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

Sadegh Jafari
Department of Computer Science
Iowa State University
Ames, IA, USA
sadeghj@iastate.edu

Mohiuddin Bilwal
Department of Computer Science
Iowa State University
Ames, IA, USA
moh596@iastate.edu

Fan Zhou
Department of Computer Science
Iowa State University
Ames, IA, USA
fanzhou@iastate.edu

Brian Gelder
Department of Agricultural And Biosystems Engineering
Iowa State University
Ames, IA, USA
bkgelder@iastate.edu

Ali Jannesari
Department of Computer Science
Iowa State University
Ames, IA, USA
jannesar@iastate.edu

***Abstract*—Deep neural networks are powerful but too large to run cheaply on edge devices. Pruning and quantization help, but deciding how and where to compress a model is still manual, or relies on algorithms that are hard to apply across architectures. We introduce APQF, an agentic profiling-guided framework that combines structured pruning, mixed-precision quantization-aware training, and accuracy recovery in one automated pipeline. A profiling agent measures cost distribution and pruning sensitivity, which then drive per-group pruning ratios, per-layer bit-widths, and the recovery strategy. Our experiments on ImageNet and CIFAR-10 show that APQF compresses convolutional and transformer models further than a joint pruning and quantization baseline at comparable or higher accuracy, and stays stable across different planners, including free open-weight ones.**

I. Introduction

Deep neural networks are expensive to adapt and deploy, yet compressing them through pruning and quantization is still largely manual, relying on heuristics that rarely transfer across architectures. Three gaps motivate this work. Uniform settings apply one pruning ratio and bit-width to every layer, ignoring that sensitivity varies widely. Joint optimization methods retrain the whole network and need large data budgets. And profiling measures where cost and redundancy lie, but drives inspection rather than compression.

II. Approach

We present APQF, an agentic profiling-guided framework combining structured pruning, mixed-precision quantization-aware training (QAT), and accuracy recovery in one automated pipeline (Fig. 1). Five agents carry this out, with LLMs planning from measured evidence rather than heuristics.

The profiling agent measures how parameters and operations are distributed across the model, and measures pruning sensitivity by pruning each dependency group at several ratios. This yields a per-group sensitivity map and an architecture brief.

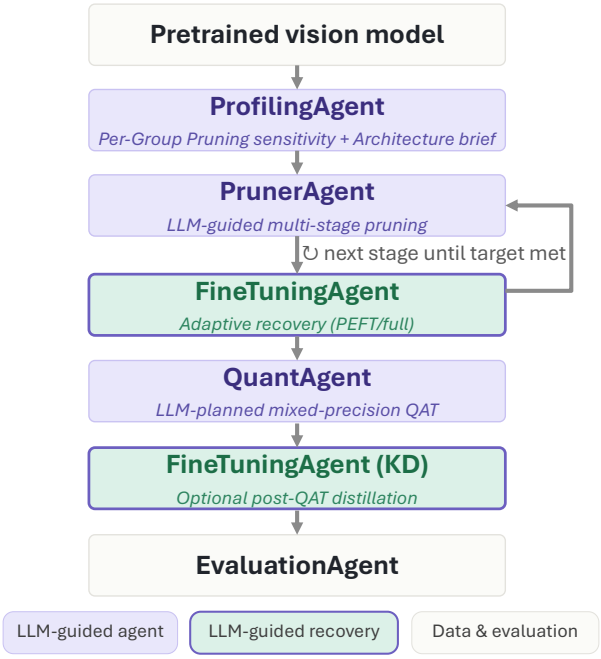


Fig. 1. Overall workflow of the APQF framework

The pruning agent uses that evidence to request a pruning strategy from the LLM at each of several stages. It builds on dependency-graph structured pruning, but instead of one global ratio the LLM assigns a per-group ratio. Decisions accumulate across stages and are replayed from the original model, while the fine-tuning agent recovers accuracy between stages so the planner sees what is recoverable. Recovery is skipped when the drop is small, uses parameter-efficient fine-

TABLE I
APQF VS. GETA ON IMAGENET, FULL-DATA TRAINING. APQF ADDS ∼5 KD EPOCHS.

