Tensor-Parallelism with Partially Synchronized Activations

Itay Lamprecht, Asaf Karnieli, Yair Hanani, Niv Giladi and Daniel Soudry

The Thirty-Ninth Annual Conference on Neural Information Processing Systems

(NeurIPS 2025)

The research of DS was funded by the European Union (ERC, A-B-C-Deep, 101039436).

Background and Motivation

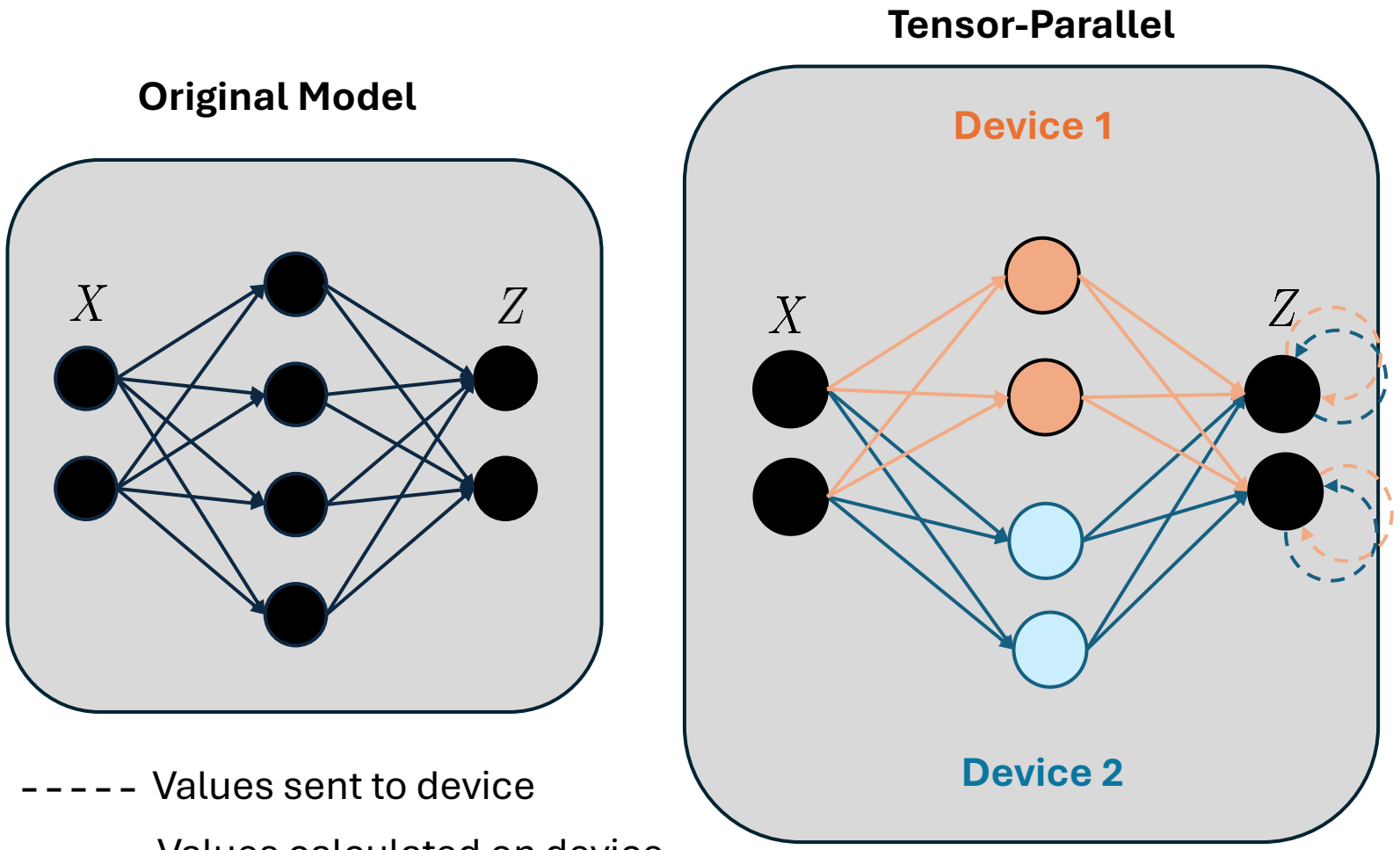
Background and Motivation

- Parallelism for Deep Neural Networks
  - Data Parallelism
  - Context Parallelism
  - Pipeline Parallelism
  - Tensor Parallelism

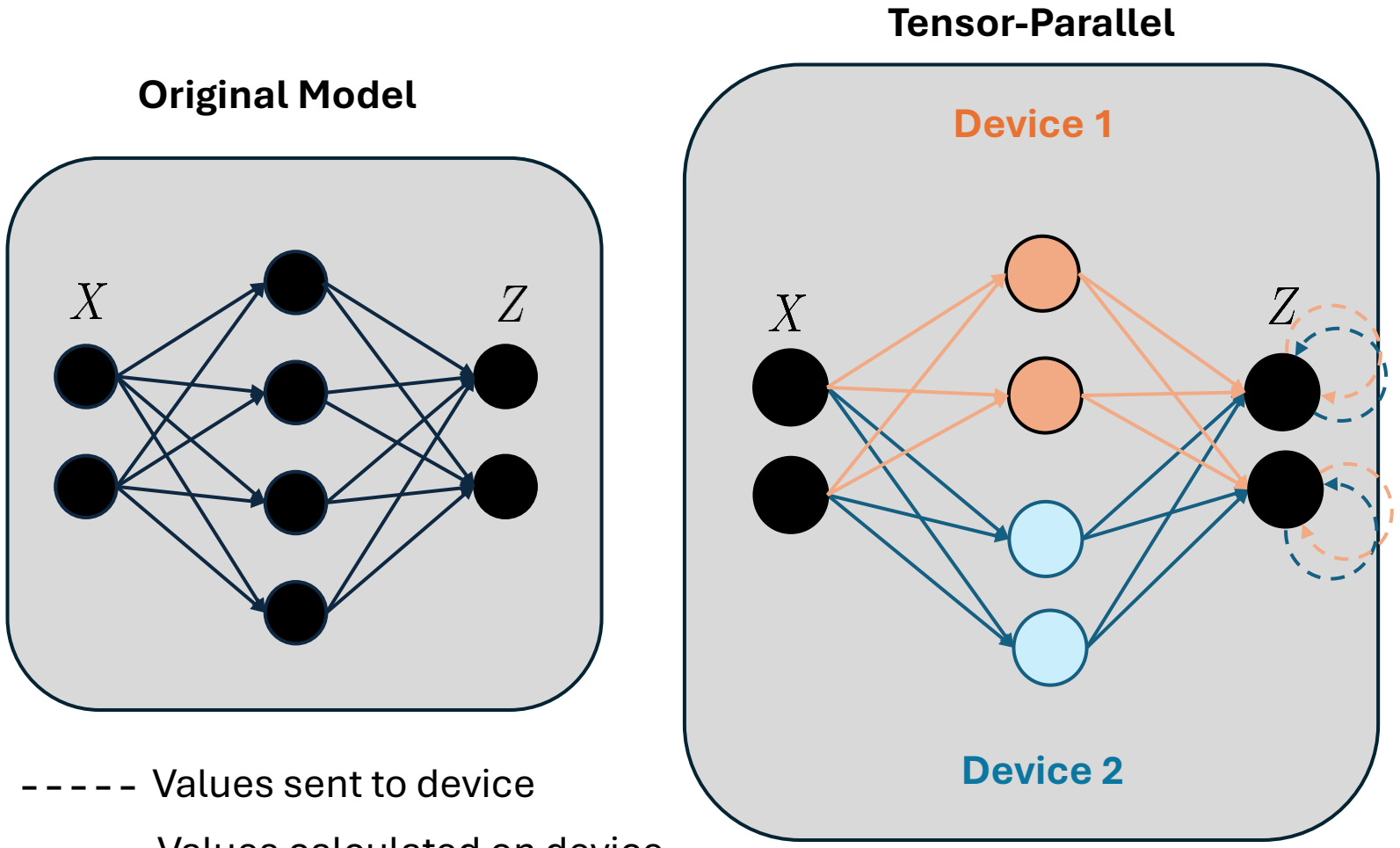
Background and Motivation

- Parallelism for Deep Neural Networks
  - Data Parallelism
  - Context Parallelism
  - Pipeline Parallelism
  - Tensor Parallelism

