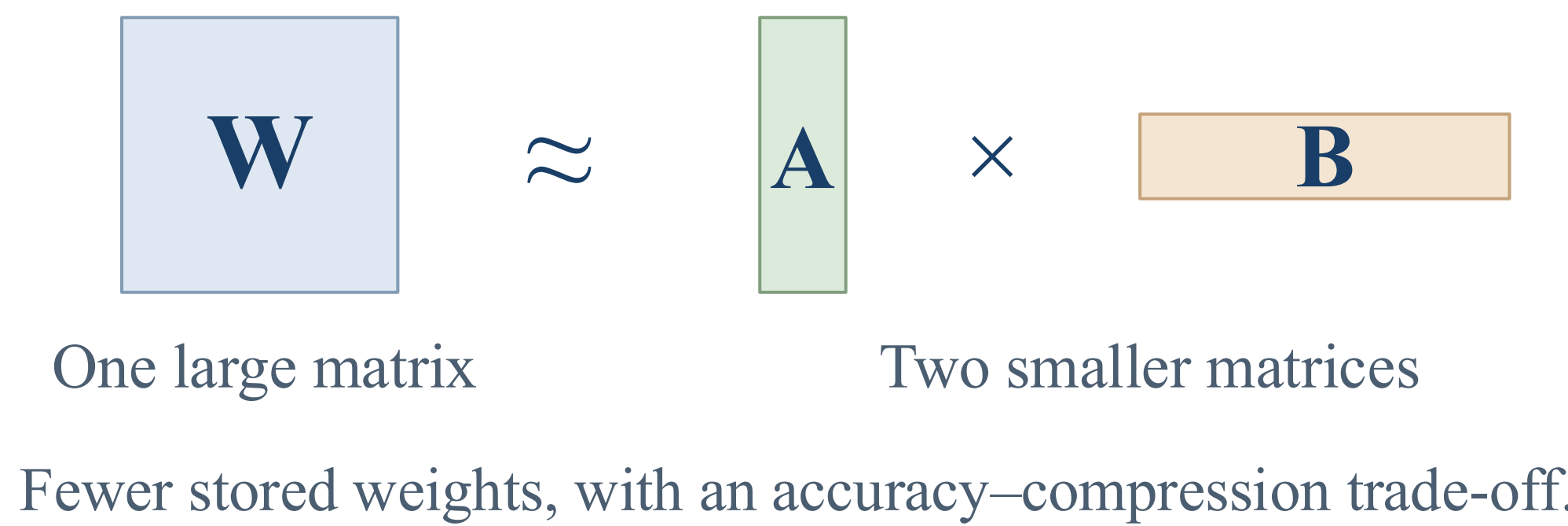


Taking Low-Rank LLM Compression a Step Further: A Global Perspective with Fused Inference

Introduction

Background

- Model compression reduces the memory and computation needed to run large language models.
- Low-rank decomposition approximates one large weight matrix with two smaller matrices.



Motivation

- Existing low-rank methods (e.g., SVD-LLM and SkipCat) rely on **greedy, layer-wise reconstruction**.
- Local truncation errors accumulate across layers, inducing **Cascading Error Propagation** and perplexity explosion at high compression.
- Sequential evaluation $Y = A(BX)$ adds **CPU kernel-launch overhead**.

Can we optimize low-rank compression globally and achieve real wall-clock inference speedup?

