NEURAL INFORMATION
PROCESSING SYSTEMS

ReplaceMe: Training-Free Depth Pruning via Transformer Block Linearization

D. Shopkhoev[1,2], A. Ali[1,2], M. Zhussip[1], V. Malykh[1,2,3], S. Lefkimmiatis[1], N. Komodakis[4,5,6], S. Zagoruyko[7]
[1]MTS AI, [2]ITMO University, [3]IITU, [4]University of Crete, [5]IACM-Forth, [6]Athena RC, [7]Polynome

Motivation & Problem

LLMs are powerful but computationally expensive—limiting their real-world use due to high latency, energy consumption, and hardware demands.

Prior works (e.g., UIDL, LLM-Streamline) have shown that LLM contain redundant blocks that can be removed. **BUT**, such pruning often leads to

(1) **performance degradation**

(2) **architectural modifications with fine-tuning**.

ReplaceMe – Core Idea

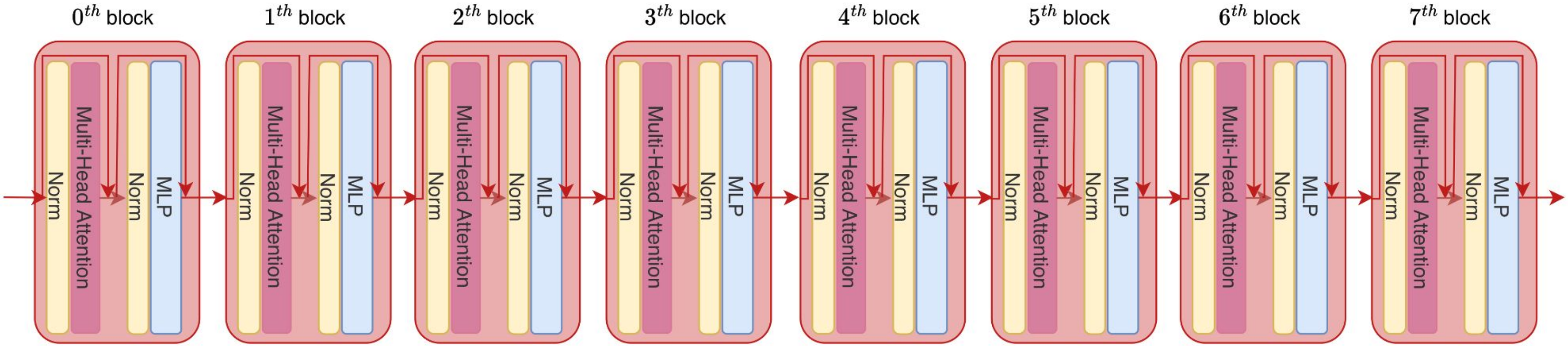
We propose **ReplaceMe** – a training-free depth-pruning method that replaces pruned blocks with a single linear transformation. Up to **25% depth reduction** with **>90% original performance retained** SoTA in accuracy, speed, and sustainability

Identify redundant consecutive transformer blocks.

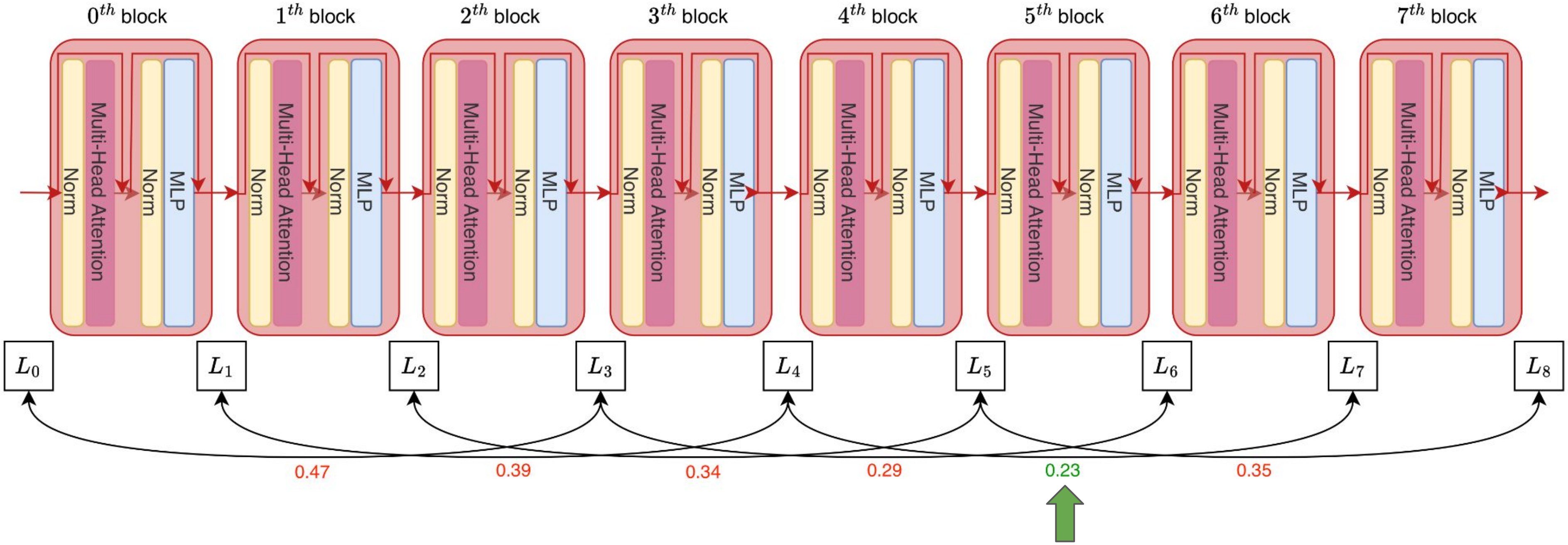
Replace them with a single linear transformation estimated from a small calibration dataset.

