WHERE COMPRESSION HURTS: FUNCTION-SPACE SENSITIVITY FOR MIXED-PRECISION NEURAL NETWORK QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting layer-wise precision is a central combinatorial problem in mixed-precision post-training quantization and requires a reliable surrogate for the damage induced by each compression action. Prior work has studied this problem from several perspectives, focusing on how compression-induced perturbations affect model parameters, optimization objectives, or model outputs. We study a complementary question: how should a quantization direction be measured in output space, and what additional information is preserved by retaining its first-order tangent response rather than reducing it to a scalar sensitivity score? For a compression action inducing $\Delta W_{l,a}$, we analyze the first-order logit perturbation $J_l \Delta W_{l,a}$ and the associated $J_l^\top J_l$ quadratic form. The resulting sensitivity is specific to the chosen compression direction, is label-independent once the calibration inputs are fixed, and can be reused across different first-order analyses. Across ResNet-18, MobileNetV3-Small, and MobileNetV2, this function-space sensitivity provides an informative ranking of realized compression damage. Its local linear response closely tracks directly measured output distortion under isolated compression actions, while severity-controlled analyses show that the ranking cannot be explained solely by compression magnitude. We further find that the benefit of modeling tangent interference is network-dependent rather than universal. Under budgeted precision allocation, improved prediction of function-space distortion can also lead to better allocation decisions when compression actions are sufficiently distinguishable. These results identify Euclidean logit geometry as a reusable function-space complement to parameter-based, loss-based, and directly measured output-space sensitivity.

1 INTRODUCTION

Quantization is one of the most effective routes to efficient neural-network inference, reducing memory traffic and enabling low-precision arithmetic on commodity and specialized hardware (Han et al., 2016; Jacob et al., 2018; Banner et al., 2019). Uniform precision, however, allocates the same numerical budget everywhere, even though layers and individual compression actions can differ substantially in their tolerance to perturbation. Mixed-precision quantization exploits this heterogeneity by assigning different bit widths across a network, but in doing so converts quantization into a combinatorial selection problem. With L quantizable tensors and three candidate precisions, the search space grows as 3^L; in our setting, this corresponds to roughly 10^{10} assignments for ResNet-18 and 10^{25} for MobileNetV3-Small. Exhaustive evaluation is therefore computationally infeasible, and the quality of the final allocation depends critically on the surrogate used to predict compression damage.

Many works have addressed this problem from different perspectives. Hardware-aware and differentiable search methods learn bit assignments jointly with resource objectives (Wang et al., 2019; Wu et al., 2018; Cai & Vasconcelos, 2020; Uhlich et al., 2020); Hessian- and gradient-based methods rank layers using task-loss sensitivity (Dong et al., 2019; 2020; Chauhan et al., 2023); and PTQ methods reduce local reconstruction error through rounding, equalization, or block-wise optimization (Nagel et al., 2019; 2020; Li et al., 2021; Wei et al., 2022). More recent work also scores

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

final-model output distortion directly or through Jacobian-mediated approximations (Chen et al., 2021; Wang et al., 2026). These approaches are effective, but they operate under different geometries and retain different forms of information. This leaves a narrower but important direction, which motivates two questions: (1) when the same concrete quantization direction is propagated to the network output, how does the choice of output-space metric affect the resulting sensitivity? and (2) what additional information is retained when the vector-valued tangent response is preserved rather than immediately collapsed to a scalar cost?

To address these questions, we formulate quantization sensitivity directly in the function space of the pretrained network. We take the pretrained function itself as the object whose local stability should be measured. Quantizing layer l with action a produces the concrete perturbation $\Delta W_{l,a} = Q_a(W_l) - W_l$. Rather than asking whether that layer is important in the abstract, we propagate this exact perturbation through the network's logit Jacobian:

$$g_{l,a}(x) = J_l(x)\,\mathrm{vec}(\Delta W_{l,a}), \qquad S^F_{l,a} = \mathbb{E}_x\left[\|g_{l,a}(x)\|_2^2\right]. \tag{1}$$

This *functional sensitivity* corresponds to the pullback $\Delta w^\top \mathbb{E}[J^\top J]\Delta w$ of Euclidean output geometry. It is action-specific, label-independent once the calibration inputs are fixed, and computable without explicitly materializing the Jacobian by using Jacobian–vector products (JVPs). Retaining the tangent responses $g_{l,a}$ additionally reveals how first-order perturbations from different layers reinforce or cancel one another. The same responses can therefore support both additive mixed-precision scoring and first-order tangent-interference analysis.

This perspective separates three notions that are often conflated in compression: parameter displacement, task-loss displacement, and function displacement. These notions can be expressed through different geometries within the general form

$$S_M(\Delta w) = \Delta w^\top J^\top M J \Delta w. \tag{2}$$

Loss-aware Gauss–Newton constructions use a loss-dependent output metric M, whereas our criterion sets $M = I$ and directly measures how far the pretrained logits move under the specified compression direction. This distinction leads to the central empirical questions of the paper: Does compression-direction function displacement predict realized mixed-precision damage better than parameter- or loss-space alternatives? Is its local linearization faithful to the nonlinear effect of quantization? When does first-order tangent interference matter? And does the resulting signal improve constrained bit allocation? We answer these questions on ResNet-18 and MobileNetV3-Small, with MobileNetV2 providing additional architecture-level validation. Our main contributions are:

- **Euclidean output-space sensitivity.** We study the $M = I$ member of the general $J^\top M J$ family along the actual quantization direction and retain the underlying JVP responses, allowing the same profiles to support multiple downstream analyses.
- **Geometric and local-fidelity comparisons.** We compare parameter error, gradient sensitivity, directional Hessian curvature, cross-entropy generalized Gauss–Newton geometry, and directly measured isolated-output fragility. The results show that function-space and loss-space metrics capture distinct notions of compression damage and that the JVP approximation is highly faithful in the local regime.
- **Cross-architecture tangent-interference analysis.** We decompose the squared norm of the summed first-order response into additive and pairwise cross terms, showing that these interactions provide substantial benefit on MobileNetV3-Small but are not universally beneficial across architectures. We further examine the transfer of additive prediction to MobileNetV2.
- **Decision-level validation.** We use the resulting layer–action costs in budget-constrained mixed-precision allocation and evaluate the selected configurations on the full validation set. This exposes both regimes in which functional costs improve allocation decisions and regimes in which improved ranking does not necessarily translate into a better constrained allocation.