Model	Method	Base → Final	Δ	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**

tuning at intermediate stages, and adds full fine-tuning when the final drop is large.

The quantization agent assigns a bit-width to each surviving layer, then trains the model under simulated quantization with distillation from a frozen full-precision teacher. Every proposed strategy is validated against the model before execution, and a deterministic evaluation agent measures every checkpoint on a fixed validation set.

III. RESULTS

Efficiency is reported as relative bit-operations (Rel. BOPs), the compressed model's bit-operations as a fraction of the FP32 original, capturing pruning and quantization together. We compare against GETA, a recent joint pruning and quantization method, run from its released configuration. Given the same 200K-image budget, GETA falls to 51–59% Top-1 while APQF retains 68–77% (Fig. 2). Both start from the same pre-trained weights, but GETA re-optimizes the whole network and needs large-scale data to converge, whereas APQF compresses from measured sensitivity and then recovers.

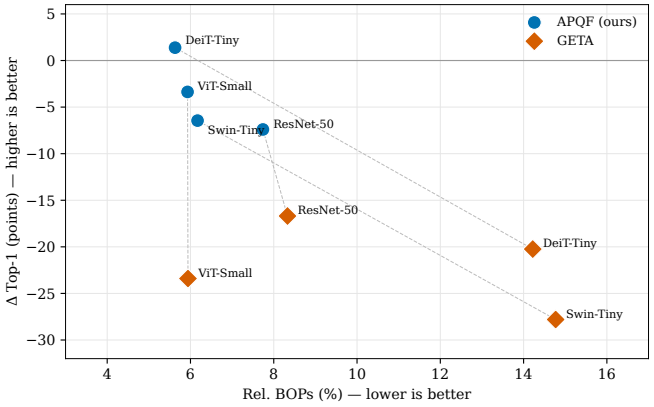


Fig. 2. Accuracy–compute trade-off under a 200K ImageNet budget.

With the full training set the gap narrows but does not close (Table I). GETA trains on all the data for its entire schedule, while APQF keeps its 200K-image model and adds only about five epochs of post-quantization distillation, yet reaches GETA's accuracy within a few points at three times lower Rel. BOPs.

On CIFAR-10 both methods start from identical fine-tuned checkpoints, so the baselines are matched (Table II). The small differences in base accuracy come from each method's own evaluation pipeline. APQF compresses further on four of the five architectures and ends more accurate than its starting point on three of them.

TABLE II
APQF VS. GETA ON CIFAR-10.

Model	Method	Base → Final	Δ	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**

Table III compares APQF with published joint pruning–quantization results on VGG7. APQF exceeds GETA's accuracy both at equal cost and at a lower cost, and it is the only method below 0.5% Rel. BOPs that improves on its own full-precision baseline. Bayesian Bits reaches marginally higher accuracy, but only by spending more bit-operations.

TABLE III
VGG7 ON CIFAR-10 VS. PUBLISHED JOINT METHODS.

Method	Acc. (%)	ΔAcc	Rel. BOPs (%)
FP32 baseline	93.05	–	100.00
DJPQ	91.54	−1.51	0.48
DJPQ-restrict	91.43	−1.62	0.46
BB (μ=0.01)	93.23	+0.18	0.51
BB (μ=0.1)	91.96	−1.09	**0.29**
GETA	92.57	−0.48	0.41
APQF†	92.68	−0.27	0.36
APQF†	**93.15**	**+0.20**	0.41

†Compressed from our VGG7 reproduction (92.95% FP32).

IV. SIGNIFICANCE AND ONGOING WORK

To our knowledge, APQF is the first agentic framework to combine structured pruning with mixed-precision quantization through LLM-guided, profiling-grounded decisions. Two ablations isolate where the gains come from. Replacing APQF's per-layer decisions with a single pruning ratio and one bit-width, at matched compute, costs the most accuracy. Keeping the LLM but withholding the profiling evidence is less damaging, yet still lowers accuracy and raises bit-operations on every model tested. It is also planner-agnostic, with six models from free to premium reaching comparable accuracy. APQF is therefore practical when data or budget is limited, and ongoing work extends it beyond image classification.