Fuse this linear transformation into MLP. **Say NO!** to (1) extra parameters! (2) retraining!

Original model

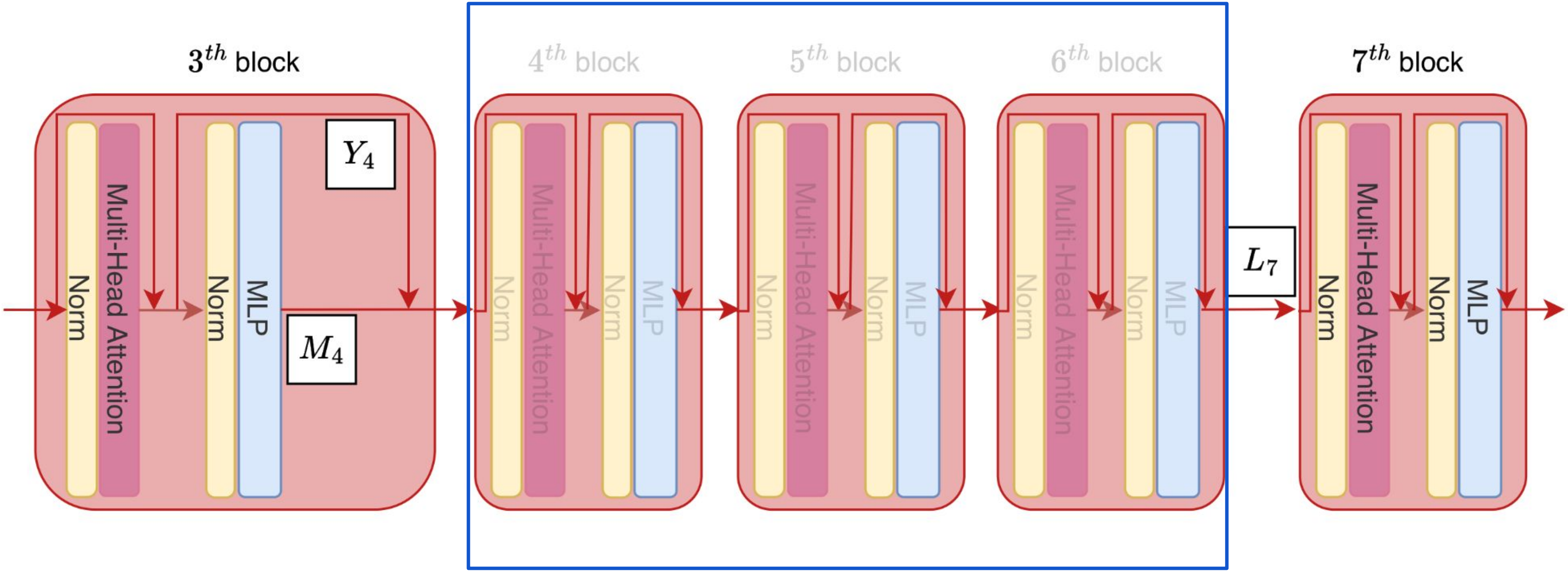


Layer Selection

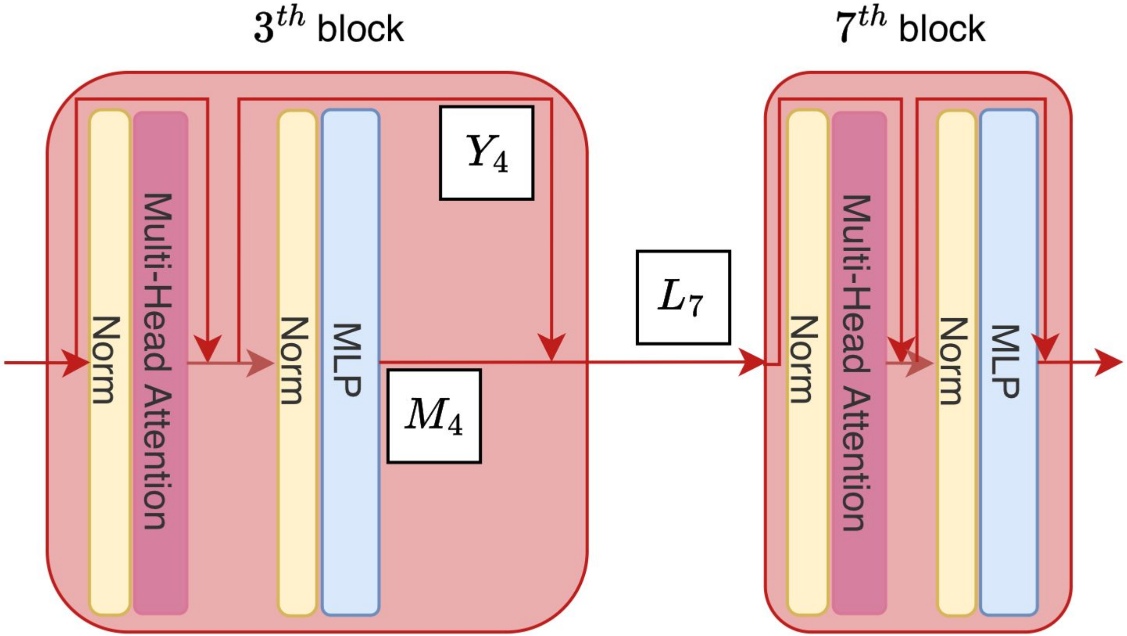


$$i^* = \arg\min_i h\left(\mathbf{L}_i, \mathbf{L}_{i+n}\right)$$

Removing layers

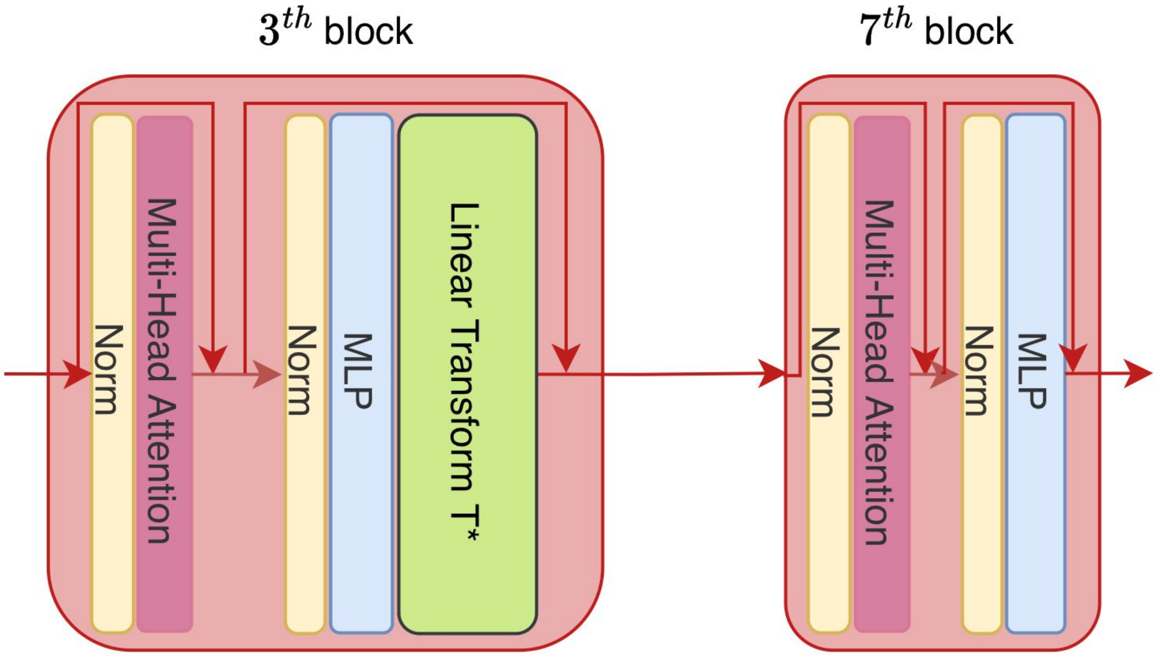