Tensor-Parallelism

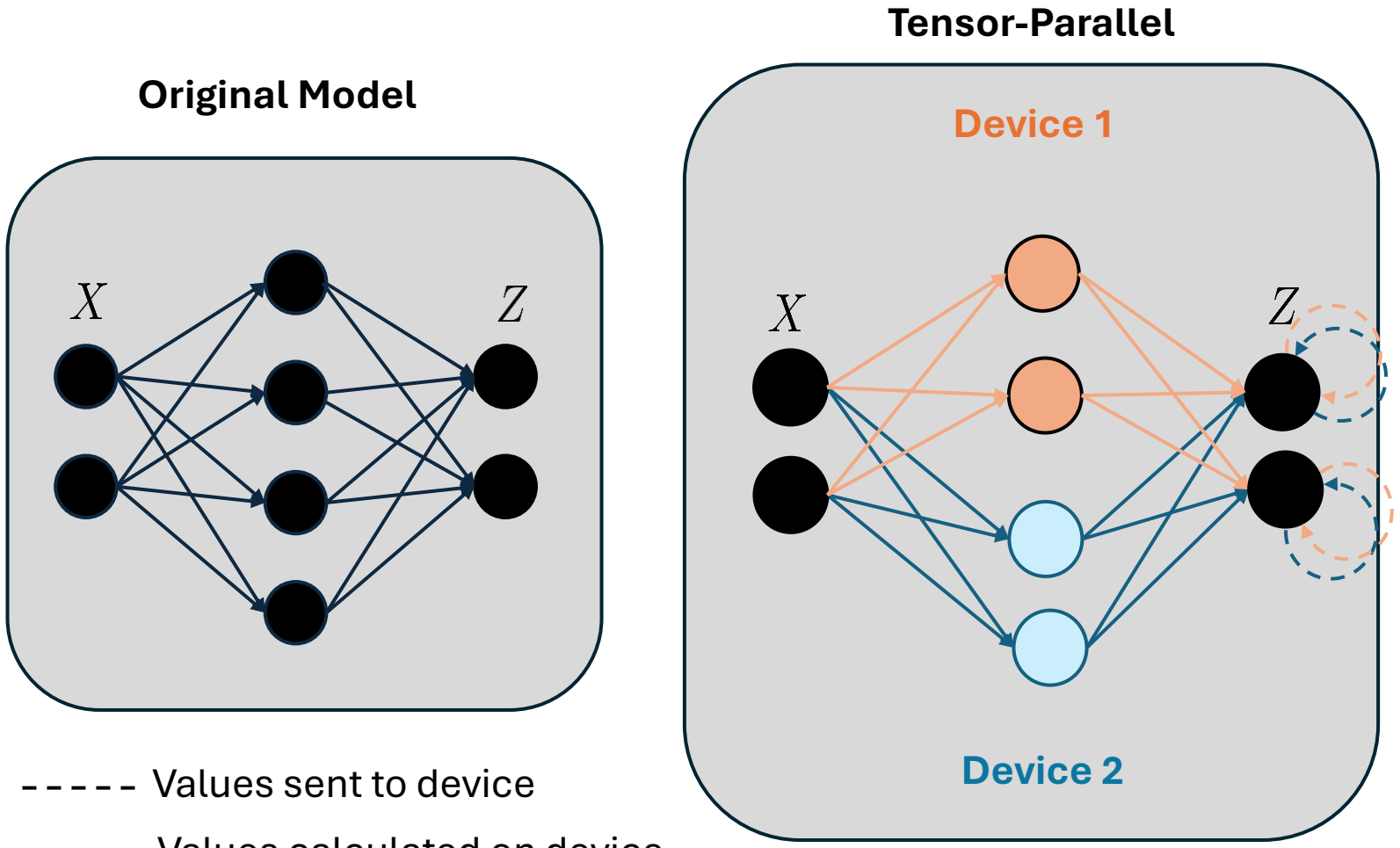


Tensor-Parallelism



Memory Efficient

Tensor-Parallelism



Memory Efficient

Costly Communication
On Critical Path

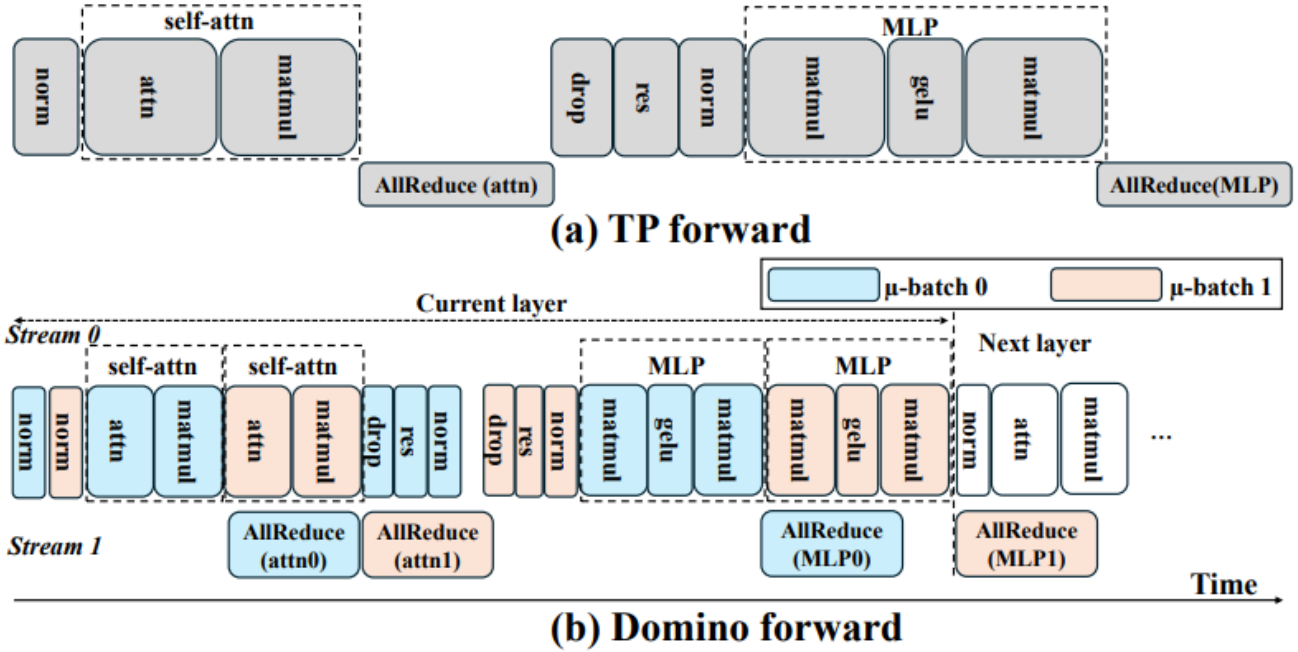
Recent works

- Pipelining Communication and Computation
- Asynchronous Training
- Compression methods
- Inference time methods