2 Related Work

Quantization and post-training adaptation. Neural-network quantization has progressed from trained low-precision representations and integer-only inference (Han et al., 2016; Jacob et al., 2018)

to practical low-bit PTQ and quantization-aware training. Analytical clipping and bias correction demonstrated 4-bit PTQ (Banner et al., 2019), while data-free equalization and bias correction improved deployment without retraining (Nagel et al., 2019). Learned quantizers such as LSQ optimize quantizer parameters during training (Esser et al., 2020). In PTQ, AdaRound optimizes discrete rounding choices (Nagel et al., 2020), BRECQ reconstructs network blocks rather than isolated layers (Li et al., 2021), and QDrop regularizes low-bit reconstruction by randomly dropping activation quantization (Wei et al., 2022). ZeroQ further shows that calibration and mixed-precision selection can be performed without access to the original data by synthesizing statistics-matched inputs (Cai et al., 2020). These works primarily improve the quantizer, calibration process, or reconstruction objective; our focus is the complementary question of how to score a fixed candidate compression action.

Mixed-precision search and allocation. Mixed precision introduces a discrete bit-allocation problem on top of quantization. HAQ uses reinforcement learning with hardware latency/energy feedback (Wang et al., 2019); DNAS and EdMIPS relax discrete bit choices into differentiable architecture-search problems (Wu et al., 2018; Cai & Vasconcelos, 2020); and Uhlich et al. learn bit width indirectly through differentiable quantizer parametrization (Uhlich et al., 2020). Proxy-based alternatives reduce search cost through attribution preservation or network orthogonality, as in GMPQ and OMPQ (Wang et al., 2021; Ma et al., 2023). Such methods optimize allocation efficiently, whereas our focus is the sensitivity representation itself.

Sensitivity, curvature, and inter-layer dependence. A second line of work treats mixed precision as sensitivity estimation. HAWQ ranks layers using Hessian information, and HAWQ-V2 uses average Hessian trace and Pareto selection (Dong et al., 2019; 2020); first-order alternatives use gradients to reduce estimation cost (Chauhan et al., 2023). Chen et al. derive a second-order constrained formulation and reduce bit assignment to a multiple-choice knapsack problem (Chen et al., 2021). Their analysis contains a Jacobian-mediated loss geometry of the form $J^\top M_\ell J$, so neither Jacobians nor knapsack allocation are novel here. Schaefer et al. further show that layer-wise Hessian scores can miss inter-layer dependencies and explicitly augment the sensitivity model with such information (Schaefer et al., 2023). Our distinction is geometric and diagnostic rather than the use of Jacobians itself: we isolate $M = I$ to measure Euclidean logit displacement for the *actual* quantization direction, compare it directly with loss-weighted geometries, and retain cached tangent responses so first-order cross terms can be inspected explicitly.

Output fidelity and function-space views. Output-preserving objectives appear throughout compression, including reconstruction, distillation, and recent fragility-based mixed precision. MixFrag directly quantizes isolated components and measures output-distribution KL before mixed-precision allocation (Opi et al., 2026). Closest to our formulation, BLOB-Q minimizes final-model output distortion for ViTs using Jacobian-projected quantization perturbations, an additive approximation, and dynamic programming over weight/activation bit widths (Wang et al., 2026); Chen et al. likewise derive Jacobian-mediated loss geometry in a constrained mixed-precision formulation (Chen et al., 2021). Therefore neither final-output sensitivity nor propagation of quantization directions through Jacobians is claimed as novel here. Our contribution is to separate Euclidean logit geometry from loss-weighted geometry under the same perturbations, test local linear fidelity against direct nonlinear fragility, and retain vector-valued tangent responses to study first-order cross-layer interference. More broadly, Jacobian-induced metrics are classical in analyses of neural-network function geometry (Jacot et al., 2018).

3 METHOD

Problem setup. Let a pretrained classifier be $f_\theta : \mathcal{X} \to \mathbb{R}^C$, where $f_\theta(x)$ denotes the pre-softmax logits. We partition the quantizable weights into L tensors,

$$\theta_q = \{W_1, \ldots, W_L\}. \tag{3}$$

For each layer l, a compression action $a \in \mathcal{A}$ applies a deterministic quantizer Q_a and induces

$$\Delta W_{l,a} = Q_a(W_l) - W_l. \tag{4}$$

In our experiments, $\mathcal{A} = \{\mathrm{FP32}, \mathrm{INT8}, \mathrm{INT4}\}$. A mixed-precision configuration is $\mathbf{a} = (a_1, \ldots, a_L)$.

The goal is to estimate the degradation caused by $\mathbf{a}$ without evaluating $f_{\theta+\Delta\theta(\mathbf{a})}$ for every candidate. Because our downstream use is configuration selection, we primarily care about preserving the *ranking* of configurations by true distortion.

From parameter perturbation to function perturbation. For an input x, a first-order expansion around the pretrained parameters gives

$$f_{\theta+\Delta\theta}(x) - f_\theta(x) \approx J_\theta(x)\Delta\theta, \tag{5}$$

where $J_\theta(x)$ is the Jacobian of the logits with respect to the parameters. Restricting the perturbation to layer l yields

$$g_{l,a}(x) = J_l(x)\,\mathrm{vec}(\Delta W_{l,a}), \qquad J_l(x) = \frac{\partial f_\theta(x)}{\partial \mathrm{vec}(W_l)}. \tag{6}$$

We call $g_{l,a}(x)$ the *functional response* of action a at layer l.

Given a calibration set $\mathcal{C} = \{x_i\}_{i=1}^N$, we define the layer–action functional sensitivity as

$$S^F_{l,a} = \frac{1}{NC}\sum_{i=1}^{N} \|J_l(x_i)\,\mathrm{vec}(\Delta W_{l,a})\|_2^2\,. \tag{7}$$

The normalization by C only fixes the scale and has no effect on ranking. Equation 7 has the equivalent quadratic-form interpretation

$$G_l = \frac{1}{NC}\sum_{i=1}^{N} J_l(x_i)^\top J_l(x_i), \tag{8}$$

so that

$$S^F_{l,a} = \mathrm{vec}(\Delta W_{l,a})^\top G_l \mathrm{vec}(\Delta W_{l,a}). \tag{9}$$

Thus, G_l induces a local *function-space metric* on parameter perturbations. While raw weight error uses the Euclidean metric I, our criterion weights each perturbation direction according to how strongly it affects the network output over the data distribution.

Additive and first-order tangent-interference scores. For a configuration $\mathbf{a}$, the simplest approximation neglects cross-layer coupling:

$$S^F_{\mathrm{ind}}(\mathbf{a}) = \sum_{l=1}^{L} S^F_{l,a_l}. \tag{10}$$

This additive form is especially attractive for allocation because each layer–action pair can be preprofiled once and reused.