Linear Transform Estimation

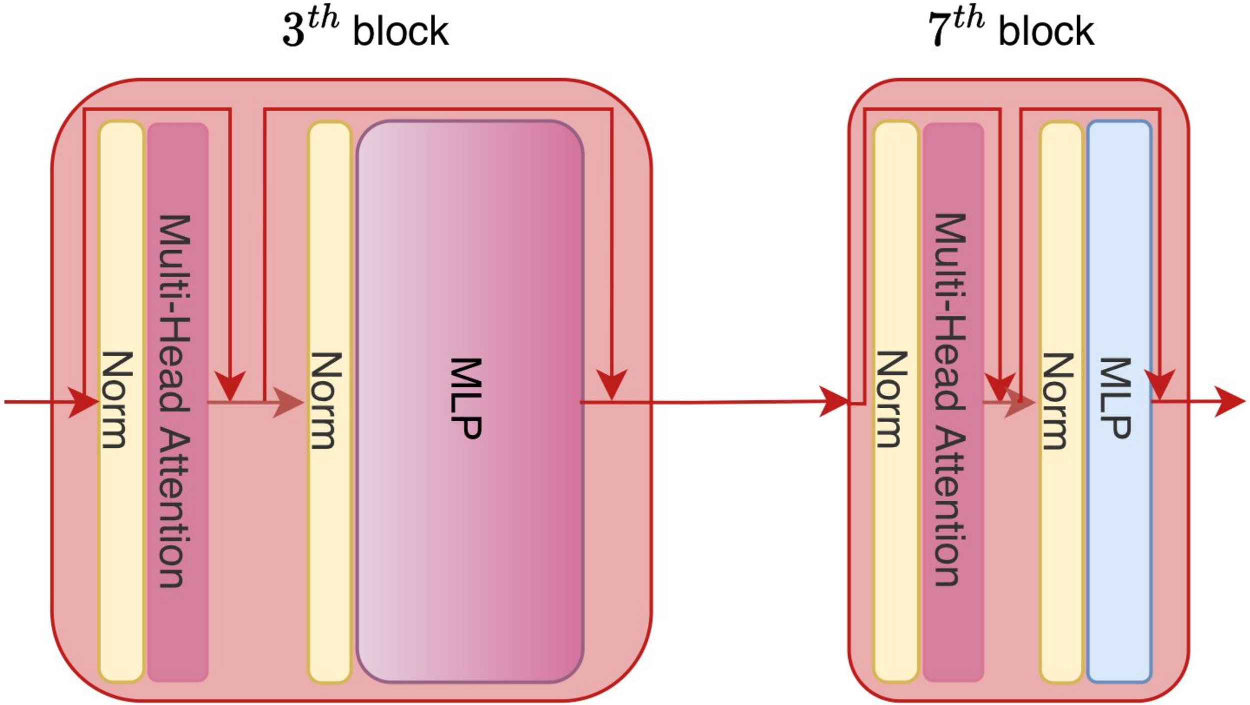


$$\mathbf{T}^* = \arg\min_{\mathbf{T}} h\big(\mathbf{M}_i \cdot \mathbf{T} + \mathbf{Y}_i; \mathbf{L}_{i+n}\big)$$

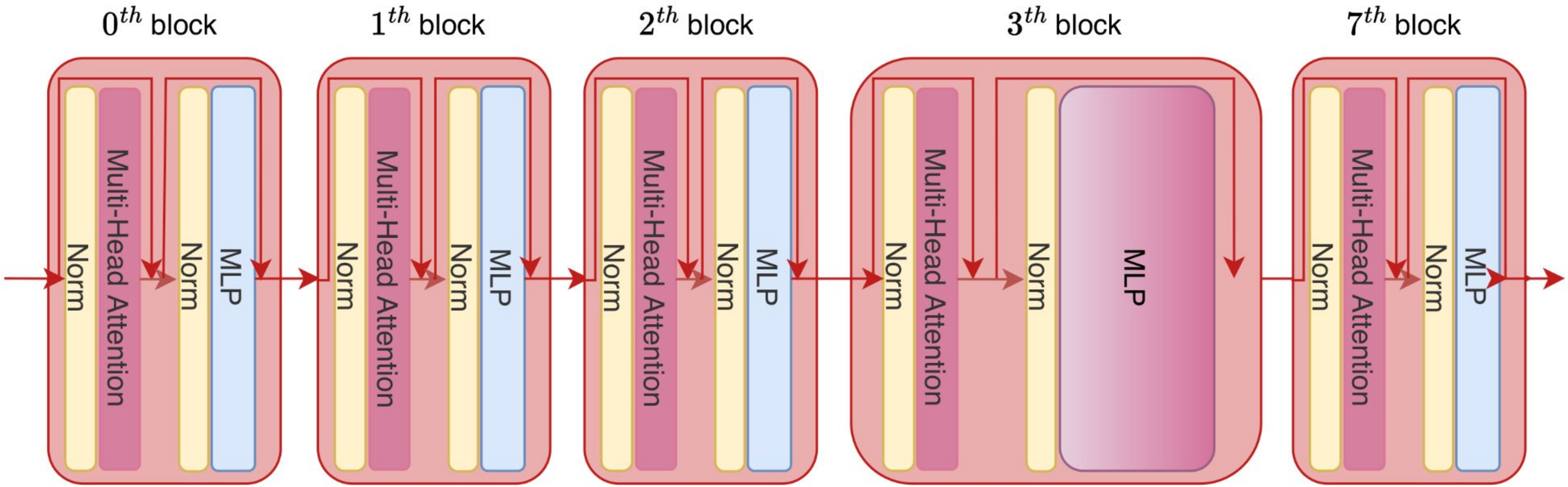
Linear Transform Estimation



Fusing Linear Transform



Final Model


0^{th} block
1^{th} block
2^{th} block
3^{th} block
7^{th} block
Norm
Multi-Head Attention
Norm
MLP

Method Overview

- **Step 1 – Layer Selection**: find cut index $i^* = \arg\min_i h\left(\mathrm{L}_i, \mathrm{L}_{i+n}\right)$
- **Step 2 – Linear Transform Estimation** $\mathbf{T}^* = \arg\min_{\mathbf{T}} h\big(\mathbf{M}_i \cdot \mathbf{T} + \mathbf{Y}_i; \mathbf{L}_{i+n}\big)$