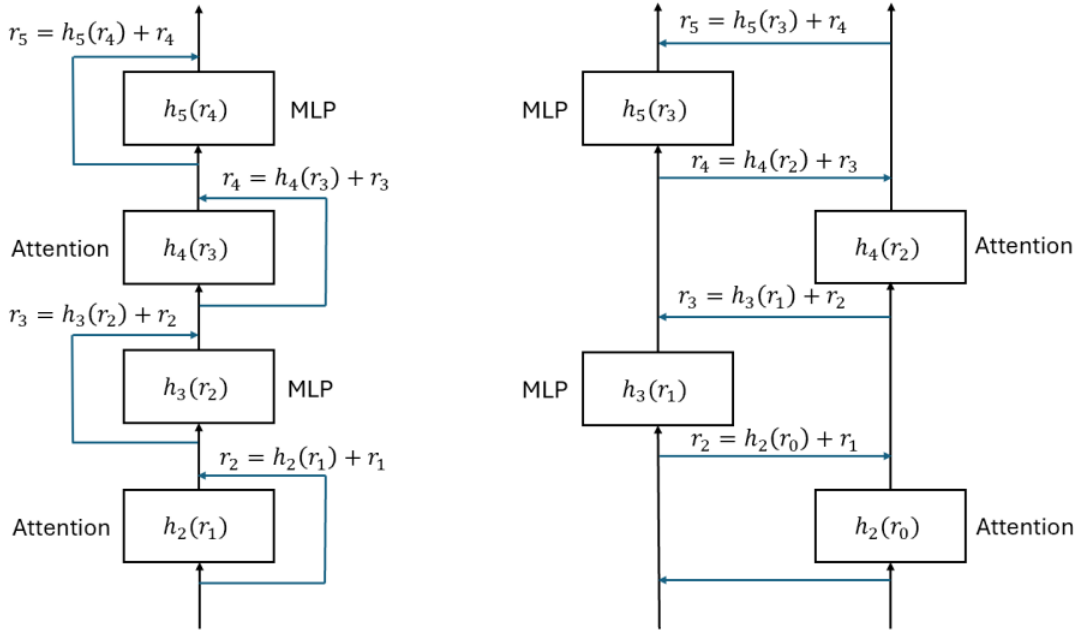
Recent works

- Pipelining Communication and Computation
- Asynchronous Training
- Compression methods
- Inference time methods

Pipelining Approaches



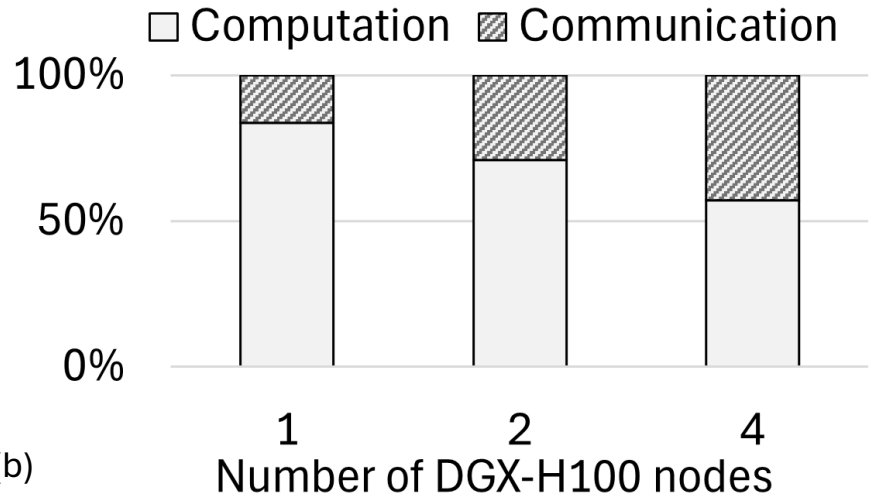
Domino: Eliminating communication in llm training via generic tensor slicing and overlapping Guanhua Wang, Chengming Zhang, Zheyu Shen, Ang Li, and Olatunji Ruwase. 2024.



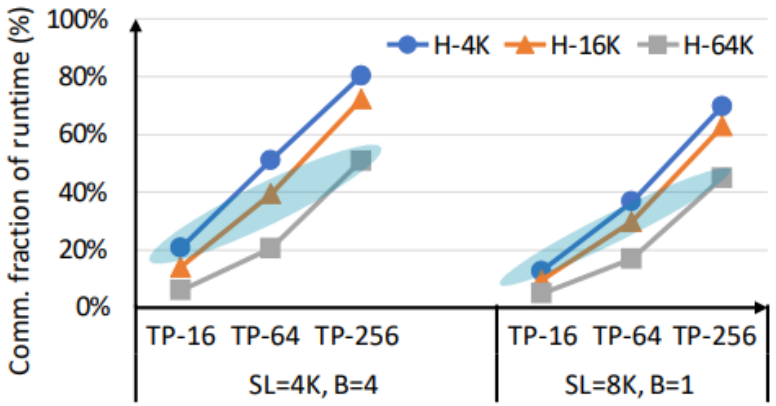
Ladder-residual: parallelism-aware architecture for accelerating large model inference with communication overlapping. Muru Zhang, Mayank Mishra, Zhongzhu Zhou, William Brandon, Jue Wang, Yoon Kim, Jonathan Ragan-Kelley, Shuaiwen Leon Song, Ben Athiwaratkun, and Tri Dao. 2025.

Pipelining Approaches

(a)



(b)

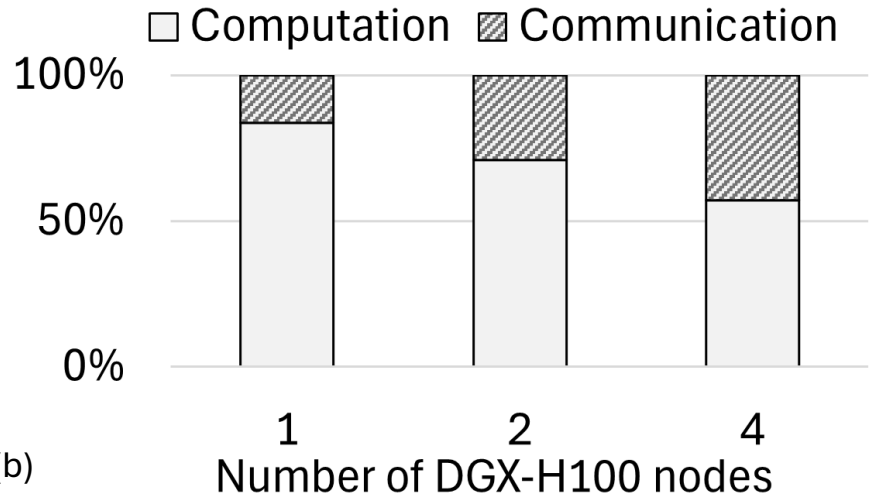


(a) **Domino: Eliminating communication in llm training via generic tensor slicing and overlapping** Guanhua Wang, Chengming Zhang, Zheyu Shen, Ang Li, and Olatunji Ruwase. 2024.