We also study the first-order response of the simultaneous perturbation:

$$S^F_{\mathrm{int}}(\mathbf{a}) = \frac{1}{NC}\sum_{i=1}^{N} \left\| \sum_{l=1}^{L} J_l(x_i)\,\mathrm{vec}(\Delta W_{l,a_l}) \right\|_2^2. \tag{11}$$

Expanding gives

$$S^F_{\mathrm{int}}(\mathbf{a}) = S^F_{\mathrm{ind}}(\mathbf{a}) + \frac{2}{NC}\sum_{i=1}^{N}\sum_{l<k} g_{l,a_l}(x_i)^\top g_{k,a_k}(x_i). \tag{12}$$

The second term measures whether the *first-order tangent responses* reinforce or cancel one another. We call this tangent interference. It does not include mixed second derivatives of the nonlinear network and therefore should not be interpreted as the full cross-layer coupling of the jointly quantized model.

Efficient computation with JVPs. Directly materializing $J_l(x)$ is infeasible. Equation 6, however, requires only the product of the Jacobian with the known direction $\Delta W_{l,a}$. We therefore compute each response using forward-mode automatic differentiation as a Jacobian–vector product. For every layer–action pair, we store the resulting $N \times C$ tangent logits and the scalar score in Eq. 7. Once these profiles are computed, arbitrary mixed-precision configurations can be scored without another forward pass through a quantized model.

Profiling scales with the number of candidate layer–action pairs rather than with the exponentially large number of configurations. The stored tangent vectors support both the additive score in Eq. 10 and the tangent-interference score in Eq. 11.

Relationship to parameter-, loss-, and output-space criteria. For a fixed compression perturbation Δw, different sensitivity criteria measure its magnitude under different geometries. The simplest is the parameter-space displacement

$$S_{\text{weight}}(\Delta w) = \|\Delta w\|_2^2, \tag{13}$$

which treats all parameter directions isotropically.

More generally, a quadratic geometry on the first-order output response can be written as

$$S_M(\Delta w) = \mathbb{E}_x\left[(J(x)\Delta w)^\top M(x)(J(x)\Delta w)\right]. \tag{14}$$

Our functional criterion corresponds to the Euclidean logit geometry $M(x) = I/C$.

For softmax cross-entropy, the output-space Hessian is

$$H_f(x) = \operatorname{diag}(p(x)) - p(x)p(x)^\top, \qquad p(x) = \operatorname{softmax}(f_\theta(x)), \tag{15}$$

and the corresponding generalized Gauss–Newton score is

$$S_{\text{GGN}}(\Delta w) = \mathbb{E}_x\left[(J(x)\Delta w)^\top H_f(x)(J(x)\Delta w)\right]. \tag{16}$$

Thus, S_F and S_{GGN} propagate the same compression direction through the same Jacobian but weight the resulting output displacement differently: S_F uses Euclidean logit geometry, whereas S_{GGN} uses the local geometry induced by the task loss.

For comparison, the directional parameter-Hessian criterion is

$$S_H(\Delta w) = \frac{1}{2}\Delta w^\top H_\theta \Delta w. \tag{17}$$

Unlike the generalized Gauss–Newton term, the full parameter Hessian may also contain contributions from second derivatives of the network itself. We therefore treat it as a separate loss-space comparator rather than identifying it directly with Eq. 14.

Our choice of Euclidean pre-softmax logit geometry targets fidelity to the pretrained representation. It is not invariant to common shifts of all logits, so we interpret it as one explicit function-space geometry rather than a unique metric on predictive distributions. This distinction also explains why S_F and loss-based curvature can rank the same perturbation differently: output directions with equal Euclidean magnitude need not incur equal local task loss. Equation 14 provides the conceptual connection to prior constrained mixed-precision formulations (Chen et al., 2021).

Quantization operator. We use deterministic symmetric per-output-channel weight quantization with candidate precisions $\{\text{FP32}, \text{INT8}, \text{INT4}\}$. Activations remain full precision, and no fine-tuning is performed, isolating the sensitivity criterion from training and hardware-kernel effects. The exact quantizer is specified in Appendix C.

Budget-constrained mixed-precision allocation. Let n_l be the number of parameters in layer l and $b(a_l)$ the bit width assigned by action a_l. We define the parameter-weighted average precision as

$$\bar{b}(\mathbf{a}) = \frac{\sum_{l=1}^{L} n_l b(a_l)}{\sum_{l=1}^{L} n_l}. \tag{18}$$

Using the additive functional score, mixed-precision selection becomes

$$\begin{aligned} \min_{\mathbf{a}} \quad & \sum_{l=1}^{L} S^{F}_{l,a_l} \\ \text{s.t.} \quad & \sum_{l=1}^{L} n_l b(a_l) \le B \sum_{l=1}^{L} n_l, \\ & a_l \in \mathcal{A}, \quad l = 1, \dots, L. \end{aligned} \tag{19}$$

This is a multiple-choice knapsack problem in which exactly one precision is selected for each layer. We solve it with dynamic programming using a 10^{-3}-bit budget resolution. The additive form gives a globally optimal solution under the discretized budget for the chosen sensitivity objective.

Because Eq. 11 is non-additive, it does not admit the same decomposition. For the tangent-interference ablation, we initialize from the additive optimum and apply one- and two-layer local changes that reduce the full first-order interference objective while respecting the budget. We therefore treat this allocation as a local refinement rather than as a globally optimal solution.

4 Experimental Setup

We evaluate ImageNet-pretrained ResNet-18, MobileNetV3-Small, and MobileNetV2 on Imagenette. ResNet-18 and MobileNetV3-Small are used for the primary predictor and allocation experiments, while MobileNetV2 provides an external architecture check. The classifier heads are left unchanged, and the ten Imagenette classes are evaluated through their corresponding ImageNet-1K output indices.

We study weight-only post-training quantization with $\{\mathrm{FP32}, \mathrm{INT8}, \mathrm{INT4}\}$ candidate precisions; activations remain in full precision and no retraining or fine-tuning is performed. Functional profiles use 64 calibration images for ResNet-18 and 32 for both MobileNet models. Predictor validation uses 512 disjoint evaluation images, while selected ResNet-18 and MobileNetV3-Small allocations are re-evaluated on the complete 3,925-image Imagenette validation split.