  Solve:

  a. Analytical (L2): Least squares → fast

$$\mathbf{T}^* = (\mathbf{M}_i^T \cdot \mathbf{M}_i)^{-1} \cdot \mathbf{M}_i^T \cdot \big(\mathbf{L}_{i+n} - \mathbf{Y}_i\big)$$

  b. Numerical (Cosine): Adam optimizer → higher accuracy

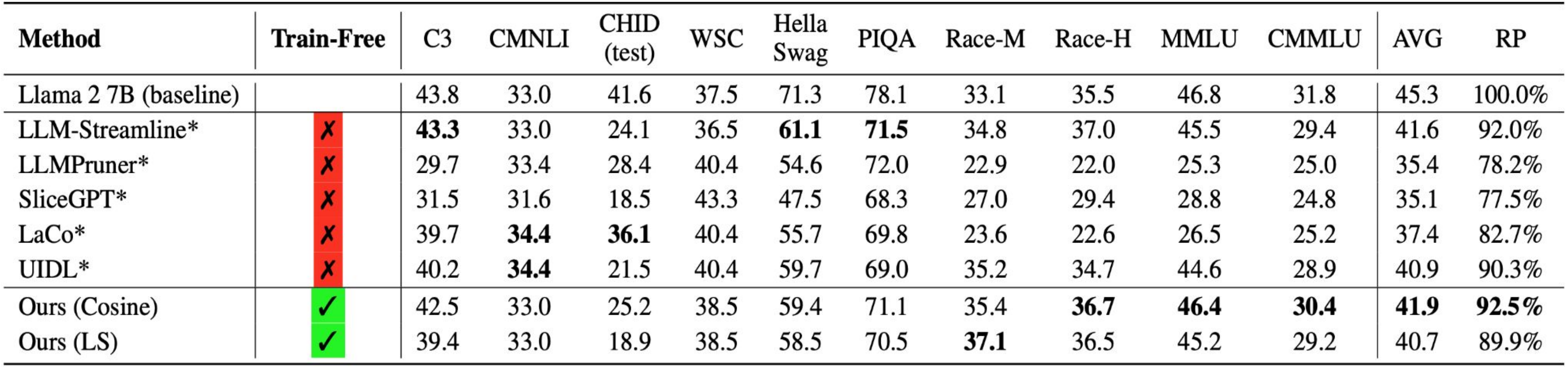
$$\mathbf{T}^* = \arg\min_{\mathbf{T}} \sum_{k=1}^{N} \left(1 - \frac{(\mathbf{M}_{i,k} \cdot \mathbf{T} + \mathbf{Y}_{i,k})^{\mathsf{T}} \cdot \mathbf{L}_{i+n,k}}{\|\mathbf{M}_{i,k} \cdot \mathbf{T} + \mathbf{Y}_{i,k}\|_2 \, \|\mathbf{L}_{i+n,k}\|_2}\right)$$

- **Regularization & Multi-LT:**

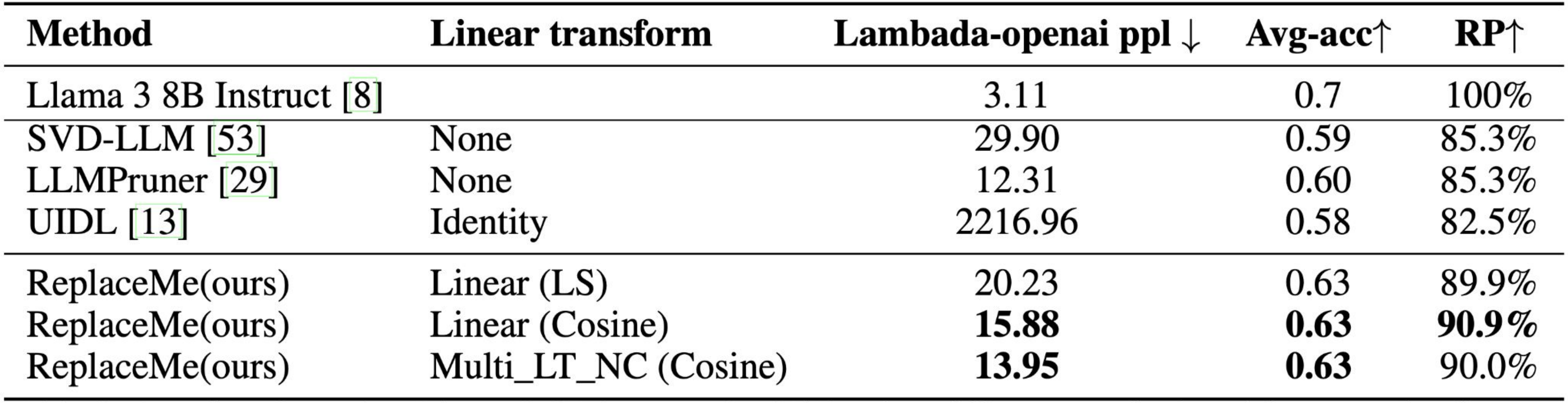
  a. L1/L2 regularization balances accuracy vs. perplexity

  b. Multi-LT: multiple non-overlapping pruned segments → better for high compression