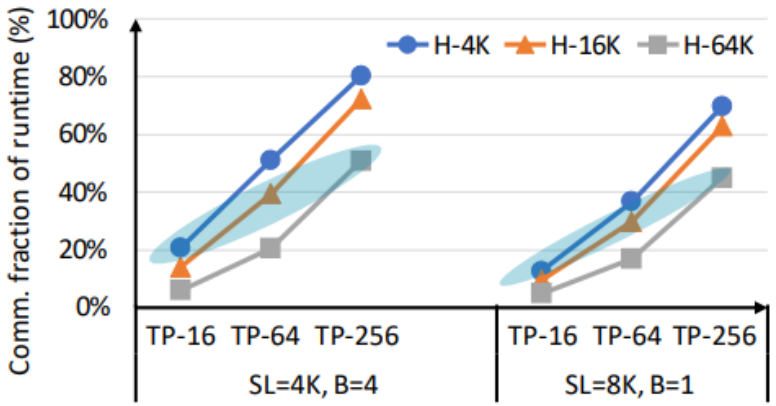
(b) **Computation vs. communication scaling for future transformers on future hardware.** Suchita Pati, Shaizeen Aga, Mahzabeen Islam, Nuwan Jayasena, and Matthew D. Sinclair. 2023

Pipelining Approaches

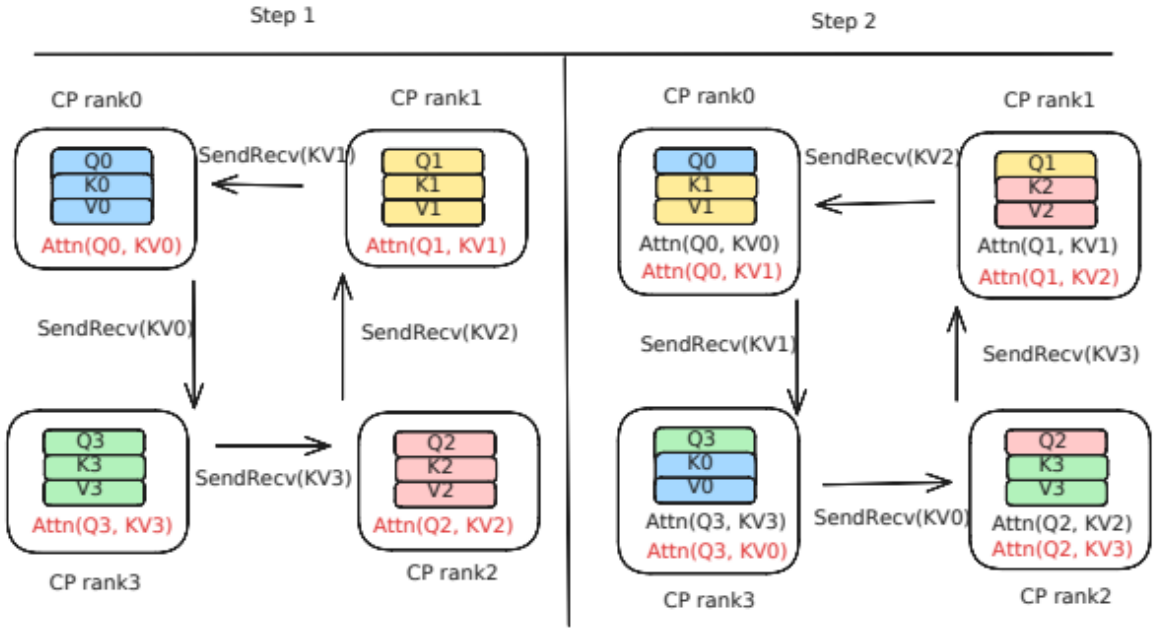
(a)



(b)



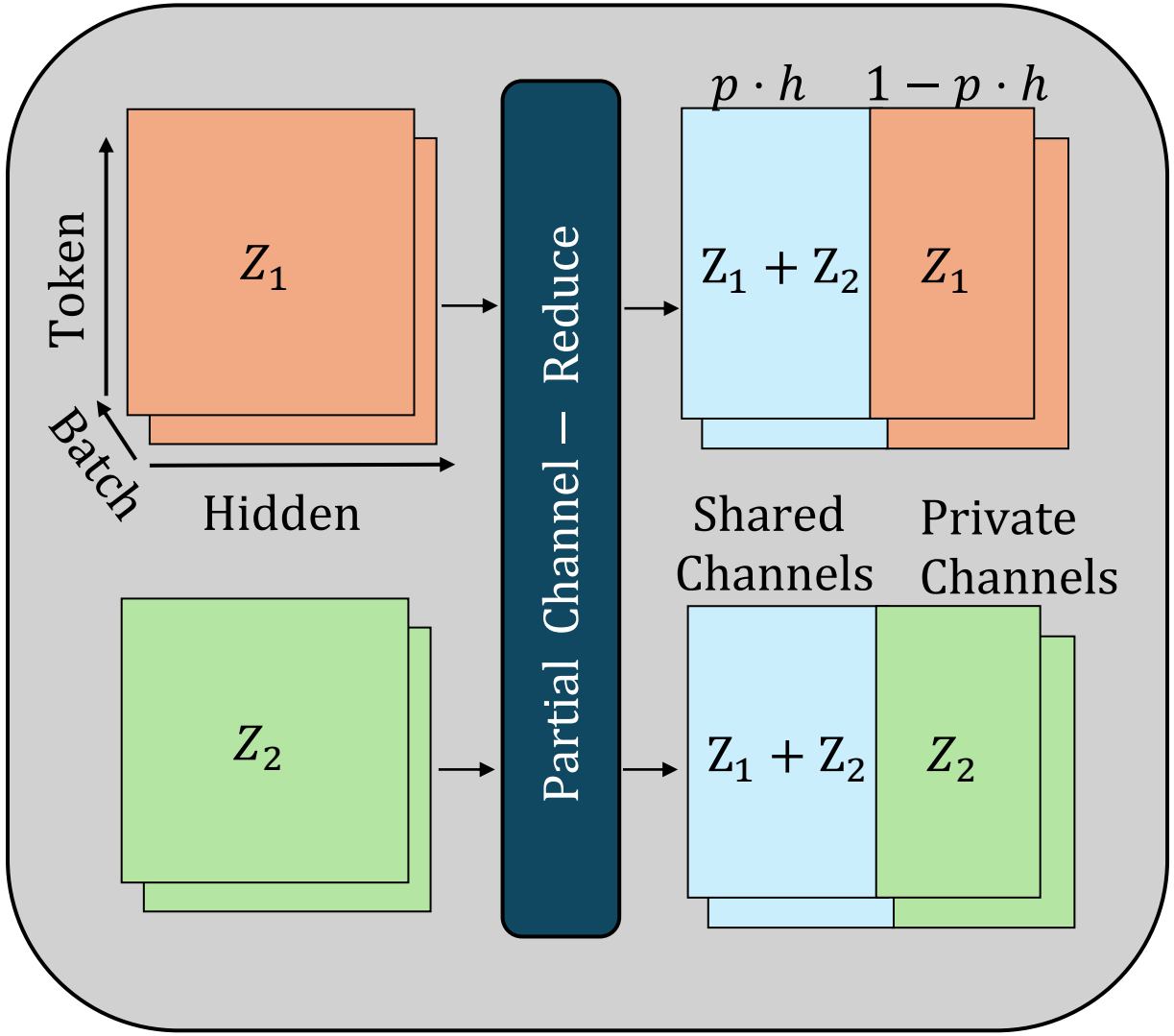
(c)