For predictor evaluation, we use 125 heterogeneous mixed-precision configurations per architecture, with the three uniform-precision configurations treated separately. Compression damage is measured using teacher–student KL divergence, logit MSE, cross-entropy change, and Top-1 change. Spearman's ρ is the primary ranking statistic, with paired bootstrap confidence intervals for predictor comparisons. Complete dataset, preprocessing, quantization, sampling, and reproducibility details are provided in Appendix A.

5 Results

5.1 Function-space sensitivity predicts realized compression distortion

Table 1 reports Spearman correlations on the 125 heterogeneous configurations. On ResNet-18, additive functional sensitivity is nearly monotonic with functional distortion: $\rho = 0.991$ for teacher–student KL and 0.988 for logit MSE. It also predicts downstream cross-entropy and Top-1 degradation substantially better than parameter-space or first-order loss baselines. MobileNetV3-Small is more challenging, but the same pattern remains: the additive functional score outperforms gradient-square and weight-space criteria for all four targets.

The Hessian comparison is particularly informative because it separates function geometry from loss curvature. On ResNet-18, additive functional sensitivity exceeds the directional block-Hessian correlation by 0.050 for KL, 0.048 for logit MSE, and 0.046 for cross-entropy change. Paired 95% bootstrap intervals are $[0.031, 0.080]$, $[0.030, 0.077]$, and $[0.013, 0.090]$, respectively, excluding zero. For Top-1 degradation, the difference is smaller (0.027) and its interval $[-0.007, 0.071]$ crosses zero. We therefore do not claim a statistically reliable Top-1 ranking advantage over the Hessian baseline.

To test whether configuration-level correlations merely recover overall compression severity, we additionally stratify the MobileNetV3-Small heterogeneous configurations into tertiles using summed

Table 1: Spearman rank correlation on 125 heterogeneous configurations for the two primary architectures. "Func.-ind." is additive; "Func.-int." includes first-order tangent-interference cross terms. Hessian results are reported for ResNet-18.

Architecture	Predictor	KL	Logit MSE	ΔCE	ΔTop-1
ResNet-18	Weight SSE	0.223	0.237	0.160	0.087
	Relative SSE	0.486	0.518	0.411	0.339
	Grad.-square	0.623	0.654	0.507	0.404
	Block Hessian	0.941	0.940	0.893	0.846
	Func.-ind.	**0.991**	**0.988**	**0.939**	**0.873**
MobileNetV3-S	Weight SSE	0.184	0.161	0.170	0.178
	Relative SSE	0.159	0.141	0.142	0.171
	Grad.-square	0.635	0.701	0.612	0.589
	Func.-ind.	0.705	0.762	0.680	0.662
	Func.-int.	**0.776**	**0.802**	**0.752**	**0.726**

Table 2: Strong output-space comparators on 125 heterogeneous configurations. Values are Spearman correlations; higher is better. "Direct KL/logit" sums directly measured isolated layer–action distortions.

Architecture	Predictor	KL	Logit MSE	ΔCE	ΔTop-1
ResNet-18	Func.-ind.	0.991	0.988	0.939	0.873
	CE-GGN-ind.	0.984	0.973	0.943	0.888
	Direct KL	0.992	0.981	0.955	0.897
	Direct logit	0.991	0.989	0.943	0.875
MobileNetV3-S	Func.-ind.	0.705	0.762	0.680	0.662
	CE-GGN-ind.	0.722	0.766	0.697	0.681
	Direct KL	0.805	0.789	0.806	0.803
	Direct logit	0.796	0.799	0.784	0.773

relative weight SSE, a separate parameter-space severity variable that does not use the functional score. Within low/medium/high severity bins, tangent-interference KL correlations remain $0.663/0.760/0.828$ in the primary configuration sample and $0.724/0.810/0.812$ in an alternate sample; the corresponding logit-MSE correlations are $0.757/0.778/0.825$ and $0.775/0.817/0.819$. In both samples, paired low-versus-high bootstrap intervals cross zero for all four targets, indicating that the ranking signal is not explained solely by moving from mild to severe compression (Appendix D).

5.2 Output geometry and direct isolated-output fragility

The Jacobian alone does not determine a unique sensitivity criterion: the output-space metric in Eq. equation 14 also matters. Table 2 compares the additive Euclidean functional score with cross-entropy GGN geometry and with a stronger nonlinear comparator that explicitly evaluates each isolated compressed layer. On ResNet-18, Euclidean functional sensitivity is significantly stronger than CE-GGN for the primary additive KL and logit targets: the paired Spearman differences are $+0.0067$ (95% CI $[0.0006, 0.0177]$) and $+0.0148$ ($[0.0060, 0.0310]$), respectively. CE-GGN is significantly stronger for Top-1 degradation, while the cross-entropy difference is not significant. On MobileNetV3-Small, all additive functional-versus-GGN bootstrap intervals cross zero. Hence neither $M = I$ nor $M = H_f$ universally dominates; they weight different notions of output damage. Full additive/tangent-interference geometry results and paired bootstrap intervals are reported in Appendix E.

The direct baseline tests whether the first-order JVP is faithful locally. Across non-FP32 layer–action pairs, functional sensitivity correlates almost perfectly with directly observed isolated damage: on ResNet-18, Spearman correlations are 0.992 with isolated KL and 0.999 with isolated logit MSE; on MobileNetV3-Small they are 0.993 and 0.999. The harder problem is composition across many simultaneously quantized layers. On ResNet-18, functional and direct isolated-logit predictors are statistically indistinguishable for KL/logit ranking and select the same effective 9–11 bit allocations

Table 3: Three-architecture functional-sensitivity summary on 125 heterogeneous configurations. R_{abs} measures departure from additive scoring. Tangent-interference scoring improves MobileNetV3-Small but not ResNet-18 or MobileNetV2.

Architecture	Func.-ind. KL	Func.-int. KL	Func.-ind. logit	Func.-int. logit	Median R_{abs}
ResNet-18	0.991	0.982	0.988	0.991	0.069
MobileNetV3-Small	0.705	**0.776**	0.762	**0.802**	0.140
MobileNetV2	**0.935**	0.928	**0.904**	0.898	0.096

on the 512-image E1 subset. On MobileNetV3-Small, summing direct isolated scores is stronger, while tangent-interference functional scoring recovers much of this gap (Appendix F).