Results – LLMs

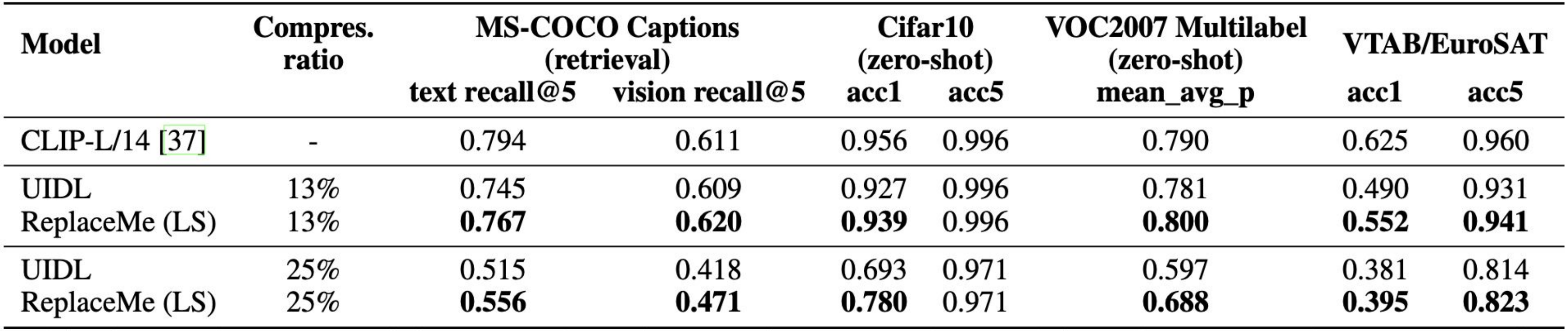


Method	**Train-Free**	C3	CMNLI	CHID (test)	WSC	Hella Swag	PIQA	Race-M	Race-H	MMLU	CMMLU	AVG	RP
Llama 2 7B (baseline)		43.8	33.0	41.6	37.5	71.3	78.1	33.1	35.5	46.8	31.8	45.3	100.0%
LLM-Streamline*	✗	**43.3**	33.0	24.1	36.5	**61.1**	**71.5**	34.8	37.0	45.5	29.4	41.6	92.0%
LLMPruner*	✗	29.7	33.4	28.4	40.4	54.6	72.0	22.9	22.0	25.3	25.0	35.4	78.2%
SliceGPT*	✗	31.5	31.6	18.5	43.3	47.5	68.3	27.0	29.4	28.8	24.8	35.1	77.5%
LaCo*	✗	39.7	**34.4**	**36.1**	40.4	55.7	69.8	23.6	22.6	26.5	25.2	37.4	82.7%
UIDL*	✗	40.2	**34.4**	21.5	40.4	59.7	69.0	35.2	34.7	44.6	28.9	40.9	90.3%
Ours (Cosine)	✓	42.5	33.0	25.2	38.5	59.4	71.1	35.4	**36.7**	**46.4**	**30.4**	**41.9**	**92.5%**
Ours (LS)	✓	39.4	33.0	18.9	38.5	58.5	70.5	**37.1**	36.5	45.2	29.2	40.7	89.9%



Method	**Linear transform**	**Lambada-openai ppl ↓**	**Avg-acc↑**	**RP↑**
Llama 3 8B Instruct [8]		3.11	0.7	100%
SVD-LLM [53]	None	29.90	0.59	85.3%
LLMPruner [29]	None	12.31	0.60	85.3%
UIDL [13]	Identity	2216.96	0.58	82.5%
ReplaceMe(ours)	Linear (LS)	20.23	0.63	89.9%
ReplaceMe(ours)	Linear (Cosine)	**15.88**	**0.63**	**90.9%**
ReplaceMe(ours)	Multi_LT_NC (Cosine)	**13.95**	**0.63**	90.0%

