PrinMix: Principled SVD-based Delta Compression via Quantization Error Minimization

Boya Xiong, Shuo Wang, Weifeng Ge, Guanhua Chen, Yun Chen

Shanghai University of Finance and Economics; Tsinghua University; Fudan University; SUSTech; MoE Key Lab of IRCE

ICML International Conference On Machine Learning

CORE IDEA **PrinMix** converts mixed-precision SVD delta compression from singular-value-based ranking into calibration-aware reconstruction-error minimization under a bit budget.

22.3% AIME2024 gain over Delta-CoMe, 7B	**26.9%** AIME2024 gain over Delta-CoMe, 14B	**6.1%** GQA gain over Delta-CoMe, 7B	**12** Qwen2.5-7B variants on one L20 GPU

1 Why Delta-Compression?

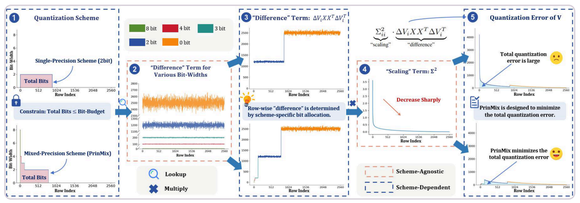
Multi-tenant serving often deploys many customized LLMs that share the same base model. Delta-compression stores one base model plus compressed delta weights, reducing memory and disk cost while preserving full fine-tuning behavior.

- **Full-model compression** still pays repeated storage for every customized model.
- **Existing delta methods** such as BitDelta and Delta-CoMe rely on empirical bit-allocation rules.
- **Hard cases** emerge when $\|\Delta W\|$ is large, such as AIME2024 and multimodal 7B backbones.

Question. Can mixed precision be chosen by minimizing the actual quantization error under a target bit budget?

2 Framework

PrinMix works in the SVD space of the delta matrix. It estimates quantization loss for candidate bit-widths, solves a constrained 0/1 optimization problem, and quantizes $\hat{V}$ and $\hat{U}$ with a shared mixed-precision scheme.



Pipeline. Given a bit budget, PrinMix derives a quantization scheme from loss terms rather than hand-written singular-value rules.

3 Core Evidence

Abs.	**22.3%** AIME2024 and **6.1%** GQA gains on 7B.
Main	**26.9%** AIME2024 gain on 14B reasoning models.
Tbl. 1-2	**2.9%** and **2.2%** relative average gains over Delta-CoMe.
Sys.	One L20 GPU supports **12** PrinMix variants versus **8** Delta-CoMe variants.

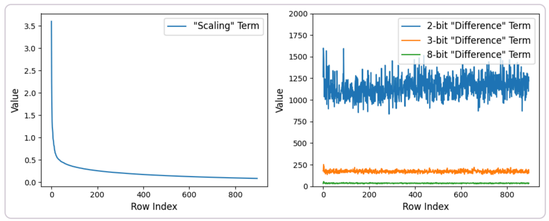
4 Why Mixed Precision?

For each row of V, the GPTQ-style loss decomposes into a singular-value scaling term and a quantization-difference term.

V-ROW LOSS

$$e_i^V = \underbrace{\Sigma_{ii}^2}_{\text{scaling}} \cdot \underbrace{\Delta V_i X X^T \Delta V_i^T}_{\text{difference}}$$

Uniform bit-widths waste budget because rows with large singular values dominate reconstruction loss.



Observation. The scaling term varies sharply across row index, creating a principled need for row-wise mixed precision.

5 Optimization Modeling

Mixed precision introduces a combinatorial allocation problem: choose one bit-width per SVD row/column pair while respecting a global bit budget.

LAYER ALLOCATION OBJECTIVE

$$\min_S \sum_i \sum_k S_{i,k}\, \mathbb{E}^{\mathbf{V}}_{i,k} \quad \text{s.t.} \quad \sum_i \sum_k S_{i,k} B_k \le G_b$$

$S_{i,k}$ selects bit-width k for row i of V and the corresponding column of $\tilde{U}$; $\mathbb{E}^{\mathbf{V}}_{i,k}$ is the corresponding quantization loss of V.

