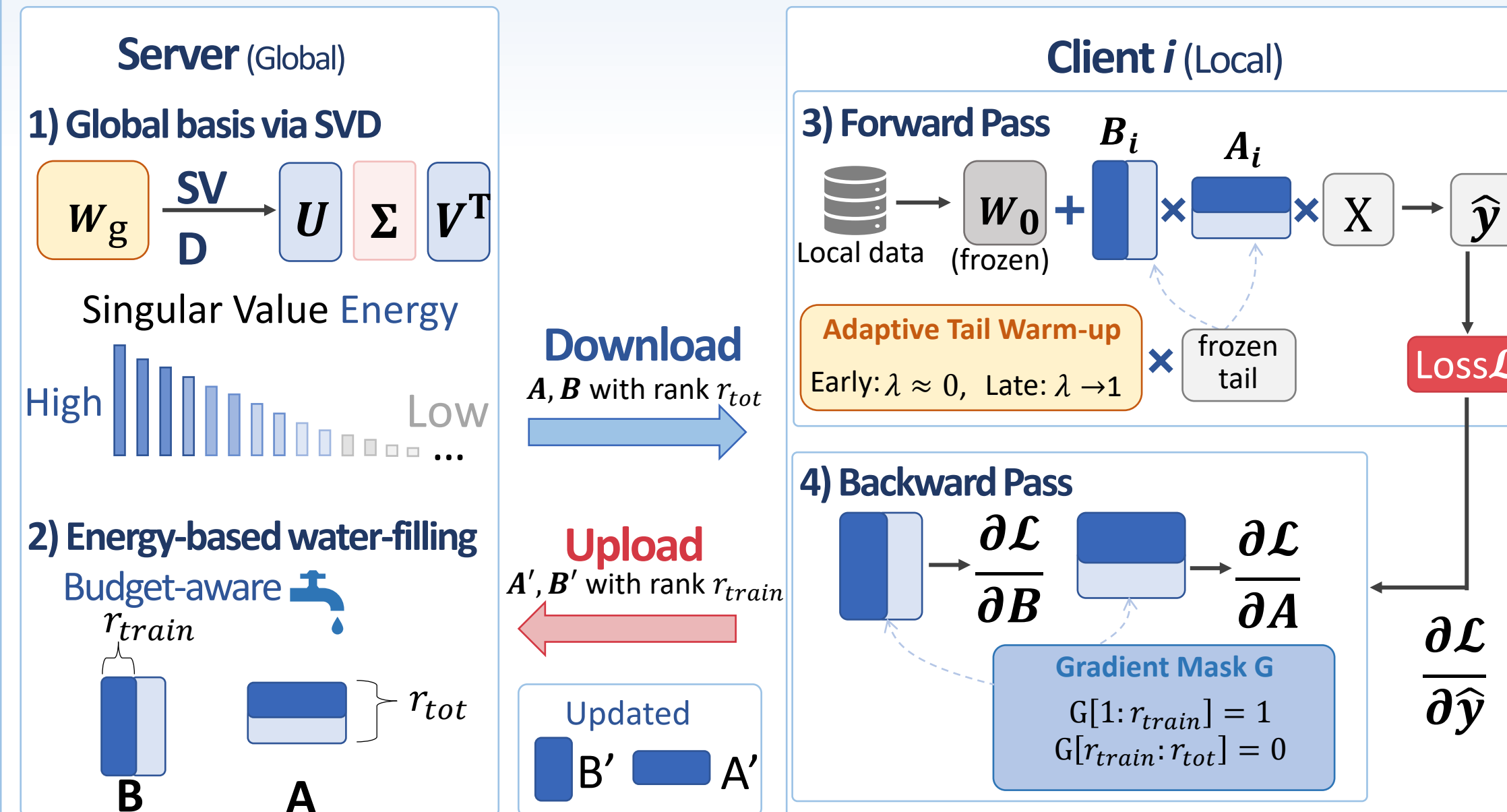


- 1 Clients may have enough bandwidth but limited VRAM
- 2 Coupling download rank to trainable rank discards useful global components
- 3 Aggressive truncation causes local-update drift

Decoupling turns unused bandwidth into useful global context.

- Key takeaways**
- Decouple communication from optimization.
 - Use frozen tail as global anchor.
 - Lower drift improves heterogeneous FFT.