5.3 Tangent-interference gains are not universal across architectures

On ResNet-18, explicitly summing tangent vectors before squaring provides no consistent advantage over the additive approximation. MobileNetV3-Small behaves differently: tangent-interference sensitivity improves KL rank correlation from 0.705 to 0.776 and logit-MSE correlation from 0.762 to 0.802. MobileNetV2 supplies an external check: additive functional sensitivity remains strong ($\rho = 0.935$ for KL and 0.904 for logit MSE), while the tangent-interference score is slightly lower (0.928 and 0.898), with paired bootstrap intervals crossing zero for all four targets (Appendix G).

Table 3 summarizes the three-architecture picture. Additive functional sensitivity remains strong on ResNet-18 and MobileNetV2, whereas MobileNetV3-Small is the only tested architecture for which the tangent-interference score yields a clear improvement. The median departure from additivity, R_{abs}, is 0.069, 0.140, and 0.096 for ResNet-18, MobileNetV3-Small, and MobileNetV2, respectively. Thus functional sensitivity generalizes to a third architecture, while the value of first-order cross-layer tangent interference is network- and target-dependent rather than universal.

5.4 Functional rankings stabilize with few calibration inputs

Mean ranking quality saturates quickly on both architectures. ResNet-18 reaches mean KL Spearman 0.978 with four calibration examples and 0.990 with 32, compared with 0.991 at 64. MobileNetV3-Small changes from 0.699 at four examples to 0.705 at 32, with larger subsets mainly reducing variance. Complete results are in Appendix C.1.

5.5 Decision-level validation

We next ask whether the sensitivity scores support useful constrained decisions. Selected configurations are re-evaluated on all 3,925 Imagenette validation images. Table 4 highlights the most informative operating points; complete budget sweeps are reported in Appendix B. On ResNet-18, the 8-bit functional, Hessian, and gradient-square objectives select the same mixed-precision configuration and are all worse than uniform INT8, showing that mixed precision is not automatically superior. At 11 bits, however, functional allocation reduces KL from 4.47×10^{-4} to 3.48×10^{-4} and logit MSE from 1.59×10^{-3} to 1.21×10^{-3} relative to the Hessian allocation while using fewer average weight bits. On MobileNetV3-Small, tangent-interference refinement helps at budget 8, but additive allocation is slightly better at budgets 9–10 and tangent interference is modestly better again at 11. Thus predictor quality and constrained decision quality are related but not equivalent.

6 Discussion

The results separate three aspects of mixed-precision sensitivity: the geometry used to measure a fixed compression perturbation, the fidelity of its local linear approximation, and the way perturbations compose across layers. Both Euclidean logit displacement and CE-GGN geometry are substantially more predictive of realized compression damage than raw parameter-space error, but neither dominates across all targets. This indicates that function displacement and task-loss curvature capture related but distinct notions of sensitivity.

Table 4: Representative full-validation allocation checkpoints. Lower KL and logit MSE are better; average bits are parameter weighted. Full budget sweeps are in Appendix B.

Architecture / budget	Method	Avg. bits	KL	Logit MSE
ResNet-18 / 11	Functional	10.789	$\mathbf{3.48\times 10^{-4}}$	$\mathbf{1.21\times 10^{-3}}$
	Hessian	10.982	4.47×10^{-4}	1.59×10^{-3}
MobileNetV3-S / 8	Functional-add.	7.971	0.01352	0.08946
	Functional-int.	7.991	**0.01264**	**0.07283**
	Uniform INT8	8.000	0.03978	0.22186

At the layer–action level, the JVP approximation is highly faithful to the nonlinear isolated perturbation, with Spearman correlations of $\rho = 0.992$–0.999. The remaining difficulty therefore lies primarily in composing many simultaneously quantized layers rather than in approximating an individual perturbation. The severity-controlled analysis further shows that the observed ranking quality is not explained simply by overall compression magnitude.

The cross-architecture results reinforce this distinction. MobileNetV3-Small benefits from first-order tangent-interference terms, whereas MobileNetV2 retains strong additive prediction without a corresponding interaction-aware gain. Thus, non-additive first-order effects can be important, but their predictive value is architecture- and target-dependent rather than universal. Moreover, stronger global ranking does not necessarily translate into a better constrained allocation at every bit budget, so predictor quality and decision quality should be evaluated separately.

Our study is limited to weight-only FP32/INT8/INT4 post-training quantization on Imagenette and three CNN architectures, with full allocation experiments on two of them. Activations remain in full precision, and we do not evaluate device-level latency or energy. Extending the analysis to larger datasets, additional architectures, activation quantization, hardware-aware objectives, and more efficient JVP implementations remains important future work.

7 Conclusion

We study the sensitivity of concrete quantization perturbations through their Euclidean logit displacement,

$$S_F(\Delta w) = \Delta w^\top \mathbb{E}[J^\top J]\Delta w,$$

viewed as the $M = I$ instance of the broader $J^\top M J$ family of output-space geometries. Across three CNN architectures, this criterion provides a strong predictor of realized mixed-precision damage, remains informative within compression-severity strata, closely tracks nonlinear isolated-output perturbations, and yields useful costs for budget-constrained weight allocation. The experiments also clarify where the approximation succeeds and where additional structure matters. Local JVP responses are highly accurate for isolated compression actions, while MobileNetV3-Small shows that first-order tangent interference can improve configuration ranking when additive composition is insufficient. MobileNetV2, however, shows that such gains are not universal. These findings support compression-direction output geometry as a useful framework for studying mixed-precision sensitivity, while distinguishing the present contribution from prior Jacobian-based output-distortion methods: the emphasis here is on the choice of output geometry, local fidelity, and cross-layer composition rather than on claiming Jacobian-based distortion itself as new.

Reproducibility Statement

The main paper specifies the quantizer, sensitivity definitions, data splits, allocation objective, and statistical protocol; the appendix provides the full ablations, interaction diagnostics, quantization details, and profiling measurements.

AI USE STATEMENT

Generative AI tools were used to assist with literature discovery, manuscript editing, and restructuring.

REFERENCES

Ron Banner, Yury Nahshan, and Daniel Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. Zeroq: A novel zero shot quantization framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13169–13178, 2020.

Zhaowei Cai and Nuno Vasconcelos. Rethinking differentiable search for mixed-precision neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2349–2358, 2020.

Arun Chauhan, Utsav Tiwari, and Vikram N R. Post training mixed precision quantization of neural networks using first-order information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 1343–1352, 2023.

Weihan Chen, Peisong Wang, and Jian Cheng. Towards mixed-precision quantization of neural networks via constrained optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5350–5359, 2021.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. Hawq-v2: Hessian aware trace-weighted quantization of neural networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha. Learned step size quantization. In *International Conference on Learning Representations*, 2020.

Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *International Conference on Learning Representations*, 2016.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. Brecq: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations*, 2021.