- **One-hot:** one selected bit-width per SVD component.
- **Budget:** total storage cannot exceed target G_b.
- **Deployment:** at most $f_{\max}$ active bit-widths.
- **Solver:** SCIP solves ILP; DP trades constraints for speed.

6 Reconstruction Target Correction

Since quantizing V first alters the optimization target for U, RTC updates U to $\tilde{U}$ prior to quantization to mitigate bias.

QUANTIZE V apply selected row bits > **CORRECT U** update target to $\tilde{U}$ > **QUANTIZE U** preserve ΔW reconstruction

RTC improves average performance by **2.2%** and AIME2024 by **13.1%** in ablation.

7 Main Results

7B method	AIME24	GQA	AVG
Aligned	40.0	60.5	76.0
BitDelta	0.0	0.0	40.5
Delta-CoMe	30.0	49.4	71.9
PrinMix	**36.7**	**52.4**	**74.0**

13-14B method	AIME24	GSM8K	AVG
Aligned	40.0	71.0	66.6
BitDelta	23.3	65.8	63.0
Delta-CoMe	24.5	70.2	63.3
PrinMix	**31.1**	**71.2**	**64.7**

+2.9% avg gain 7B	**+2.2%** avg gain 13-14B	**+22.3%** AIME24 7B	**+26.9%** AIME24 14B

8 Where PrinMix Helps Most

Gains concentrate on challenging settings where baselines are far from the aligned model. The paper links this to larger delta-weight norms: DeepSeek-R1-Distill-Qwen-7B has a median norm 6.5x that of Qwen-Coder-Instruct-7B.

- AIME2024 exposes the largest gap between empirical heuristics and loss-minimizing bit selection.
- MBPP and HumanEval are near-lossless for strong baselines, leaving less room for improvement.
- PrinMix supports arbitrary compression ratios; BitDelta and Delta-CoMe use fixed schemes in the paper.

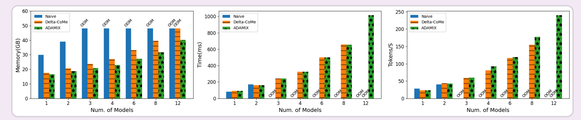
9 Error Evidence

Method	High all	High outlier
BitDelta	21.51	3162.58
Delta-CoMe	7.54	470.82
PrinMix	**6.81**	**426.20**

Conclusion. PrinMix lowers aggregate and outlier error, validating that PrinMix preserves more information.

10 Serving Cost and Latency

A Triton implementation evaluates end-to-end decoding latency for Qwen2.5-7B variants on one NVIDIA L20 GPU.



Deployment. The aligned baseline fits two variants; Delta-CoMe fits eight; PrinMix fits twelve with better generation speed.

2 aligned models on one L20	**8** Delta-CoMe variants	**12** PrinMix variants

11 Practical Advantages

- Quantization takes **1.1 hours** for 7B models and **2.4 hours** for 14B models on one L20 GPU.
- The Qwen2.5-Math-7B-Instruct optimization stage takes **29.4 minutes** and can be reduced by at least **4x** with a commercial solver and smaller solution space.
- PrinMix scales to 70B in about **8.4 hours** on one L40 GPU.
- Calibration sensitivity experiments show robust performance across tested setup choices.

Delta form	BitDelta	Delta-CoMe	PrinMix
Full tuning delta	Yes	Yes	**Yes**
LoRA delta	**No**	**No**	**Yes**

PrinMix optimizes the delta matrix itself, so the same principle extends to LoRA-style low-rank updates.

12 Takeaways

PROBLEM. Many fine-tuned LLMs share one base. | **METHOD.** Optimize mixed precision under a bit budget. | **RESULT.** Better hard-task accuracy than Delta-CoMe. | **SYSTEM.** More variants fit on one GPU.