FedHera: decoupled dual-rank federated fine-tuning



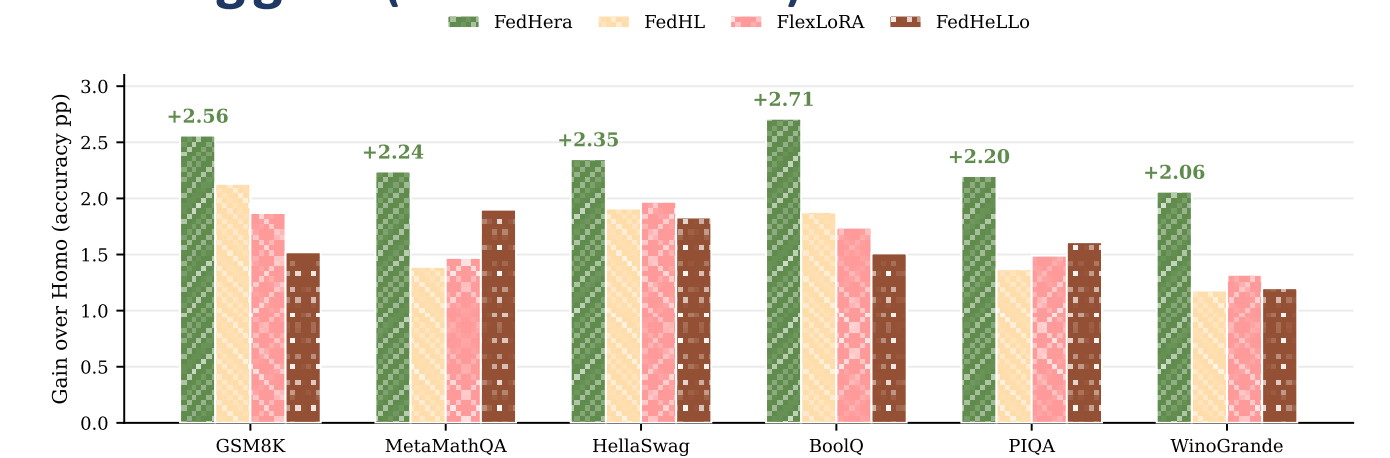
- 1 **Dual-rank allocation**: Decouple the download rank from the training rank, trading more bandwidth resources for more global information.
- 2 **Energy-based water-filling**: Under the resource budget, a high-energy-first rank allocation across layers is performed, minimizing the information loss caused by truncation.
- 3 **Prefix-gated training + ATW**: Train the prefix and use the frozen tail as a progressively activated forward anchor to reduce truncation-induced drift.

Decoupled adapter dimensions: $B_i = [B_i^{train}, B_i^{tail}]$, $A_i = \begin{bmatrix} A_i^{train} \\ A_i^{tail} \end{bmatrix}$, $r_i^{train} \leq r_i^{tot} = r_i^{train} + r_i^{tail}$

Local forward with frozen tail: $h = W_0 x + B_i^{train} A_i^{train} x + \lambda_i B_i^{tail} A_i^{tail} x$

Evidence: Better generalization, lower drift, higher utilization

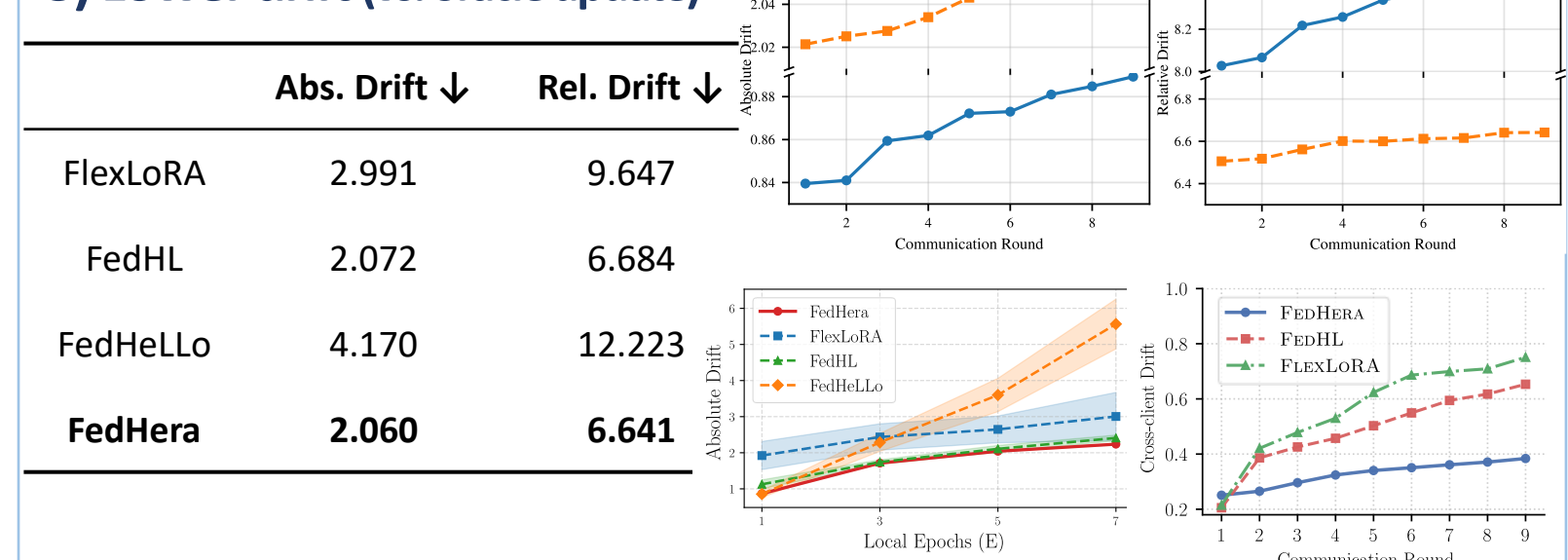
1) Reasoning gains (Uniform distribution)



2) Generative gains (Uniform distribution)

	Alpaca		E2E NLG				
	(ROUGEL)	(BLEU)	(ROUGEL)	(BLEU)	(MET.)	(NIST)	(CIDER)
FedHomoLoRA	0.609	0.175	0.684	0.430	0.677	4.490	0.176
FlexLoRA	0.701	0.227	0.864	0.642	0.872	5.722	0.345
FedHL	0.703	0.230	0.871	0.649	0.875	5.801	0.350
FedHeLlo	0.620	0.195	0.741	0.463	0.723	4.776	0.202
FedHera	0.708	0.236	0.879	0.652	0.883	5.90	0.447

3) Lower drift (vs. oracle update)



Contact:

Ke.Xiao@Glasgow.ac.uk

Supported in part by EU Horizon projects TERRA and ELLIE.