Yuexiao Ma, Taisong Jin, Xiawu Zheng, Yan Wang, Huixia Li, Yongjian Wu, Guannan Jiang, Wei Zhang, and Rongrong Ji. Ompq: Orthogonal mixed precision quantization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9029–9037, 2023. doi: 10.1609/aaai.v37i7.26084.

Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Markus Nagel, Rana Ali Amjad, Mart van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 2020.

Md. Mehrab Hossain Opi, Robiul Islam Ryad, and Md. Umar Faruk. Mixfrag: Fragility-guided mixed-precision post-training quantization for vision transformers. *arXiv preprint arXiv:2607.28589*, 2026.

Clemens J. S. Schaefer, Navid Lambert-Shirzad, Xiaofan Zhang, Chia-Wei Chou, Tom Jablin, Jian Li, Elfie Guo, Caitlin Stanton, Siddharth Joshi, and Yu Emma Wang. Augmenting hessians with inter-layer dependencies for mixed-precision post-training quantization. *arXiv preprint arXiv:2306.04879*, 2023.

Stefan Uhlich, Lukas Mauch, Fabien Cardinaux, Kazuki Yoshiyama, Javier Alonso Garcia, Stephen Tiedemann, Thomas Kemp, and Akira Nakamura. Mixed precision dnns: All you need is a good parametrization. In *International Conference on Learning Representations*, 2020.

Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. Haq: Hardware-aware automated quantization with mixed precision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Zhe Wang, Kaixin Xu, Xue Geng, Fen Fang, Mohamed M. Sabry Aly, Xulei Yang, Min Wu, and Weisi Lin. Blob-q: Boosting low bit vit quantization via global optimization on model distortion. In *Computer Vision – ECCV 2026*, pp. 385–404. Springer, 2026.

Ziwei Wang, Han Xiao, Jiwen Lu, and Jie Zhou. Generalizable mixed-precision quantization via attribution rank preservation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5271–5280, 2021.

Xiuying Wei, Ruihao Gong, Yuhang Li, Xianglong Liu, and Fengwei Yu. Qdrop: Randomly dropping quantization for extremely low-bit post-training quantization. In *International Conference on Learning Representations*, 2022.

Bichen Wu, Yanghan Wang, Peizhao Zhang, Yuandong Tian, Peter Vajda, and Kurt Keutzer. Mixed precision quantization of convnets via differentiable neural architecture search. *arXiv preprint arXiv:1812.00090*, 2018.

A Experimental Setup and Dataset Details

Dataset and preprocessing. All experiments use Imagenette, a 10-class subset of ImageNet. The evaluated networks use ImageNet-pretrained weights and the corresponding ImageNet preprocessing; the 10 Imagenette classes are mapped to their ImageNet-1K output indices and the classifier heads are not retrained. Calibration examples are drawn from the training split, whereas predictor evaluation and final allocation evaluation use the validation split, keeping calibration and evaluation data disjoint.

Architectures and evaluation protocol. ResNet-18 and MobileNetV3-Small are the two primary architectures. In our setup, they contain 21 and 54 quantized convolutional/linear weight tensors, respectively. MobileNetV2 is used only as an external architecture check for predictor and tangent-interference generalization and is not included in the allocation study. Table 5 summarizes the calibration and evaluation protocol.

Table 5: Summary of the experimental protocol. "Predictor eval." denotes the held-out subset used to measure realized degradation for the heterogeneous configuration study. "Full allocation eval." denotes re-evaluation of selected mixed-precision configurations on the complete Imagenette validation split.

Architecture	Quantized tensors	Calibration	Predictor eval.	Full allocation eval.
ResNet-18	21	64	512	3,925
MobileNetV3-Small	54	32	512	3,925
MobileNetV2	–	32	512	–

Table 6: Complete full-validation ResNet-18 allocation results on 3,925 Imagenette images. Lower KL and logit MSE are better. Average bits are parameter weighted.

Budget	Method	Avg. bits	KL	Logit MSE
8	Functional	7.969	7.805×10^{-3}	2.639×10^{-2}
	Hessian	7.969	7.805×10^{-3}	2.639×10^{-2}
	Grad.-square	7.969	7.805×10^{-3}	2.639×10^{-2}
	Uniform INT8	8.000	$\mathbf{2.762\times10^{-3}}$	$\mathbf{9.466\times10^{-3}}$
9	**Functional**	8.861	$\mathbf{6.48\times10^{-4}}$	$\mathbf{2.25\times10^{-3}}$
	Hessian	8.861	6.96×10^{-4}	2.40×10^{-3}
	Grad.-square	8.861	6.48×10^{-4}	2.25×10^{-3}
10	Functional	9.913	4.86×10^{-4}	$\mathbf{1.66\times10^{-3}}$
	Hessian	9.972	$\mathbf{4.85\times10^{-4}}$	1.72×10^{-3}
	Grad.-square	9.770	5.66×10^{-4}	1.97×10^{-3}
11	**Functional**	**10.789**	$\mathbf{3.48\times10^{-4}}$	$\mathbf{1.21\times10^{-3}}$
	Hessian	10.982	4.47×10^{-4}	1.59×10^{-3}
	Grad.-square	10.982	4.91×10^{-4}	1.76×10^{-3}

Precision configurations. Candidate weight precisions are $\{\mathrm{FP32}, \mathrm{INT8}, \mathrm{INT4}\}$. For each architecture, the predictor-validation study uses 125 unique heterogeneous layer-wise assignments generated with fixed random seeds; duplicate full-network assignments are rejected. The three uniform-precision configurations are retained separately as anchors and are excluded from the heterogeneous-configuration rank-correlation analysis. The exact sampled configurations and seeds are retained with the experimental outputs.

Quantization scope. The study is weight-only post-training quantization: activations remain in full precision and no fine-tuning or retraining is performed after quantization. INT8 and INT4 weights use deterministic symmetric per-output-channel fake quantization; the exact operator is given in Appendix C. Consequently, reported average bit widths measure parameter-weighted weight precision rather than hardware latency or energy.

Prediction targets and statistical protocol. Realized compression damage is evaluated using teacher–student KL divergence, logit MSE, change in cross-entropy, and change in Top-1 accuracy relative to the FP32 model. Predictor quality is measured primarily with Spearman rank correlation. Pairwise comparisons between predictors use 5,000 paired bootstrap resamples over the sampled heterogeneous configurations. These intervals characterize stability over the evaluated configuration set; they do not represent uncertainty over independently trained models or architecture draws. Final selected allocations for ResNet-18 and MobileNetV3-Small are re-evaluated on all 3,925 Imagenette validation images.