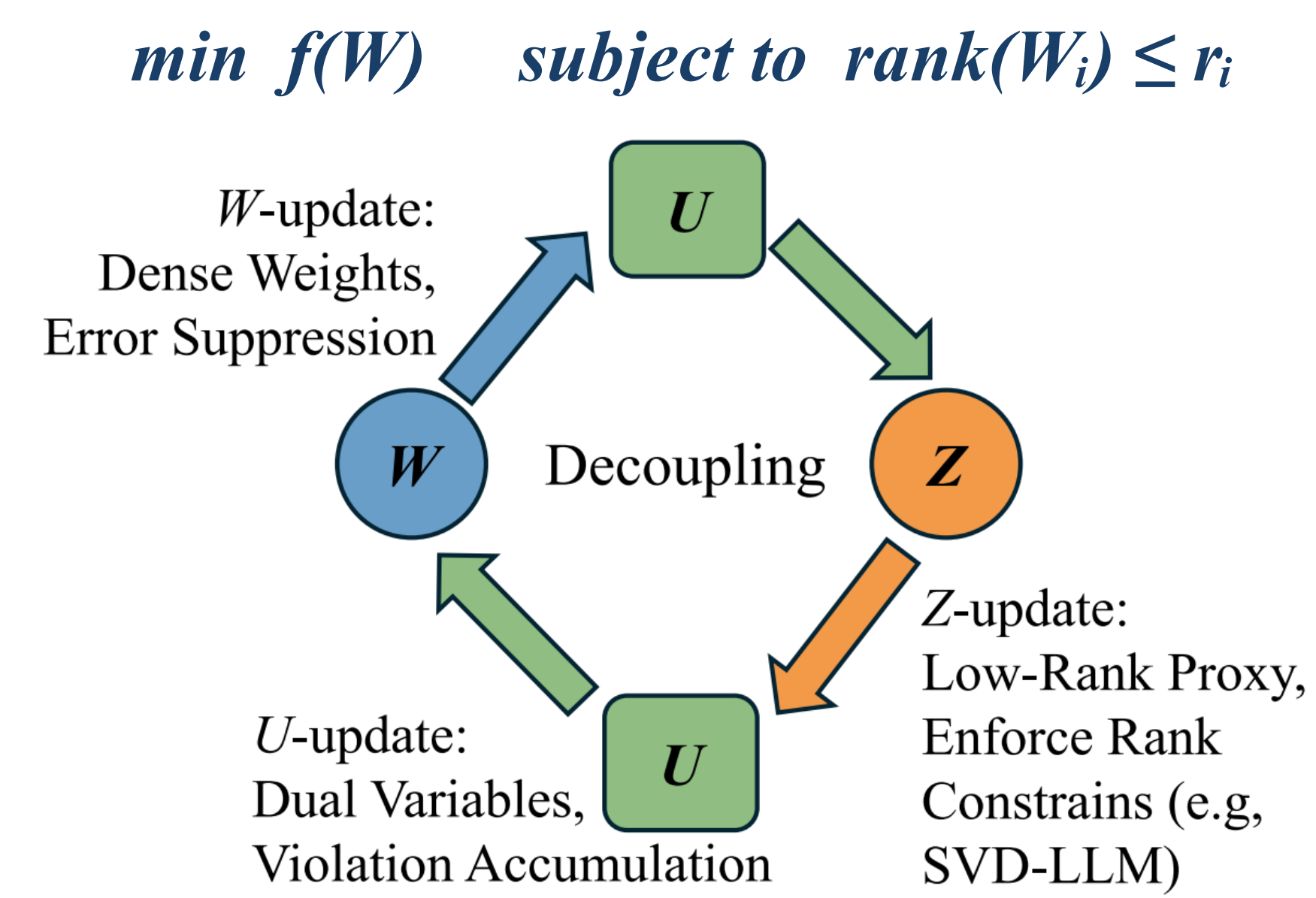
Our approach

- Global ADMM optimization** coordinates all layers through variable splitting, decoupling non-convex rank constraints from the global language modeling loss to suppress cascading error.
- Fused ABX inference:** combines low-rank projections in one kernel.

Method

Global optimization with ADMM

Separate the language-model loss from the rank constraints using a low-rank proxy Z and scaled dual variables U .



Global error suppression, low-rank projection and dual feedback form an iterative cycle.

