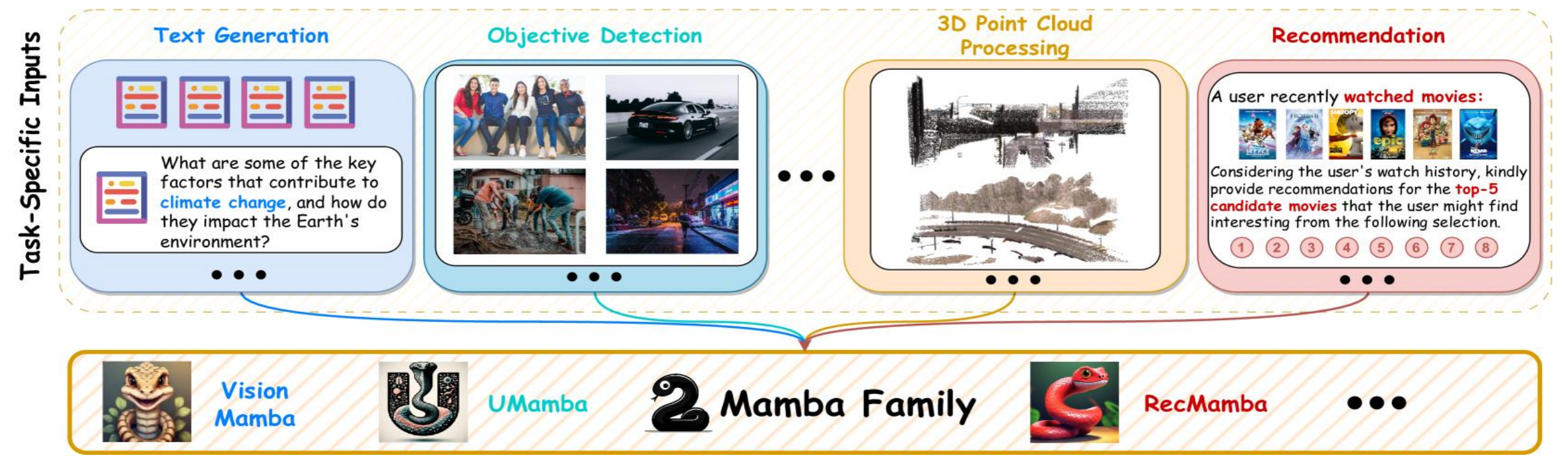
MambaQuant: Quantizing The Mamba Family With Variance Aligned Rotation Methods

Zukang Xu*, Yuxuan Yue[1,2*†], Xing Hu[1], Zhihang Yuan[1], Zixu Jiang[1,3†], Zhixuan Chen[1], Jiangyong Yu[1], Chen Xu[1], Sifan Zhou[1,4†], Dawei Yang[1] ✉

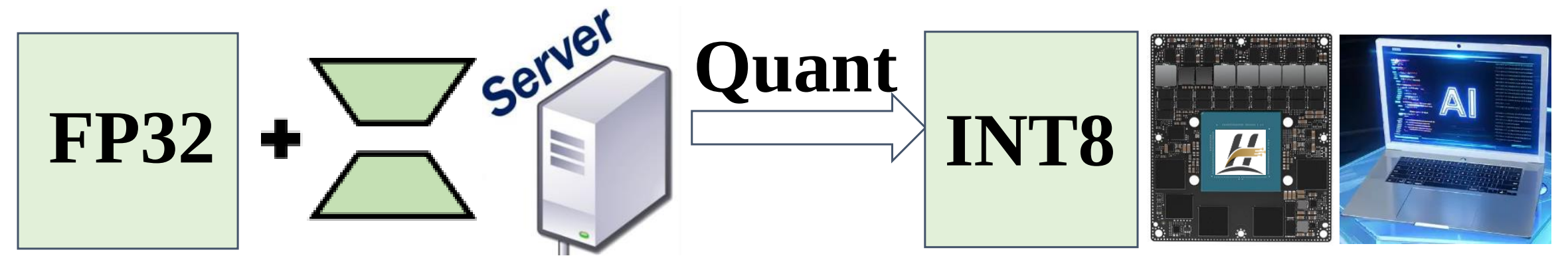
[1] Houmo AI [2] Harbin Institute of Technology (Shenzhen) [3] Nanjing University [4]Southeast University

Motivation

Mamba is widely applied across domains.