B Full Allocation Results

The complete full-validation allocation sweeps corresponding to the representative checkpoints in the main text are reported below.

Table 7: Complete full-validation MobileNetV3-Small allocation results on 3,925 Imagenette images. Lower KL, logit MSE, ΔCE, and ΔTop-1 are better as degradation measures; negative deltas indicate an improvement relative to FP32 on that discrete/task metric.

Budget	Method	Avg. bits	KL	Logit MSE	ΔCE	ΔTop-1
8	Functional-add.	7.971	0.01352	0.08946	0.01155	0.178
	Functional-int.	7.991	**0.01264**	**0.07283**	0.01926	**0.102**
	Uniform INT8	8.000	0.03978	0.22186	0.03719	0.790
9	**Functional-add.**	8.961	**0.000899**	**0.006227**	-0.00235	-0.255
	Functional-int.	8.983	0.000951	0.006296	-0.00280	-0.178
10	**Functional-add.**	9.972	**0.000652**	**0.004700**	-0.00336	-0.433
	Functional-int.	9.972	0.000695	0.004756	-0.00025	-0.204
11	Functional-add.	10.970	0.000493	0.003340	-0.00276	-0.204
	Functional-int.	10.970	**0.000468**	**0.003283**	-0.00076	-0.051

C Quantization and Calibration Details

For $b \in \{4, 8\}$ and output channel c, let

$$q_{\max} = 2^{b-1} - 1, \qquad s_c = \frac{\max_j |W_{c,j}|}{q_{\max}}. \tag{20}$$

Weights are quantized and dequantized as

$$Q_b(W_{c,j}) = s_c \,\mathrm{clip}\left(\mathrm{round}\left(\frac{W_{c,j}}{s_c}\right), -q_{\max}, q_{\max}\right). \tag{21}$$

FP32 is the identity action. All experiments are weight-only PTQ with full-precision activations and no fine-tuning.

C.1 Calibration-size ablation

Table 8 reports the complete calibration-size ablation for additive functional sensitivity.

Table 8: Calibration-size ablation. Values are mean Spearman correlations. ResNet-18 uses 100 subsamples below $N = 64$; MobileNetV3-Small uses 20 subsets below $N = 32$.

Architecture	N	KL	Logit MSE	ΔCE	ΔTop-1
ResNet-18	4	0.978	0.977	0.922	0.853
	8	0.984	0.983	0.931	0.864
	16	0.987	0.985	0.933	0.866
	32	0.990	0.987	0.937	0.871
	64	0.991	0.988	0.939	0.873
MobileNetV3-S	4	0.699	0.756	0.676	0.661
	8	0.697	0.754	0.674	0.658
	16	0.706	0.761	0.683	0.666
	32	0.705	0.762	0.680	0.662

D Severity-Controlled Ranking

A high global correlation could arise trivially if both a predictor and the observed damage only track overall compression magnitude. We therefore partition the 125 MobileNetV3-Small heterogeneous configurations into low/medium/high tertiles using configuration-level summed relative weight SSE, a parameter-space severity variable that does not use the functional scores. The bin sizes are 42/41/42. Table 9 reports within-bin Spearman correlations for the first-order tangent-interference score for the primary and alternate configuration samples.

For the primary sample, the low-minus-high 95% paired-bootstrap intervals are $[-0.415, 0.065]$ for KL and $[-0.288, 0.152]$ for logit MSE; for the alternate sample they are $[-0.318, 0.107]$ and

Table 9: Severity-controlled MobileNetV3-Small correlations. Each entry is Spearman ρ computed within a relative-weight-SSE severity tertile. The alternate sample uses a different configuration draw.

Sample	Predictor	KL-L	KL-M	KL-H	Logit-L	Logit-M	Logit-H
Primary	Func.-int.	0.663	0.760	0.828	0.757	0.778	0.825
Alternate	Func.-int.	0.724	0.810	0.812	0.775	0.817	0.819

$[-0.242, 0.142]$. The corresponding intervals for ΔCE and ΔTop-1 also cross zero in both samples. Thus the functional ranking remains substantial within severity strata and does not rely only on ordering mild versus severe configurations.

E Additional Output-Geometry Results

Table 10 gives the complete additive and tangent-interference comparison between Euclidean output displacement ($M = I$) and cross-entropy generalized Gauss–Newton geometry ($M = H_f$). On ResNet-18, additive functional sensitivity exceeds additive CE-GGN for teacher–student KL by 0.0067 with paired 95% bootstrap CI $[0.0006, 0.0177]$ and for logit MSE by 0.0148 with CI $[0.0060, 0.0310]$. The additive cross-entropy difference is not significant, while CE-GGN is significantly stronger for Top-1 degradation: functional-minus-GGN is -0.0146 with CI $[-0.0315, -0.0019]$. In the tangent-interference comparison, functional-minus-GGN is significant for logit MSE ($+0.0147$, CI $[0.0058, 0.0316]$) and favors GGN for Top-1 (-0.0245, CI $[-0.0435, -0.0072]$); the KL and cross-entropy intervals cross zero. On MobileNetV3-Small, all functional-versus-GGN paired intervals cross zero for both additive and tangent-interference scores.

Table 10: Complete output-geometry ablation. Values are Spearman correlations over the 125 heterogeneous configurations.

Architecture	Predictor	KL	Logit MSE	ΔCE	ΔTop-1
ResNet-18	Func.-ind.	0.9909	0.9879	0.9387	0.8732
	GGN-ind.	0.9842	0.9731	0.9431	0.8878
	Func.-int.	0.9819	0.9907	0.9173	0.8476
	GGN-int.	0.9796	0.9760	0.9290	0.8721
MobileNetV3-S	Func.-ind.	0.7050	0.7618	0.6797	0.6622
	GGN-ind.	0.7224	0.7656	0.6975	0.6809
	Func.-int.	0.7758	0.8023	0.7522	0.7263
	GGN-int.	0.7754	0.7946	0.7400	0.6982

F Direct Isolated-Output Fragility

The direct baseline explicitly constructs a model with only one non-FP32 layer–action perturbation and measures its nonlinear output damage on the calibration set. Table 11 shows that the JVP layer–action score closely tracks this direct damage on both architectures. Table 12 then compares additive configuration rankings formed from the local scores.

Table 11: Layer–action agreement between functional sensitivity and directly measured isolated-output damage. Spearman correlations are computed over all non-FP32 layer–action pairs.