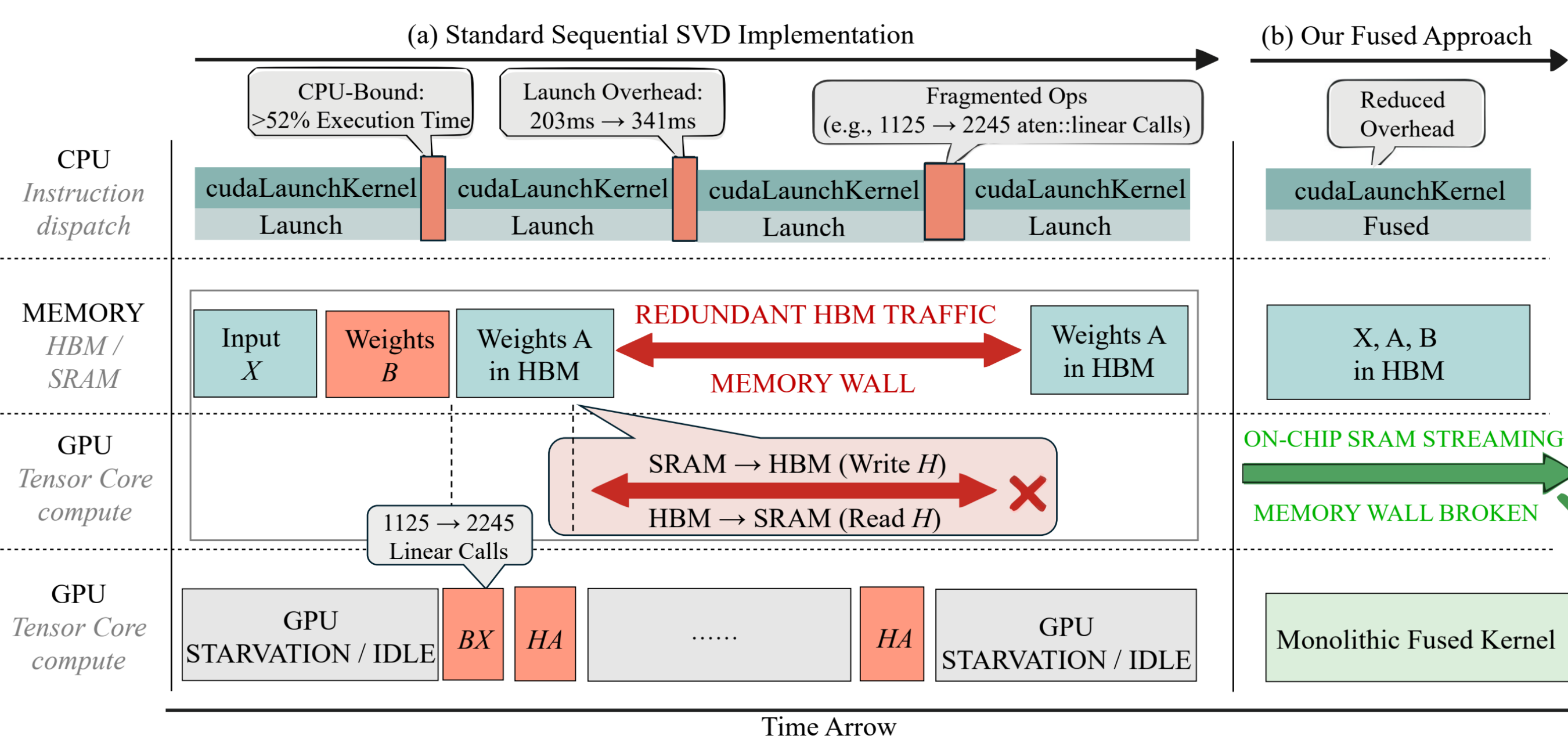
W-update: minimize the global loss with a proximal penalty.

Z-update: project onto the rank constraint with SVD-LLM or SkipCat.

U-update: accumulate the residual between W and Z .

Hardware-aware fused ABX

Compute $Y = A(BX)$ in one CUDA kernel. Keep $H = BX$ on-chip to avoid writing and reading the intermediate activation in HBM.



