Iowa State University

Department of Computer Science

Sadegh Jafari, Mohiuddin Bilwal, Fan Zhou, Brian Gelder, Ali Jannesari

SC26 CHICAGO NOV 15–20

APQF: Agentic Profiling-Guided Structured Pruning and Mixed-Precision Quantization with Adaptive Fine-Tuning

Introduction

Deep neural networks such as vision models are powerful but huge, which makes their inference slow and expensive to run, especially on edge devices. Compression techniques such as pruning, quantization, and knowledge distillation make them more efficient, but deciding how and where to compress a given model is still manual or relies on complex algorithms that keep the approach from being model-agnostic.

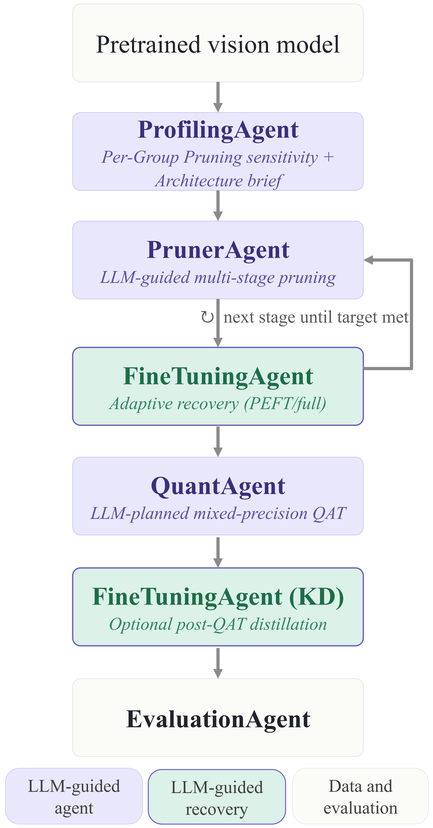
Problem

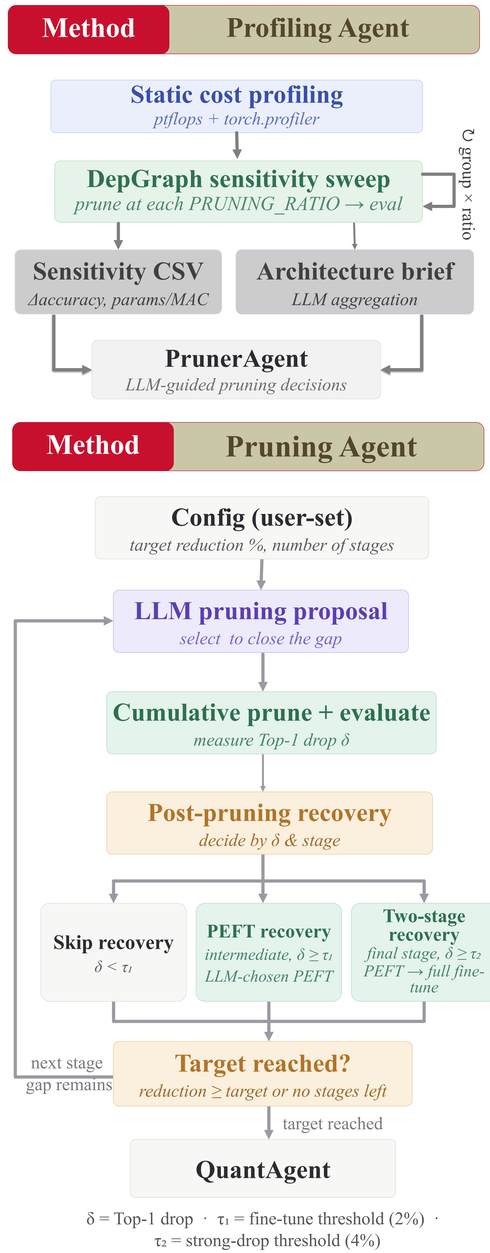
- Compression relies on hand-designed heuristics that rarely transfer across architectures
- Uniform pruning ratios and bit-widths ignore how much layer sensitivity varies
- Joint optimization retrains the whole network and needs large training data
- Profiling measures cost and redundancy, but does not drive the decisions