Architecture	Pairs	Direct KL	Direct logit MSE
ResNet-18	42	0.9924	0.9985
MobileNetV3-Small	108	0.9927	0.9992

For ResNet-18, functional-minus-direct differences for KL are statistically indistinguishable for both direct-KL and direct-logit sums. For logit MSE, functional sensitivity is higher than the direct-KL

Table 12: Configuration-level Spearman correlations for additive functional sensitivity and summed direct isolated-output costs.

Architecture	Predictor	KL	Logit MSE	ΔCE	ΔTop-1
ResNet-18	Func.-ind.	0.9909	0.9879	0.9387	0.8732
	Direct KL	0.9917	0.9811	0.9554	0.8967
	Direct logit	0.9913	0.9889	0.9427	0.8748
MobileNetV3-S	Func.-ind.	0.7050	0.7618	0.6797	0.6622
	Direct KL	0.8050	0.7886	0.8063	0.8032
	Direct logit	0.7963	0.7991	0.7838	0.7734

sum by 0.0069 with CI $[0.0008, 0.0159]$, while it is statistically indistinguishable from the direct-logit sum. Direct-KL is significantly stronger for ΔCE and ΔTop-1. On MobileNetV3-Small, direct-KL and direct-logit sums are significantly stronger than additive functional sensitivity for KL; direct-KL is also significantly stronger for ΔCE and ΔTop-1. The logit-MSE differences have bootstrap intervals that cross zero.

Table 13 reports the direct-baseline allocation decisions on the original 512-image E1/E2 evaluation subset. On ResNet-18, direct isolated-logit selects the same effective configurations as additive functional sensitivity at budgets 9–11. On MobileNetV3-Small, direct-logit coincides with the additive functional solution at budgets 9 and 10, while the 8-bit direct objectives select different trade-offs.

Table 13: Observed 512-image allocation results for direct isolated-output costs. Lower KL and logit MSE are better.

Architecture	Method	Budget	Avg. bits	KL	Logit MSE
ResNet-18	Direct KL	9	8.861	0.000624	0.002262
	Direct KL	10	9.972	0.000479	0.001717
	Direct KL	11	10.822	0.000367	0.001206
	Direct logit	9	8.861	0.000624	0.002262
	Direct logit	10	9.913	0.000481	0.001684
	Direct logit	11	10.789	0.000351	0.001221
MobileNetV3-S	Direct KL	8	7.969	0.009033	0.078210
	Direct KL	9	8.961	0.000907	0.006060
	Direct KL	10	9.972	0.000739	0.004540
	Direct logit	8	7.971	0.013899	0.074490
	Direct logit	9	8.961	0.000907	0.006060
	Direct logit	10	9.972	0.000760	0.004706

G MobileNetV2 External-Architecture Check

MobileNetV2 is evaluated as an external predictor-generalization check using the same 125 heterogeneous FP32/INT8/INT4 configurations, 32 calibration images, and 512 evaluation images as the MobileNetV3-Small predictor study. Additive functional sensitivity remains strongly rank-correlated with all four observed degradation targets. Tangent-interference scoring is slightly lower for every target, but none of the paired tangent-interference-minus-additive differences is statistically distinguishable from zero under 5,000 paired bootstrap resamples.

Table 14: MobileNetV2 external-architecture results. $\Delta\rho = \rho_{\mathrm{int}} - \rho_{\mathrm{ind}}$; confidence intervals are paired 95% bootstrap intervals over the 125 heterogeneous configurations.

Target	ρ_{ind}	ρ_{int}	$\Delta\rho$	95% CI
Teacher–student KL	0.9353	0.9283	-0.0071	$[-0.0196, 0.0043]$
Logit MSE	0.9043	0.8976	-0.0067	$[-0.0217, 0.0069]$
ΔCE	0.9196	0.9099	-0.0097	$[-0.0244, 0.0031]$
ΔTop-1	0.9192	0.9138	-0.0054	$[-0.0201, 0.0080]$

The median configuration-level relative departure from additivity is $R_{\text{abs}} = 0.096$, intermediate between ResNet-18 (0.069) and MobileNetV3-Small (0.140). This result separates two claims that would otherwise be conflated: the additive functional metric generalizes well to MobileNetV2, but a larger non-additive term does not by itself imply that tangent-interference ranking will be better.

H TANGENT-INTERFERENCE DIAGNOSTICS

We summarize the configuration-level departure from additivity with

$$R_{\text{abs}} = \frac{|S_{\text{int}} - S_{\text{ind}}|}{S_{\text{ind}} + \epsilon}, \tag{22}$$

and, for fixed INT4 actions, normalized pairwise tangent interference with

$$I_{lk} = \frac{\mathbb{E}[g_l^\top g_k]}{\sqrt{\mathbb{E}\|g_l\|_2^2 \, \mathbb{E}\|g_k\|_2^2}}. \tag{23}$$

Table 15 reports the corresponding diagnostics. MobileNetV3-Small exhibits the largest median configuration-level departure from additivity and broader pairwise extremes than ResNet-18. MobileNetV2 is intermediate in R_{abs}, reinforcing that a larger non-additive term does not by itself guarantee a ranking gain from tangent-interference scoring.

Table 15: Tangent-interference diagnostics over the 125 heterogeneous configurations. Pairwise statistics use fixed INT4 actions; they were not computed for the external MobileNetV2 check.

Architecture	Median R_{abs} (IQR)	Median $\lvert I_{lk}\rvert$	Pairwise range
ResNet-18	0.069 $[0.031, 0.133]$	0.0368	$[-0.302, 0.218]$
MobileNetV3-Small	0.140 $[0.064, 0.247]$	0.0434	$[-0.458, 0.544]$
MobileNetV2	0.096 $[0.049, 0.196]$	–	–

I PROFILING-TIME MEASUREMENTS

All timings below use the same CPU implementation and machine used for the corresponding experiments. The Hessian timing is listed for completeness but is not directly comparable because its protocol uses one 8-image calibration batch rather than the full calibration sets used by functional and direct-isolated profiling.

Table 16: Profiling and cached decision costs. Functional/direct profiling uses 64 calibration images for ResNet-18 and 32 for MobileNetV3-Small.

Quantity	ResNet-18	MobileNetV3-Small
Functional JVP profile (s)	194.07	93.81
Direct isolated profile (s)	56.89	16.35
Directional Hessian profile (s)	25.46†	–
Cached scoring, 128 configs (ms)	2.21	3.47
Additive MCKP solve (ms)	7.33	14.46

†One 8-image calibration batch; not a like-for-like full-calibration timing.