Task-Specific Inputs
Text Generation
What are some of the key factors that contribute to climate change, and how do they impact the Earth's environment?
Objective Detection
3D Point Cloud Processing
Recommendation
A user recently watched movies:
Considering the user's watch history, kindly provide recommendations for the top-5 candidate movies that the user might find interesting from the following selection.
1 2 3 4 5 6 7 8
Vision Mamba
UMamba
Mamba Family
RecMamba

Quantization compresses models, cuts costs.

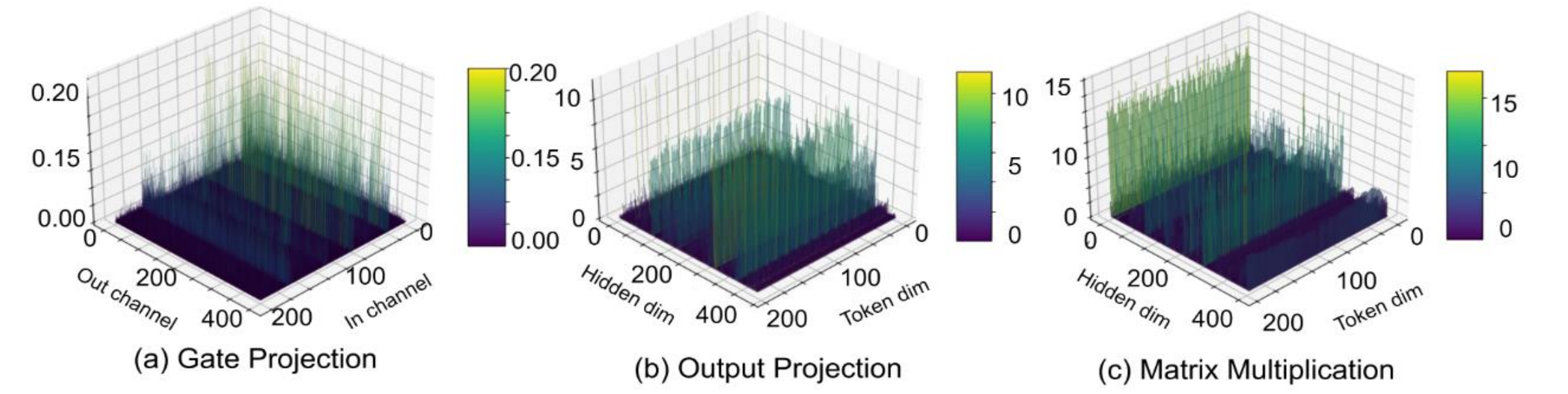


Mamba quantization under-researched, solns urgent.

- ➢Lack of systematic exploration.
- ➢Ineffectiveness of existing methods.
- ➢Unique challenges in Mamba.