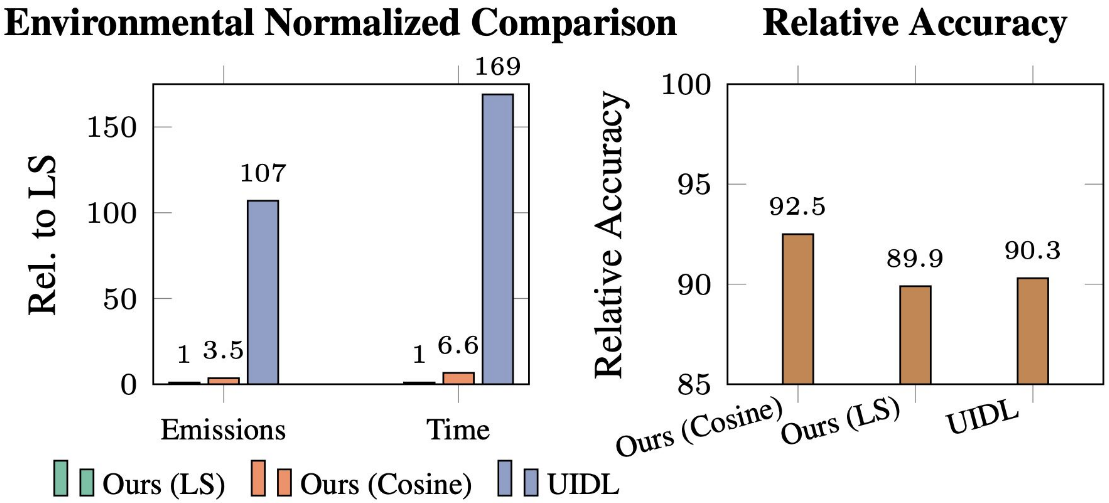
Vision Transformers – CLIP ViT



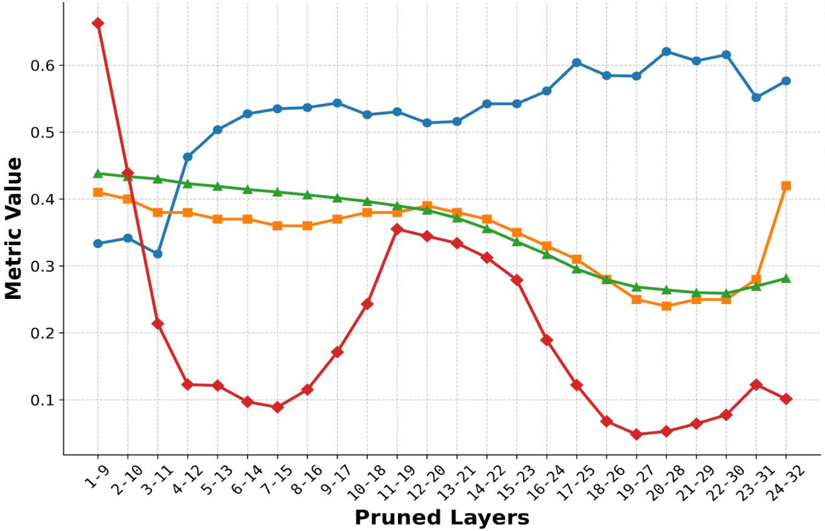
Model	Compres. ratio	MS-COCO Captions (retrieval)		Cifar10 (zero-shot)		VOC2007 Multilabel (zero-shot)	VTAB/EuroSAT	
		text recall@5	vision recall@5	acc1	acc5	mean_avg_p	acc1	acc5
CLIP-L/14 [37]	-	0.794	0.611	0.956	0.996	0.790	0.625	0.960
UIDL	13%	0.745	0.609	0.927	0.996	0.781	0.490	0.931
ReplaceMe (LS)	13%	**0.767**	**0.620**	**0.939**	0.996	**0.800**	**0.552**	**0.941**
UIDL	25%	0.515	0.418	0.693	0.971	0.597	0.381	0.814
ReplaceMe (LS)	25%	**0.556**	**0.471**	**0.780**	0.971	**0.688**	**0.395**	**0.823**

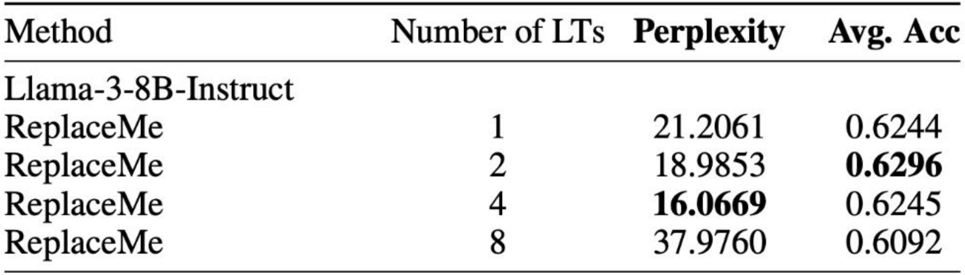
Efficiency & Sustainability

- No retraining → **~100× less CO_2** vs. UIDL (Fig. 2)
- Compression time: **minutes vs. hours/days**
- Memory: stores only 2 activations per token (optimized cosine loss)
- Fused architecture: **no inference overhead**