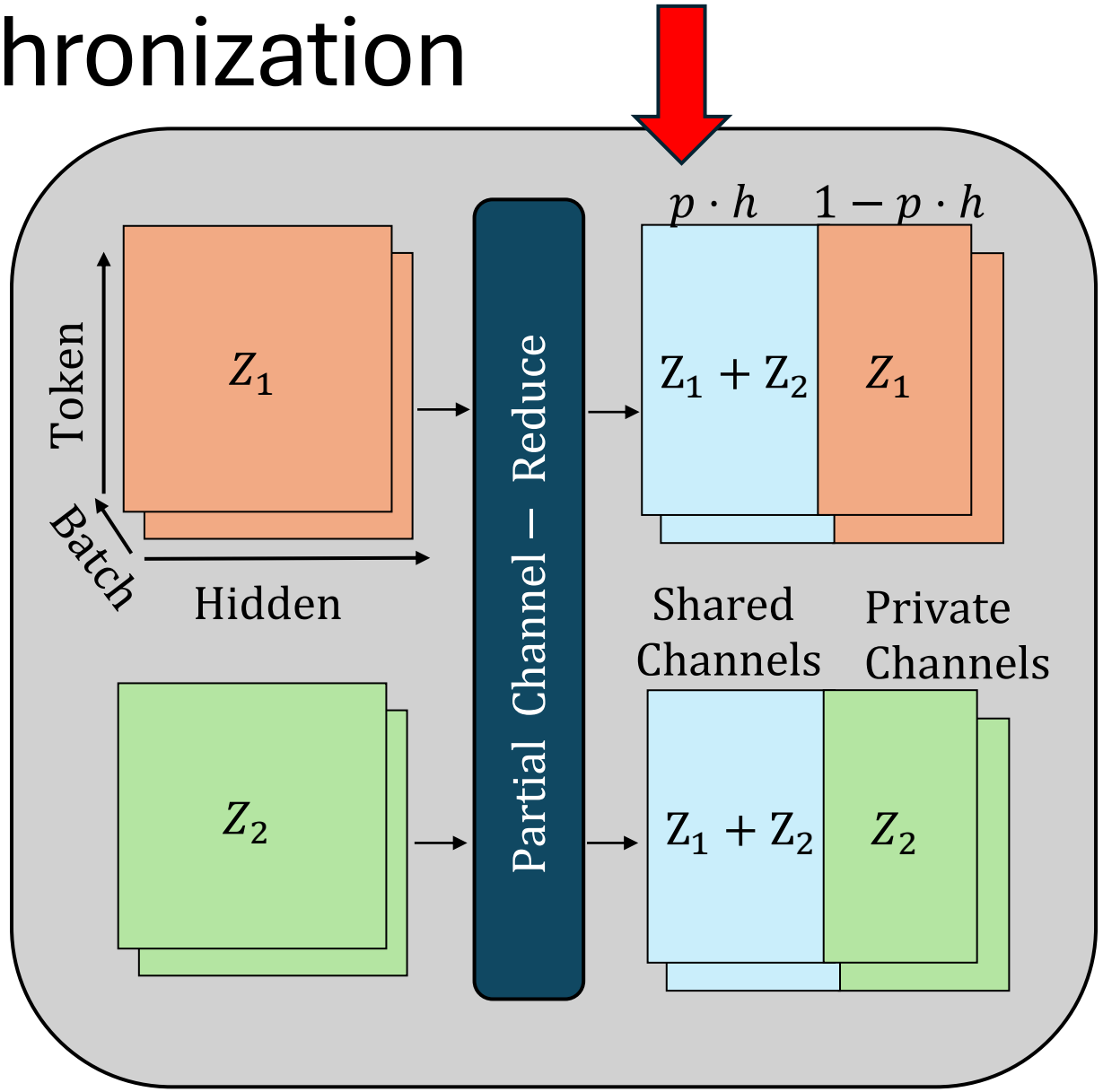
(a) **Domino: Eliminating communication in llm training via generic tensor slicing and overlapping** Guanhua Wang, Chengming Zhang, Zheyu Shen, Ang Li, and Olatunji Ruwase. 2024.
(b) **Computation vs. communication scaling for future transformers on future hardware.** Suchita Pati, Shaizeen Aga, Mahzabeen Islam, Nuwan Jayasena, and Matthew D. Sinclair. 2023

(c) **Context parallelism for scalable million-token** Amy Yang, Jingyi Yang, Aya Ibrahim, Xinfeng Xie, Bangsheng Tang, Grigory Sizov, Jeremy Reizenstein, Jongsoo Park, and Jianyu Huanginference, 2025