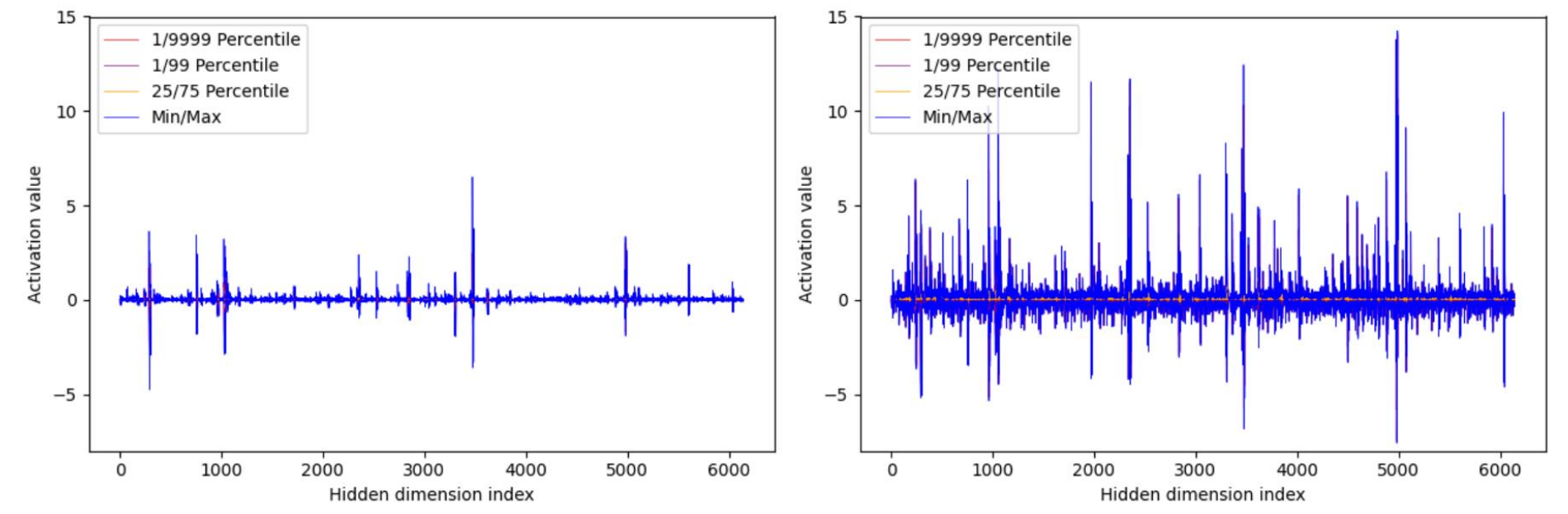
Challenge

1. Significant outliers in Mamba models

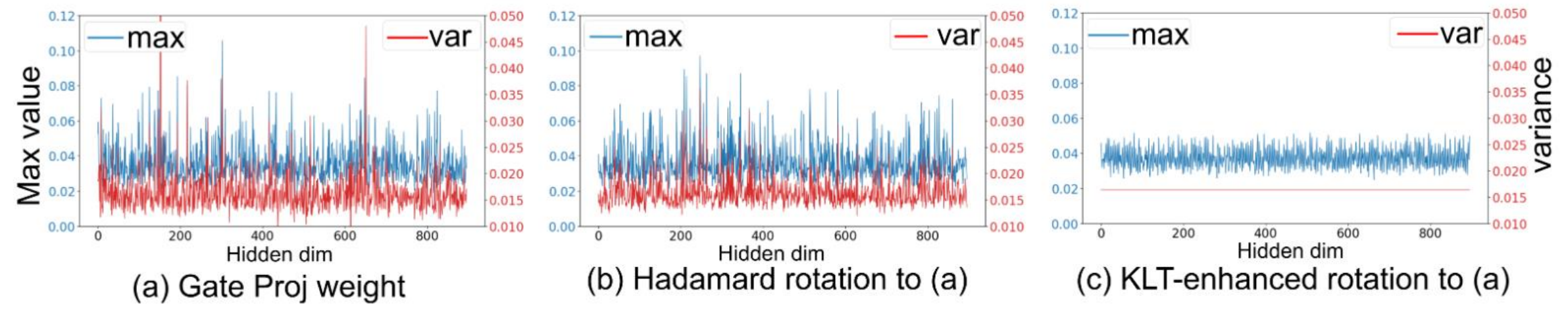


(a) Gate Projection (b) Output Projection (c) Matrix Multiplication

2. PScan amplifies the outliers



3. Hadamard rotation fails to align variance



(a) Gate Proj weight (b) Hadamard rotation to (a) (c) KLT-enhanced rotation to (a)