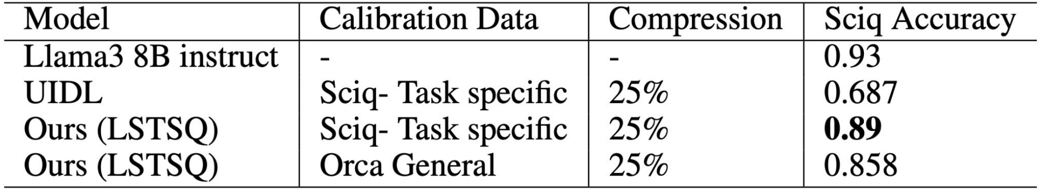


Ablations & Insights

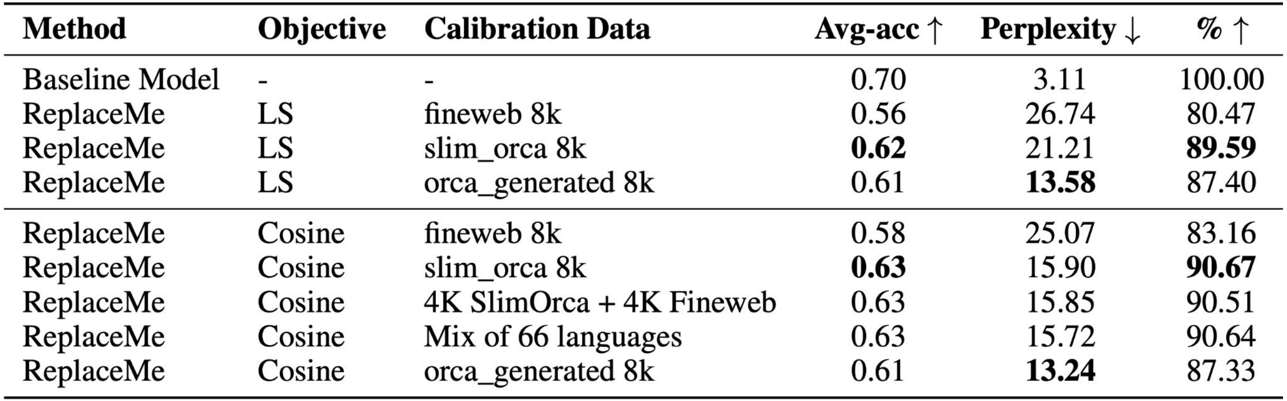




Method	Number of LTs	**Perplexity**	**Avg. Acc**
Llama-3-8B-Instruct			
ReplaceMe	1	21.2061	0.6244
ReplaceMe	2	18.9853	**0.6296**
ReplaceMe	4	**16.0669**	0.6245
ReplaceMe	8	37.9760	0.6092



Model	Calibration Data	Compression	Sciq Accuracy
Llama3 8B instruct	-	-	0.93
UIDL	Sciq- Task specific	25%	0.687
Ours (LSTSQ)	Sciq- Task specific	25%	**0.89**
Ours (LSTSQ)	Orca General	25%	0.858



Method	**Objective**	**Calibration Data**	**Avg-acc ↑**	**Perplexity ↓**	**% ↑**
Baseline Model	-	-	0.70	3.11	100.00
ReplaceMe	LS	fineweb 8k	0.56	26.74	80.47
ReplaceMe	LS	slim_orca 8k	**0.62**	21.21	**89.59**
ReplaceMe	LS	orca_generated 8k	0.61	**13.58**	87.40
ReplaceMe	Cosine	fineweb 8k	0.58	25.07	83.16
ReplaceMe	Cosine	slim_orca 8k	**0.63**	15.90	**90.67**
ReplaceMe	Cosine	4K SlimOrca + 4K Fineweb	0.63	15.85	90.51
ReplaceMe	Cosine	Mix of 66 languages	0.63	15.72	90.64
ReplaceMe	Cosine	orca_generated 8k	0.61	**13.24**	87.33

Conclusion & Impact

- **ReplaceMe is a simple, training-free, effective depth pruning method**
- >90% original performance at 25% compression
- Works across LLMs (Llama, Qwen, Falcon) and Vision Transformers (CLIP-ViT)
- Open-source: https://github.com/mts-ai/ReplaceMe
- Enables sustainable, accessible AI without retraining

Conclusion & Impact

- ReplaceMe is a simple, training-free, effective depth pruning method
- **>90% original performance at 25% compression**
- Works across LLMs (Llama, Qwen, Falcon) and Vision Transformers (CLIP-ViT)
- Open-source: https://github.com/mts-ai/ReplaceMe
- Enables sustainable, accessible AI without retraining

Conclusion & Impact

- ReplaceMe is a simple, training-free, effective depth pruning method
- >90% original performance at 25% compression
- **Works across LLMs (Llama, Qwen, Falcon) and Vision Transformers (CLIP-ViT)**
- Open-source: https://github.com/mts-ai/ReplaceMe
- Enables sustainable, accessible AI without retraining

Conclusion & Impact

- ReplaceMe is a simple, training-free, effective depth pruning method
- >90% original performance at 25% compression
- Works across LLMs (Llama, Qwen, Falcon) and Vision Transformers (CLIP-ViT)
- **Open-source: https://github.com/mts-ai/ReplaceMe**
- Enables sustainable, accessible AI without retraining

Conclusion & Impact

- ReplaceMe is a simple, training-free, effective depth pruning method
- >90% original performance at 25% compression
- Works across LLMs (Llama, Qwen, Falcon) and Vision Transformers (CLIP-ViT)
- Open-source: https://github.com/mts-ai/ReplaceMe
- **Enables sustainable, accessible AI without retraining**

Thank you!

We invite you to our poster for more details about our work

Open-source framework: https://github.com/mts-ai/ReplaceMe