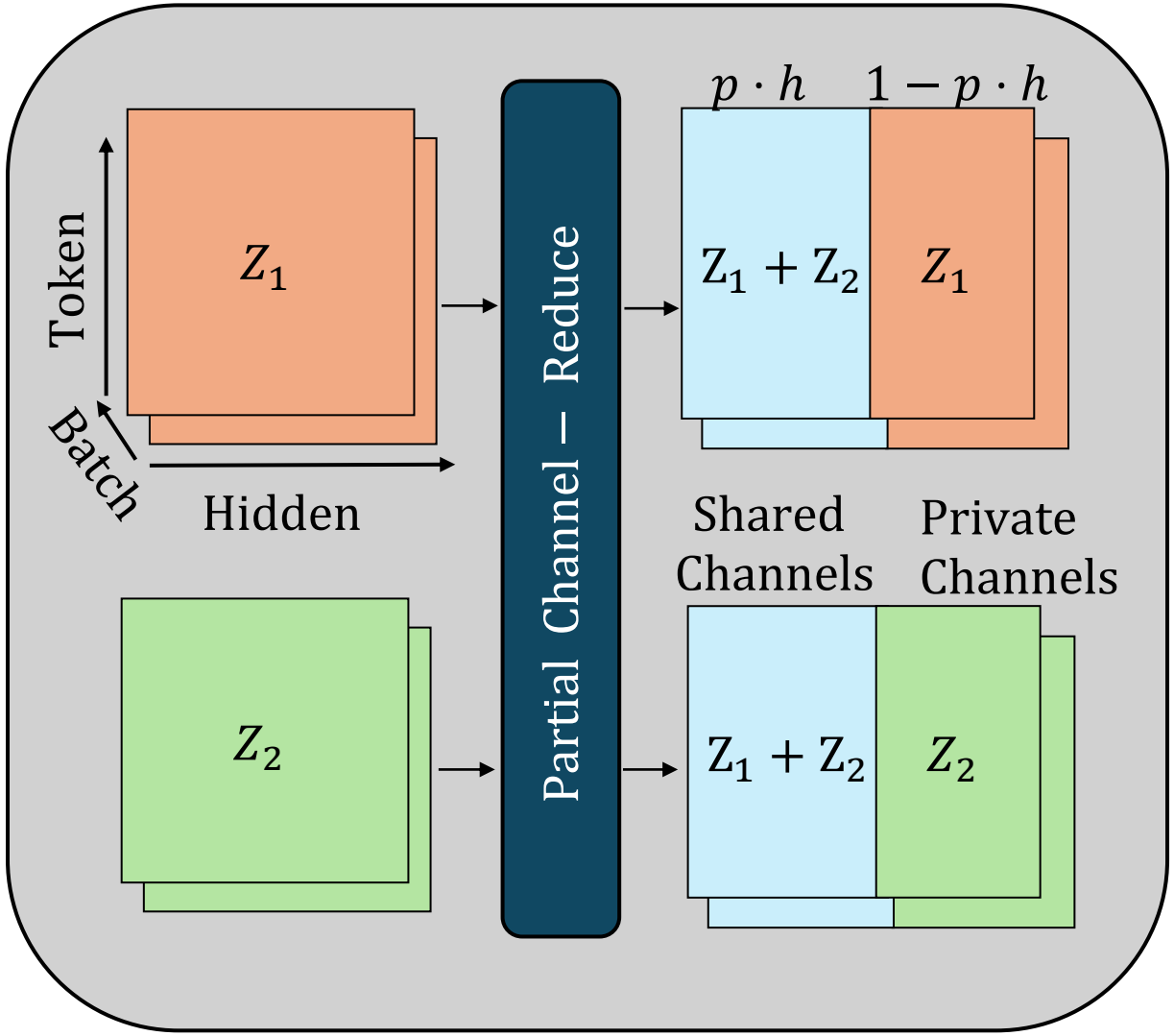
Partial Synchronization



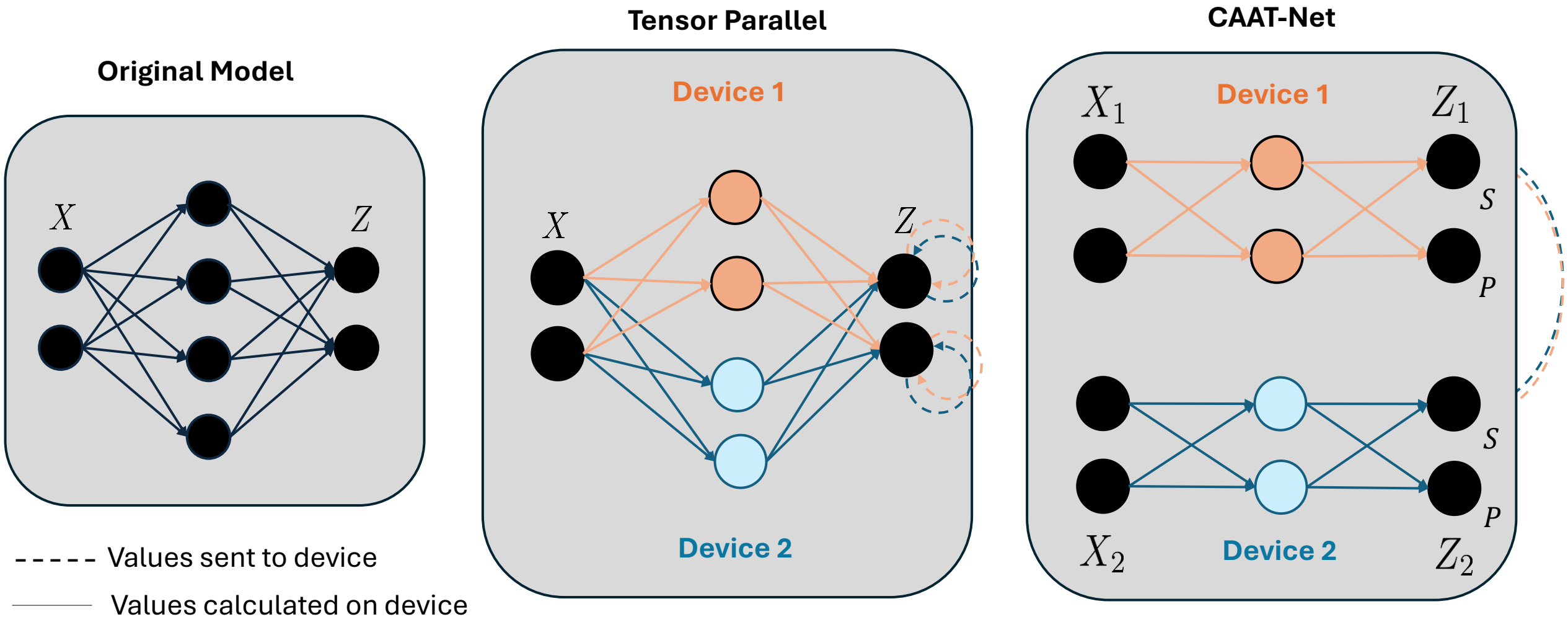
Partial Synchronization



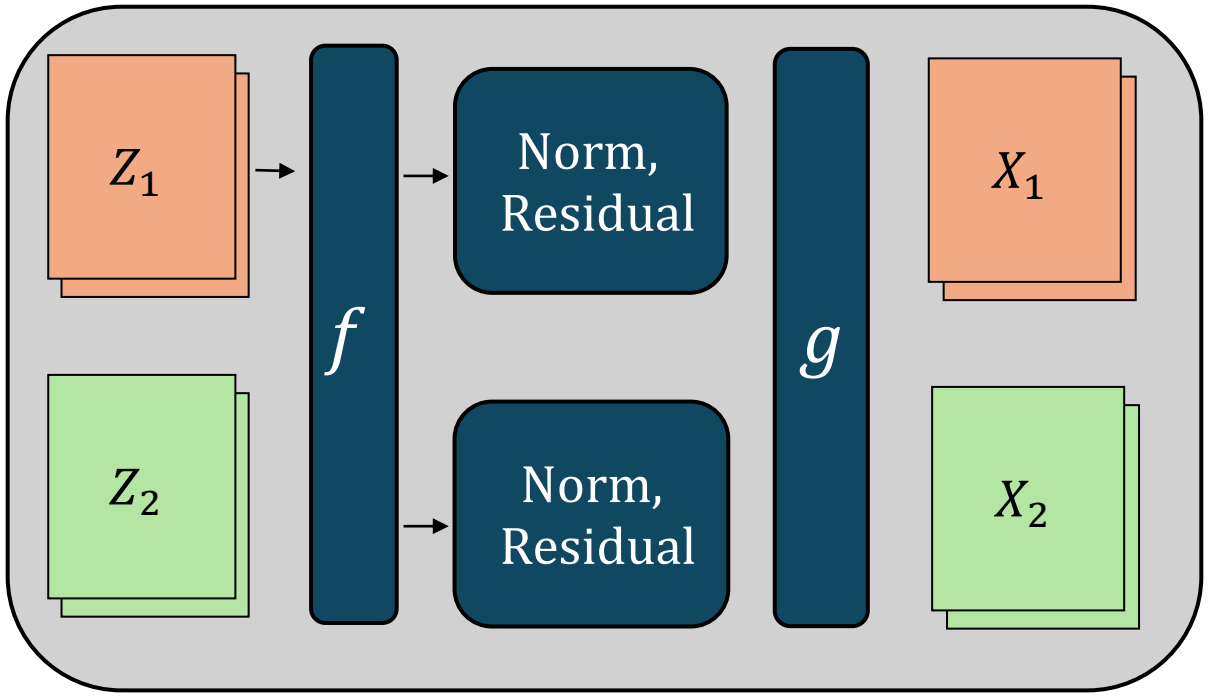
Partial Synchronization

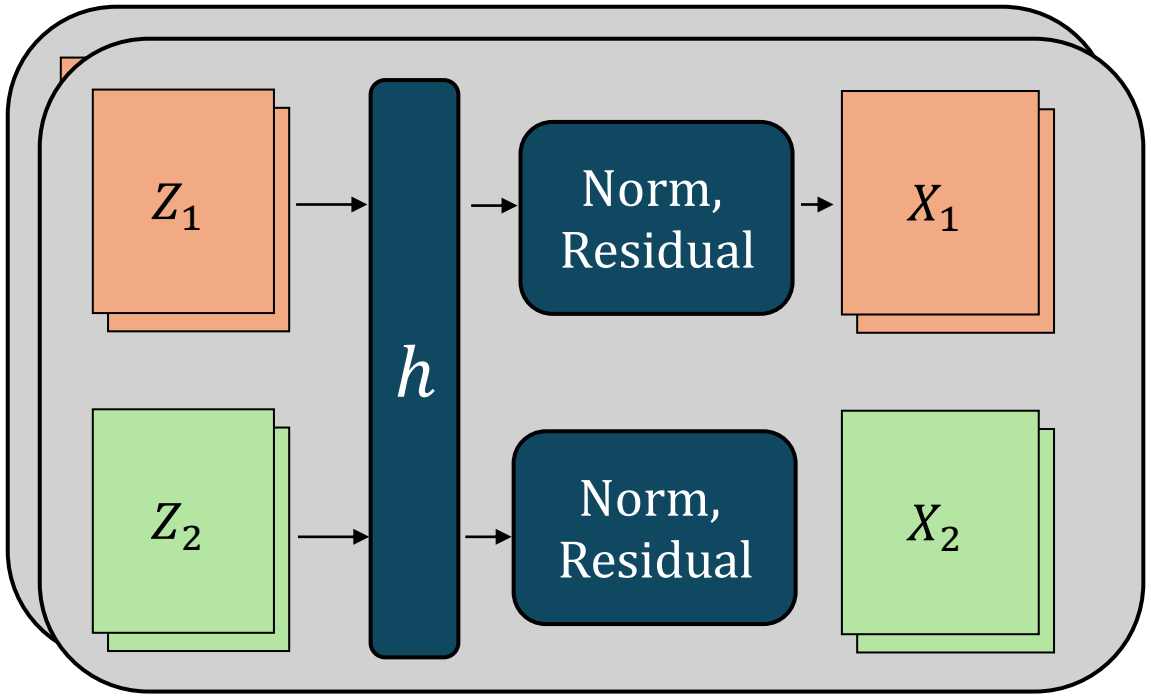


CAAT-Net


Original Model
X
Z
Tensor Parallel
Device 1
X
Z
Device 2
CAAT-Net
X_1
Device 1
Z_1
S
P
S
P
X_2
Device 2
Z_2
Values sent to device
Values calculated on device

Partial Synchronization – Implementation





Results and Evaluation

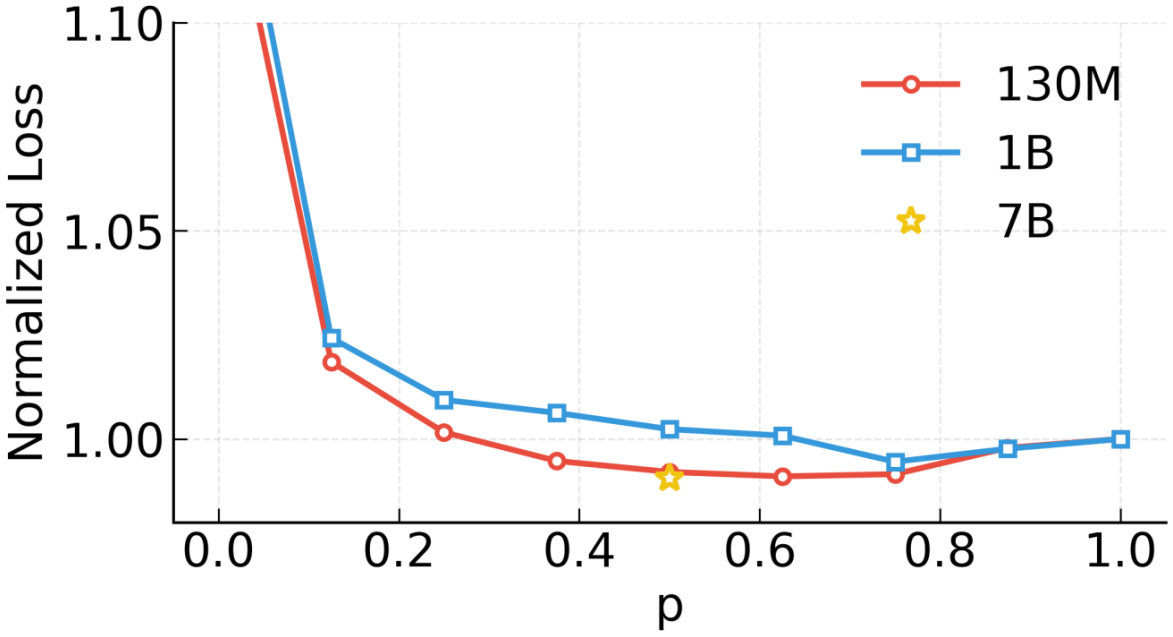