Method(Part I): Offline Rotation

KLT-Enhanced Rotation For Offline Transformation

➢Covariance matrix C_X of centered matrix X from calibration data

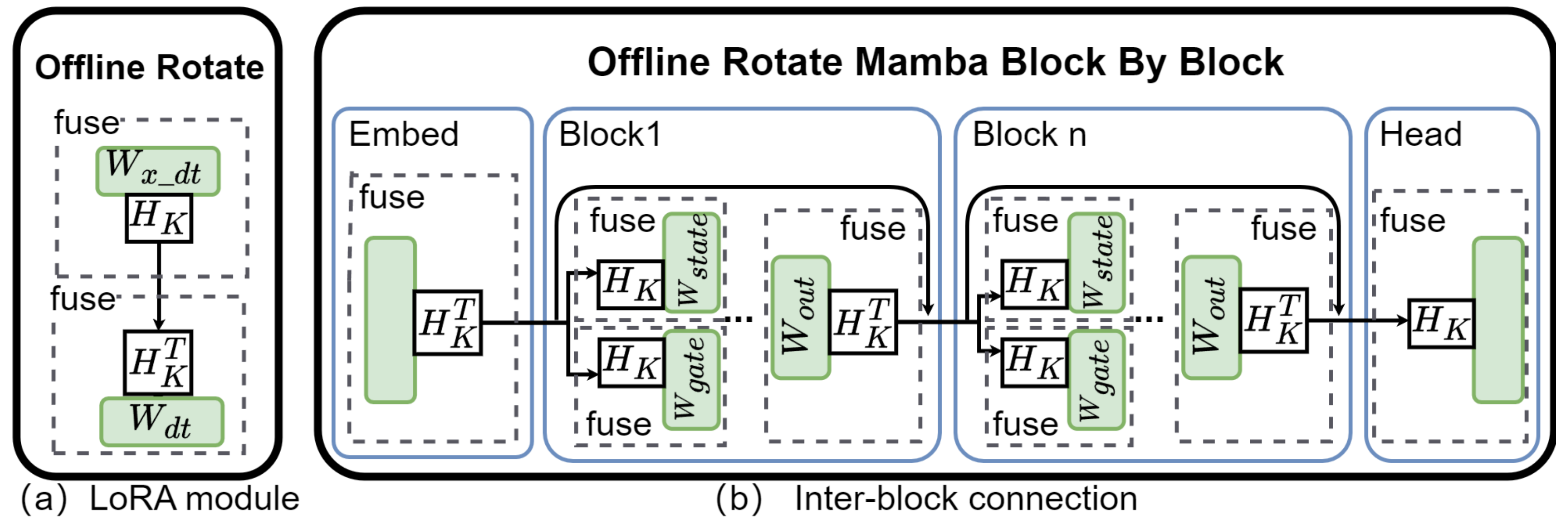
$$C_X = \frac{1}{n-1} X^T X = \frac{1}{n-1} K \Lambda K^T.$$

➢Apply KLT to Hadamard matrix H to get KLT - Enhanced rotation matrix H_K

$$H_K = KH,$$

$$C_{XH_K} = \frac{1}{n-1} H_K^T K \Lambda K^T H_K = \frac{1}{n-1} H^T K^T K \Lambda K^T K H = \frac{1}{n-1} H^T I \Lambda I H,$$

➢Offline transformation designs



(a) LoRA module (b) Inter-block connection

Method(Part II): Online Rotation

Smooth-Fused Rotation For Online Transformation

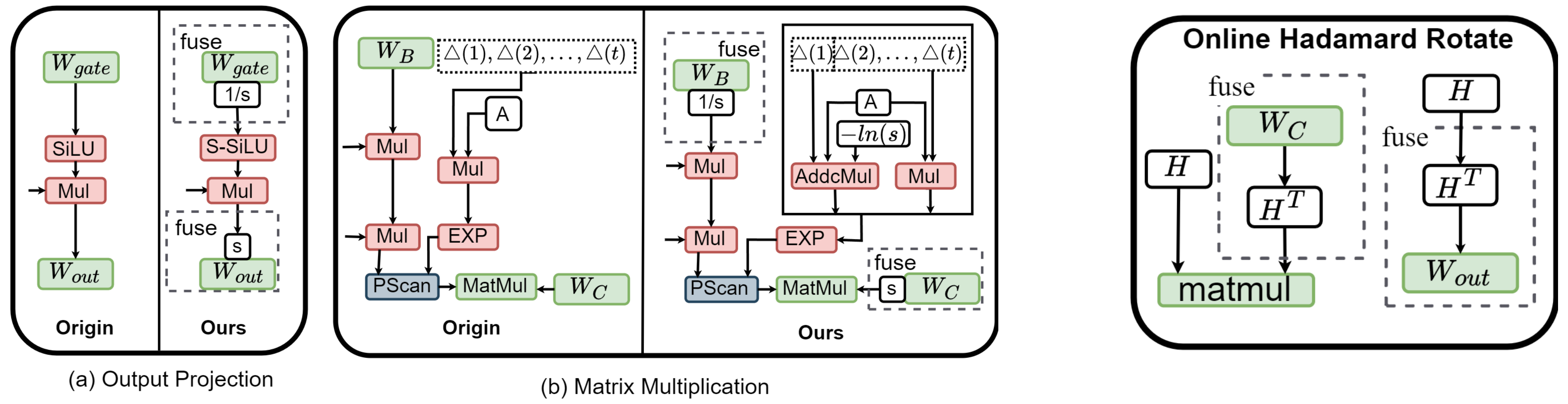
➢For the output projection layer: replace SiLU with S-SiLU to fuse parameter s.