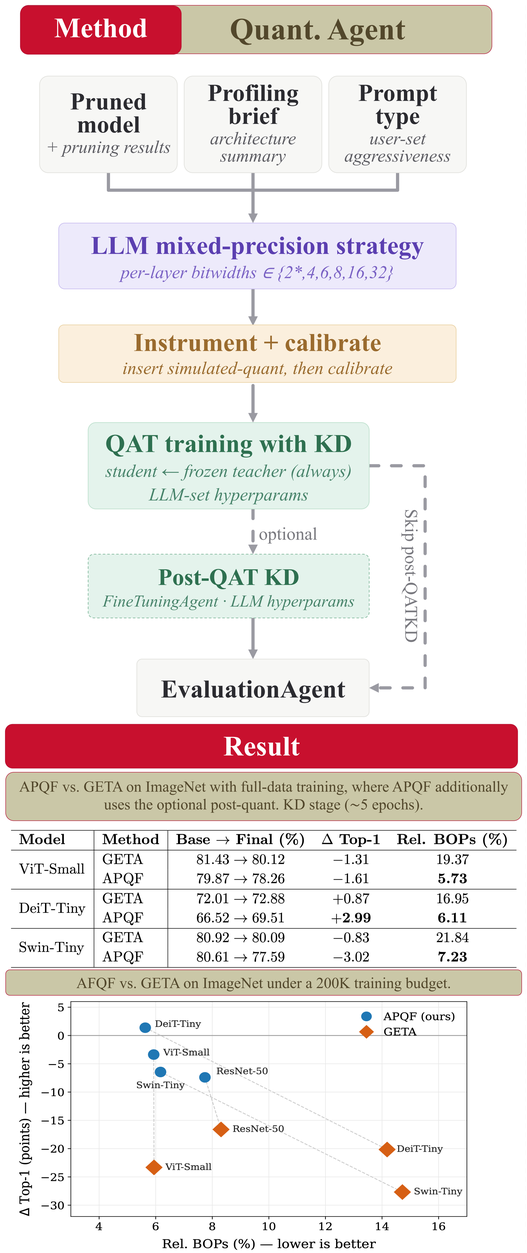
APQF

APQF is an agentic, profiling-guided framework that combines structured pruning, mixed-precision quantization, and accuracy recovery in one automated pipeline. A profiling agent measures how cost is distributed and how sensitive each part is to pruning, and LLM planners use that evidence to decide the compression.

- **Profiling agent**: sensitivity map and architecture brief ground all decisions
- **LLM-guided pruning**: per-group ratios, adaptive recovery
- **Mixed-precision QAT**: per-layer bit-widths with distillation
- **Architecture-agnostic**: one pipeline for CNNs and transformers
- **Data-efficient**: recovers a pretrained model, no retraining
- **LLM-agnostic**: runs on free or premium planners







Result

APQF vs. GETA on ImageNet with full-data training, where APQF additionally uses the optional post-quant. KD stage (~5 epochs).

Model	Method	Base → Final (%)	Δ Top-1	Rel. BOPs (%)
ViT-Small	GETA	81.43 → 80.12	−1.31	19.37
	APQF	79.87 → 78.26	−1.61	**5.73**
DeiT-Tiny	GETA	72.01 → 72.88	+0.87	16.95
	APQF	66.52 → 69.51	+**2.99**	**6.11**
Swin-Tiny	GETA	80.92 → 80.09	−0.83	21.84
	APQF	80.61 → 77.59	−3.02	**7.23**

AFQF vs. GETA on ImageNet under a 200K training budget.