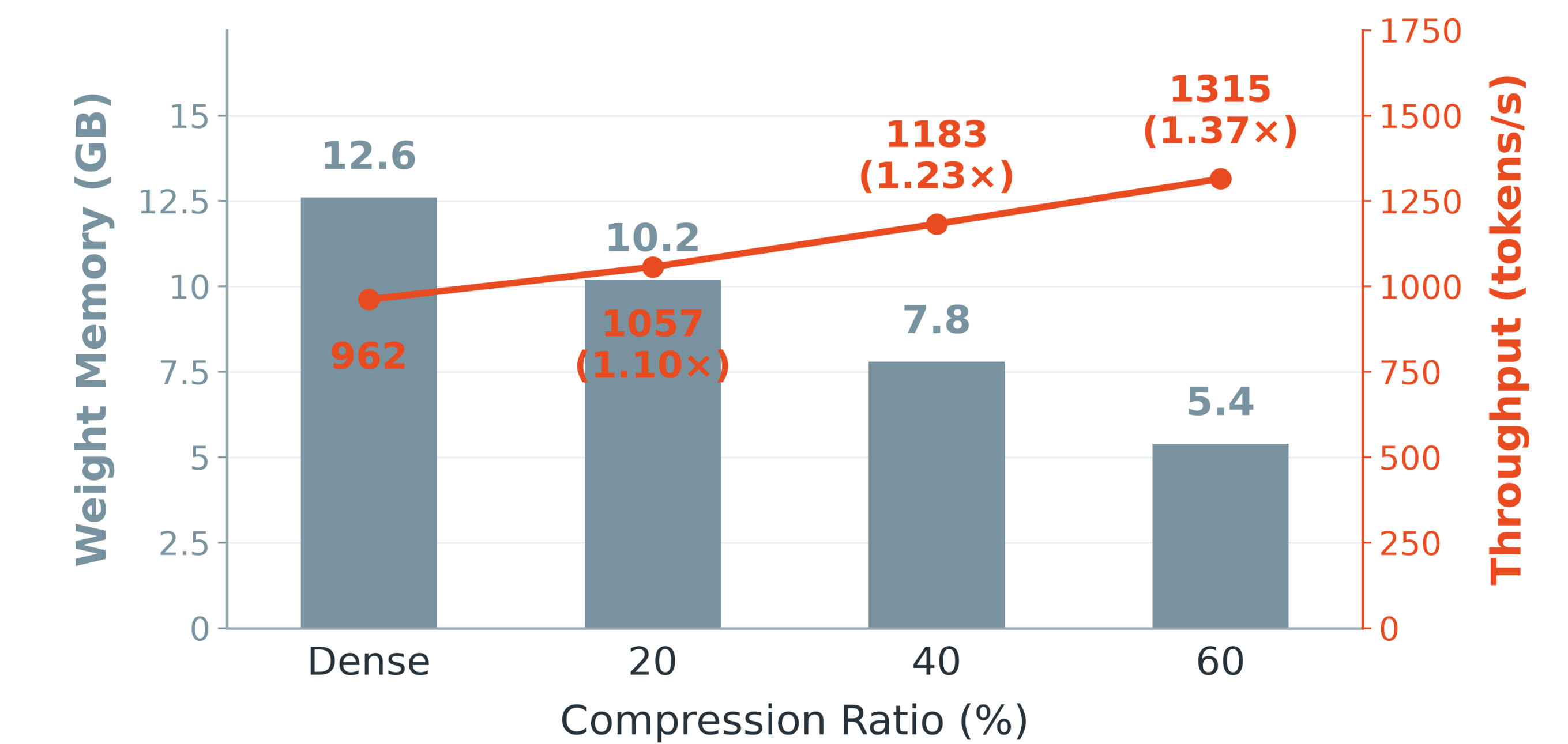
Experimental Results

Model quality at 30% compression

Model	Metric	SVD-LLM	ADMM-SVD	SkipCat	ADMM-CatSVD
Qwen3-8B	PPL ↓	20.15	12.35	14.82	10.05
	Avg. ↑	46.70	51.89	57.65	60.66
LLaMA-2-7B	PPL ↓	11.66	7.42	7.77	6.58
	Avg. ↑	50.03	54.77	56.24	57.89
LLaMA-3-8B	PPL ↓	43.91	10.97	48.24	9.69
	Avg. ↑	46.98	54.59	39.64	59.70

PPL: perplexity. Avg.: mean accuracy (%) over seven zero-shot reasoning tasks. ADMM-SVD and ADMM-CatSVD represent the application of our ADMM-based method to the SVD-LLM and SkipCat, respectively..

End-to-end inference speed



LLaMA-2-7B inference with Hugging Face Transformers on a single NVIDIA A800 80GB GPU.
Batch 128, prompt 128 tokens, generation 16 tokens.

Conclusion

Global ADMM improves model fidelity. Fused ABX turns low-rank compression into measured inference acceleration.