$$\text{S-SiLU}(\boldsymbol{x}, s) = \boldsymbol{x} \odot \sigma(\boldsymbol{s} \odot \boldsymbol{x}),$$

$$\boldsymbol{y}_{out} = [\boldsymbol{y}_{ssm} \odot \text{SiLU}(\boldsymbol{x}_g \boldsymbol{W}_g)]\boldsymbol{W}_o = [\boldsymbol{y}_{ssm} \odot \text{S-SiLU}(\boldsymbol{x}_g \boldsymbol{W}_g', s_{out})]\boldsymbol{W}_o',$$

➢For the Matmul layer: use addcmul to absorb s passed through PScan

$$\text{addcmul}(-\ln(s_{mm}), \boldsymbol{\Delta}(1), \boldsymbol{A}) = \boldsymbol{A}\boldsymbol{\Delta}(1) - \ln(s_{mm}).$$



(a) Output Projection (b) Matrix Multiplication

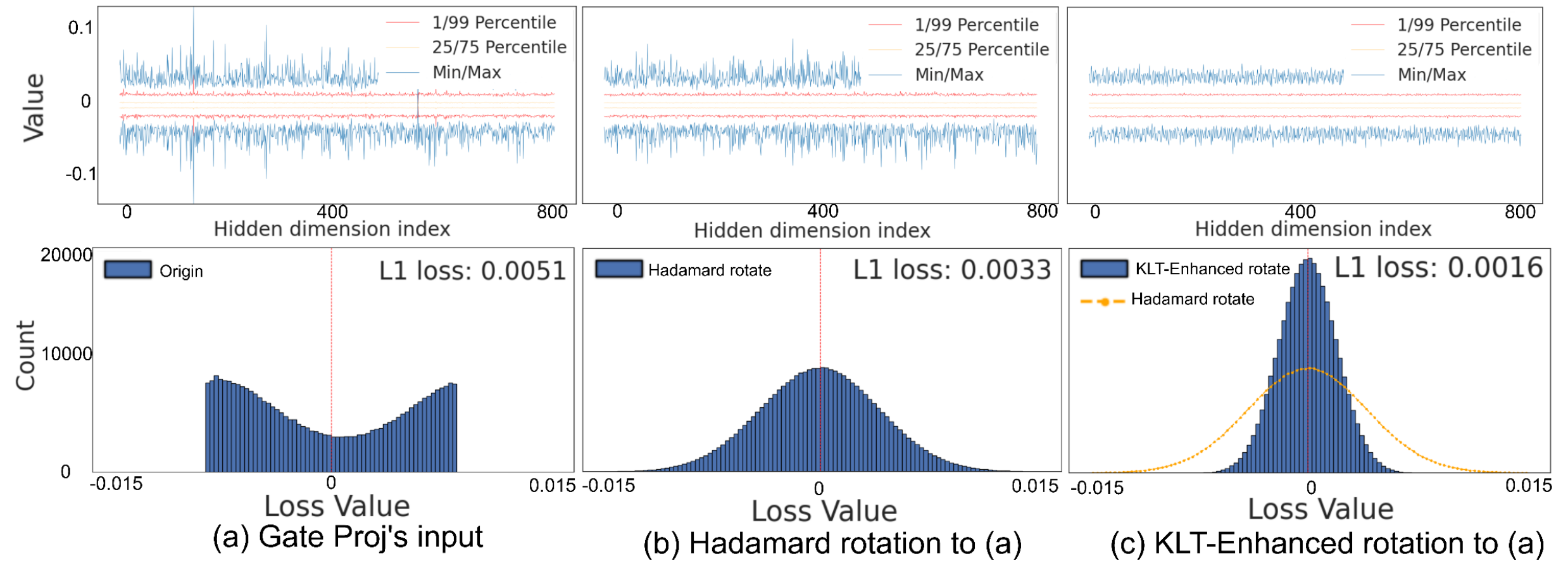
Experiments

Performance Comparison on Vision Model and Language Model

Bit Width	Methods	Vision Mamba					Mamba-ND			Mamba-LLM			
		Vim-T	Vim-T†	Vim-S	Vim-S†	Vim-B	mamba-2d S	Mamba-2d B	Mamba-3d	Mamba-370m	Mamba-790m	Mamba-1.4b	Mamba-2.8b
FP16	-	76.1	78.3	80.5	81.6	80.3‡	81.7	83.0	89.6	50.9	54.8	58.6	62.2
W8A8	RTN	37.4	32.4	68.8	68.8	52.2	80.3	82.2	87.9	45.7	44.9	53.9	58.4