Table 1: **CAAT-Net vs baseline: Zero-shot accuracy after pretraining.** 7B parameter models, with $p = 0.5$ and tensor-parallel 8.

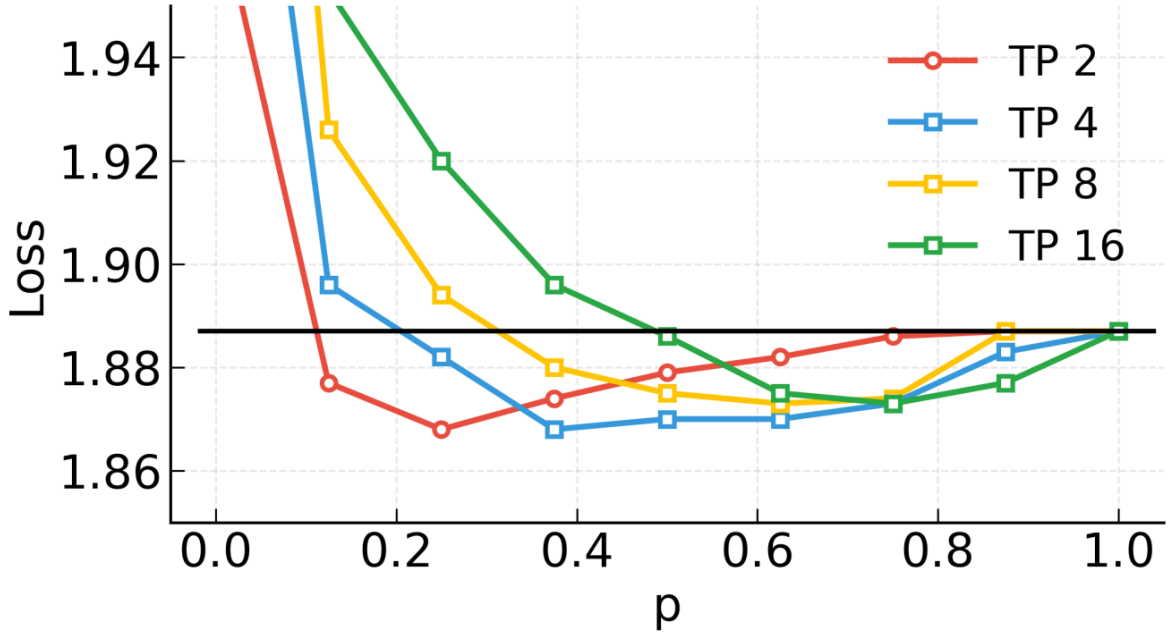
Model	LAMBADA (acc)	Hellaswag (acc)	WinoGrande (acc)	PIQA (acc)
Baseline	**61.34 ± 0.68**	45.85 ± 0.50	61.48 ± 1.37	**72.91 ± 1.06**
CAAT-Net	61.05 ± 0.68	**46.10 ± 0.50**	**62.19 ± 1.36**	72.86 ± 1.04
	OpenBookQA (acc)	BOOL-Q (acc)	WikiText (ppl)	Validation Loss
Baseline	**26.60 ± 1.98**	**64.89 ± 0.83**	12.51	1.01
CAAT-Net	24.00 ± 1.87	62.51 ± 0.85	**12.46**	**1.00**