VGG7 on CIFAR-10: comparison of quantization and joint pruning–quant. methods

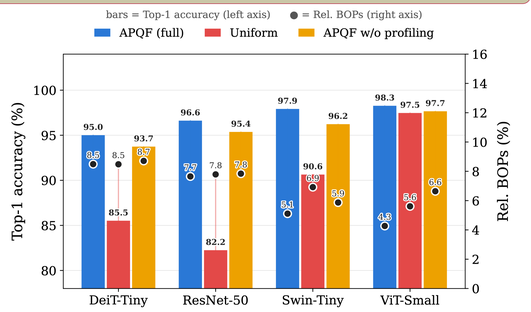
Method	Prune	Quant.	Acc. (%)	ΔAcc	Rel. BOPs (%)
Baseline	–	–	93.05	–	100.00
TWN Li et al. (2016)	×	✓ (ternary W, 2/32)	92.56	−0.49	6.24
LR-Net Shayer et al. (2017)	×	✓ (1-bit W)	93.18	+0.13	3.11
RQ (8/8) Louizos et al. (2018)	×	✓ (W+A, 8/8)	93.30	+0.25	6.24
RQ (4/4) Louizos et al. (2018)	×	✓ (W+A, 4/4)	92.04	−1.01	1.56
WAGE Wu et al. (2018)	×	✓ (W+A, 2/8)	93.22	+0.17	1.56
DQ Uhlich et al. (2019)	×	✓ (mixed W+A)	91.59	−1.46	0.48
DQ-restrict Uhlich et al. (2019)	×	✓ (mixed W+A)	91.59	−1.46	0.54
DJPQ Wang et al. (2020)	✓	✓ (mixed W+A)	91.54	−1.51	0.48
DJPQ-restrict Wang et al. (2020)	✓	✓ (mixed W+A)	91.43	−1.62	0.46
BB (μ=0.01) Van Baalen et al. (2020)	✓	✓ (mixed W+A)	93.23	+0.18	0.51
BB (μ=0.1) Van Baalen et al. (2020)	✓	✓ (mixed W+A)	91.96	−1.09	0.29
GETA Qu et al. (2025)	✓	✓ (mixed W+A)	92.57	−0.48	0.41
APQF (ours)†	✓	✓ (mixed W+A)	92.68	−0.27	0.36
APQF (ours)†	✓	✓ (mixed W+A)	93.15	+0.20	0.41

†APQF is compressed from our VGG7 reproduction (92.95% FP32); ΔAcc and Rel. BOPs are relative to that baseline.

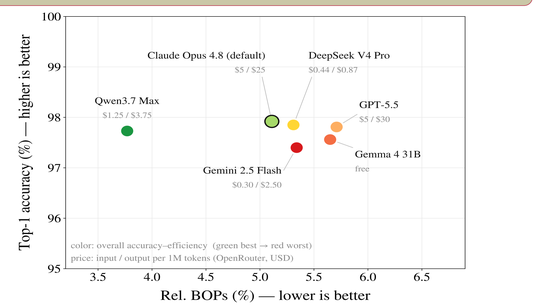
APQF vs. GETA on CIFAR-10. Δ is the accuracy change from each method's baseline

Model	Method	Base → Final (%)	Δ Top-1	Rel. BOPs (%)
DeiT-Tiny	GETA	91.75 → 91.05	−0.70	11.90
	APQF	91.90 → 95.01	+**3.11**	**8.49**
ResNet-50	GETA	96.55 → 93.09	−3.46	7.68
	APQF	95.83 → 96.62	+**0.79**	**7.65**
Swin-Tiny	GETA	97.20 → 91.46	−5.74	15.12
	APQF	96.78 → 97.92	+**1.14**	**5.11**
ViT-Small	GETA	98.51 → 92.16	−6.35	**3.59**
	APQF	98.46 → 98.27	−**0.19**	4.27
ResNet-20	GETA	91.70 → 91.42	−**0.28**	4.50
	APQF	92.12 → 91.55	−0.57	**2.90**

CIFAR-10 ablation at matched compute, four architectures. Bars are Top-1 accuracy, dots are Rel. BOPs. Uniform loses the most accuracy, removing profiling less.



LLM planner comparison (Swin-Tiny, CIFAR-10). One point per LLM, colored by accuracy–efficiency rank (green best, red worst) & labeled with its price per 1M tokens.



Qu, X., Aponte, D., Banbury, C., Robinson, D. P., Ding, T., Koishida, K., Zharkov, I., and Chen, T. (2025). Automatic joint structured pruning and quantization for efficient neural network training and compression. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 15234–15244.

Jafari, S., Sarkar, A., Bilwal, M., and Jannesari, A. (2025). Profilingagent: Profiling-guided agentic reasoning for adaptive model optimization. arXiv preprint arXiv:2509.05584.