	GPTQ+RTN	37.7	32.5	68.9	70.5	52.2	80.4	82.2	87.8	46.2	48.6	55.0	58.9
	SmoothQuant	37.7	32.3	68.7	72.9	52.1	80.3	82.2	87.9	45.2	41.7	54.2	58.7
	QuaRot	59.3	57.4	73.8	75.5	73.8	80.8	82.3	88.0	48.8	51.6	56.9	59.3
	Ours	75.6	77.8	80.3	81.4	80.1	81.2	82.8	89.0	50.0	53.8	58.3	62.1
W4A8	RTN	26.3	25.0	66.1	70.0	46.2	40.6	78.8	86.1	36.2	35.4	51.6	54.8
	GPTQ+RTN	30.4	27.9	66.5	70.6	47.7	60.3	78.9	86.8	36.7	36.0	51.1	53.6
	SmoothQuant	27.0	26.0	66.4	70.2	46.7	59.7	80.2	86.9	36.8	39.3	52.0	54.9
	QuaRot	52.7	48.5	72	74.0	72.8	80.1	82.0	86.9	43.4	40.0	53.8	58.5
	Ours	72.1	73.7	79.4	80.4	79.8	80.4	81.9	88.4	43.9	45.8	54.3	58.5

Ablation Experiment For KLT-Enhanced Rotation

Bit Width	Methods	Vim T†	Mamba-790m	Bit Width	Methods	Vim T†	Mamba-790m
FP16	-	78.3	54.8	FP16	-	78.3	54.8
W8A8	Baseline(RTN)	32.4	44.2	W4A8	Baseline(RTN)	25.0	35.4
	Hadamard Rotate	33.9(↑ 1.5)	50.8(↑ 6.6)		Hadamard Rotate	25.1(↑ 0.1)	40.2(↑ 4.8)
	KLT-Enhanced Rotate	47.7(↑ 15.3)	51.3(↑ 7.1)		KLT-Enhanced Rotate	38.9(↑ 3.9)	42.3(↑ 6.9)