Results and Evaluation

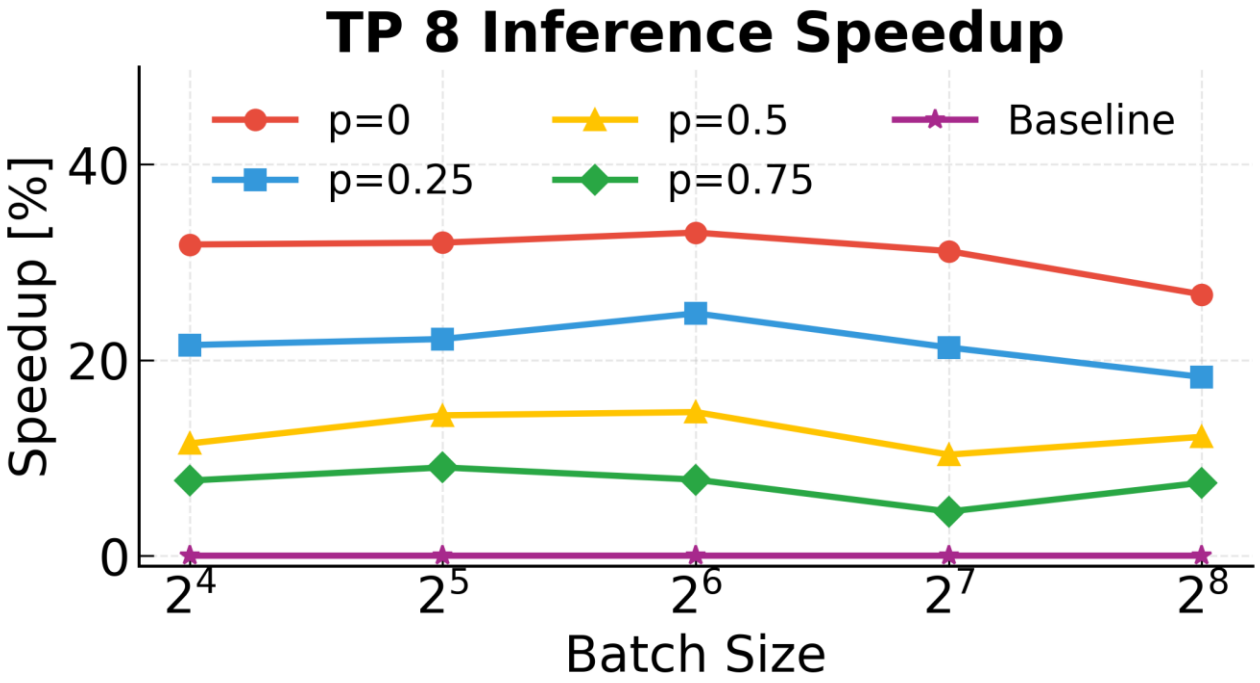
Table 1: **CAAT-Net vs baseline: Zero-shot accuracy after pretraining.** 7B parameter models, with $p = 0.5$ and tensor-parallel 8.

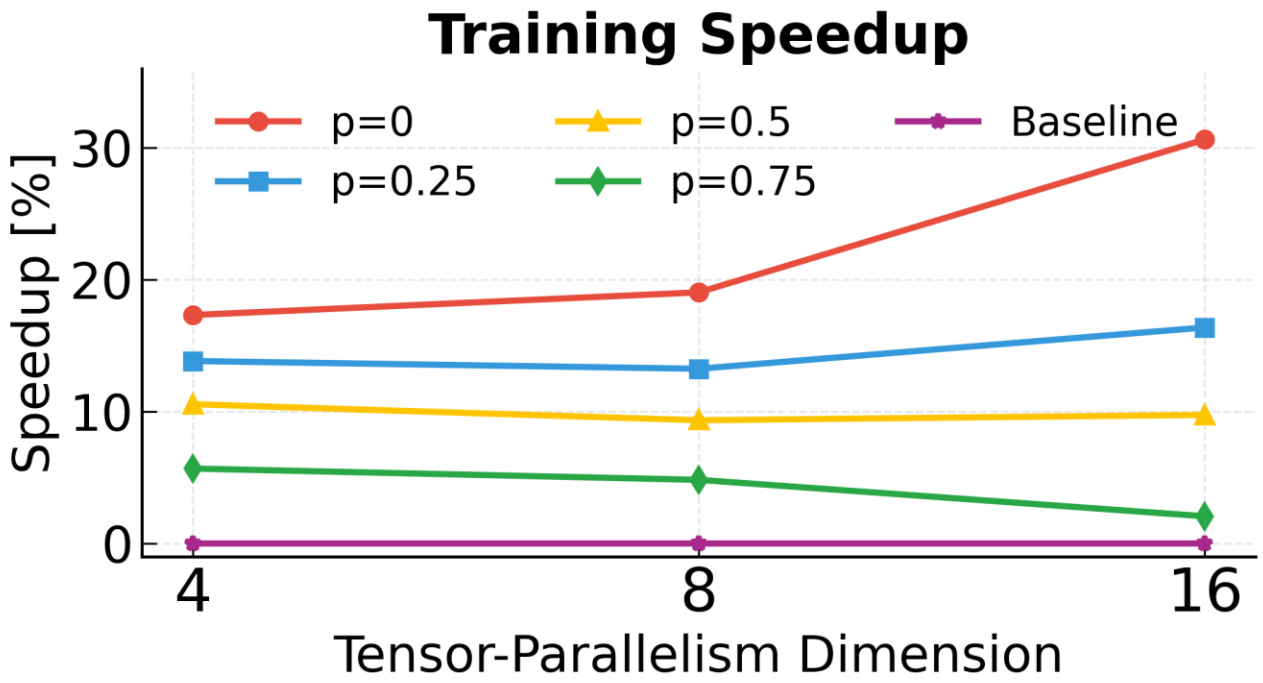
Model	LAMBADA (acc)	Hellaswag (acc)	WinoGrande (acc)	PIQA (acc)
Baseline	**61.34 ± 0.68**	45.85 ± 0.50	61.48 ± 1.37	**72.91 ± 1.06**
CAAT-Net	61.05 ± 0.68	**46.10 ± 0.50**	**62.19 ± 1.36**	72.86 ± 1.04
	OpenBookQA (acc)	BOOL-Q (acc)	WikiText (ppl)	Validation Loss
Baseline	**26.60 ± 1.98**	**64.89 ± 0.83**	12.51	1.01
CAAT-Net	24.00 ± 1.87	62.51 ± 0.85	**12.46**	**1.00**





Results and Evaluation





TECHNION Israel Institute of Technology

intel

NEURAL INFORMATION PROCESSING SYSTEMS

Thank You!

erc

The research of DS was funded by the European Union (ERC, A-B-C-Deep, 101039436).