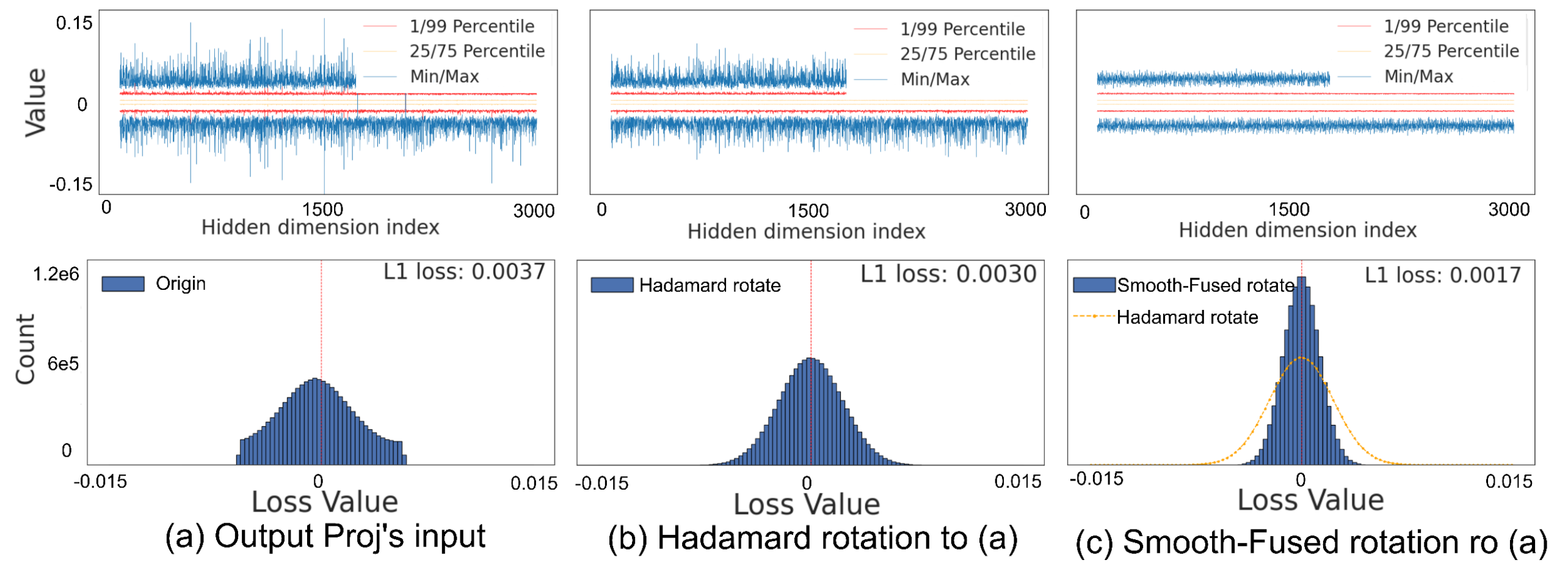


(a) Gate Proj's input (b) Hadamard rotation to (a) (c) KLT-Enhanced rotation to (a)

By comparing experiments with and without it in different models, it shows that KLT - Enhanced Rotation can balance channel variance. For example, in Vim's W4A8 setting, accuracy improves over 6%, validating its effectiveness in the MambaQuant framework.

Ablation Experiment For KLT-Enhanced Rotation

Bit Width	Methods	Vim-T†	Mamba-790M	Bit Width	Methods	Vim-T†	Mamba-790M
FP16	-	78.3	54.6	FP16	-	78.3	58.6
W8A8	Baseline(KLT-enhanced Rotation)	47.7	51.3	W4A8	Baseline(KLT-enhanced Rotation)	38.9	42.3
	Hadamard Rotation	69.7(↑ 22.0)	51.8(↑ 0.5)		Hadamard Rotation	62.0(↑ 23.1)	43.0(↑ 0.7)
	Smooth-Fused Rotation	77.8(↑ 30.1)	53.3(↑ 2.0)		Smooth-Fused Rotation	73.7(↑ 34.8)	45.8(↑ 3.5)



(a) Output Proj's input (b) Hadamard rotation to (a) (c) Smooth-Fused rotation ro (a)

It replaces SiLU with S-SiLU and absorbs the s - parameter. Experiments on models show Smoothed Rotation can equalize activation channel variances. It boosts quantized Mamba model performance, proving its worth in the framework.

Conclusion

- ➢ Unveiling the cause of performance drop in the quantization of the mamba model
- ➢**Mambaquant**: first general and effective quantization method for maba-based models
- ➢ **SOTA performance:** almost the same as the FP16 model in W8A8