WHO NEEDS MORE BITS? TASK-AWARE QUANTIZATION FOR MULTI-SOURCE COLLABORATIVE INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Collaborative AI models increasingly integrate multiple views, modalities, or agents to solve a shared task, yet different information sources can contribute unequally to the final prediction. This raises a fundamental question for model quantization: how much precision should each source receive, and what determines its precision requirement? We study this problem from an information-theoretic perspective. We first establish a local-to-global perturbation characterization that connects source-wise parameter quantization errors to collaborative output distortion. Building on this result, we derive lower and upper rate-distortion bounds for task-aware multi-source model quantization by jointly accounting for source-specific parameter statistics and task contributions. We further characterize heterogeneous source contributions from a partial information decomposition perspective. Based on these results, we derive an explicit precision-allocation law and develop a lightweight precision-allocation rule that maps the theoretical solution to hardware-supported bit-widths. Experiments across multiple collaborative models, quantization methods, and practical resource constraints show that the derived bounds capture practical performance trends and effectively guide source-wise precision allocation.

1 INTRODUCTION

Modern AI systems increasingly combine multiple information sources rather than relying on a single input stream. Examples span multimodal learning and multi-agent collaborative perception, where representations from different views, modalities, sensors, or agents are integrated to make a shared prediction (Baltrušaitis et al., 2019; Xu et al., 2022; Hu et al., 2022). A key characteristic of such systems is *source heterogeneity*: different sources often contribute unequally to a given downstream task. One source may provide unique information that resolves an ambiguous prediction, while another may be largely redundant with the remaining sources. Consequently, preserving every source with the same precision can be inefficient when their contributions to the collaborative task are substantially different.

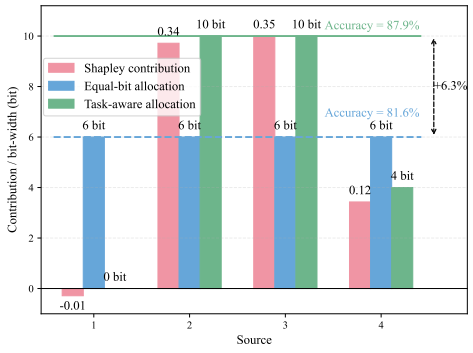


Figure 1: Motivating example of source-wise precision allocation. Under the same total source-wise precision budget of 24 bits (an average of 6 bits per source), equal-bit allocation uses $(6, 6, 6, 6)$ bits and achieves 81.6% accuracy, whereas task-aware allocation uses $(0, 10, 10, 4)$ bits and achieves 87.9%, yielding a 6.3 percentage-point improvement.

Meanwhile, the increasing size of modern AI models creates substantial memory and computation costs for resource-constrained devices and agents. Quantization alleviates these costs by representing model parameters with low-precision values, and recent post-training quantization (PTQ) methods have demonstrated strong performance without full retraining (Frantar et al., 2023; Xiao et al., 2023; Lin et al., 2024; Shao et al., 2024). Mixed-precision quantization further recognizes that different model components can have different sensitivities to

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.

quantization and therefore need not use the same bit-width (Dong et al., 2020; Chowdhury et al., 2026). However, these methods primarily determine precision according to the sensitivity of layers, tensors, or experts within an individual model. Collaborative inference (co-inference), introduces an additional dimension of heterogeneity: a source-specific branch is characterized not only by how its parameters respond to quantization, but also by how much that source contributes to the shared task. This raises a different question: *when multiple source-specific model branches jointly contribute to the same task, should they all use the same precision?*

Figure 1 illustrates this distinction. Different sources contribute differently to the final task, as quantified by their Shapley contributions. Under the same precision budget, equal-bit allocation assigns equal bits to every source branch. In contrast, task-aware allocation assigns more precision to sources with larger task contributions and less precision to those with smaller contributions, improving the performance without increasing the total bit budget. Thus, the key question is not simply whether mixed precision is beneficial, but *how the precision of each source should be determined*. This motivates the central question of this work: *given a limited precision budget, how many bits should be assigned to each source-specific model branch in co-inference?*

Answering this question requires resolving **three theoretical challenges**. First, quantization acts locally on model parameters, whereas its consequence is observed at the final collaborative output. For highly nonlinear models, relating source-wise parameter perturbations to global output distortion is nontrivial. Second, a principled allocation requires characterizing how the quantization rate of each source branch affects its collaborative task performance. Third, equal parameter distortions may have unequal task impacts because different sources provide heterogeneous task-relevant information. Existing quantization methods provide effective empirical solutions, but do not establish a unified relationship among source contribution, quantization rate, parameter distortion, and collaborative task degradation.

To address these challenges, we develop a task-aware rate-distortion characterization for multi-source model quantization. We first establish a local-to-global perturbation relation between source-wise parameter distortion and collaborative output distortion. We then derive rate-distortion bounds for source-specific model quantization. Finally, we use partial information decomposition (PID) to interpret source interactions and employ Shapley attribution to quantify source contributions in the multi-source setting. Together, these components form a unified framework for task-aware source-wise precision allocation.

Our analysis leads to **three main insights**. First, quantization-induced collaborative output distortion admits a tractable source-wise additive characterization, making source-wise parameter distortion a useful surrogate for analyzing global degradation. Second, the precision requirement of each source is jointly determined by its task contribution and its source-specific model parameter statistics. Third, uniform precision is optimal only under restrictive symmetry conditions; in general, sources with larger task contributions or less compressible parameter distributions should receive more bits, leading to an explicit source-wise precision-allocation law under a limited bit budget.

Our main contributions are summarized as follows:

- **A source-wise local-to-global distortion characterization.** We establish a perturbation relation that connects local parameter quantization errors to the global collaborative output distortion. The resulting source-wise additive structure provides a tractable basis for analyzing how local quantization perturbations propagate through co-inference.
- **A task-aware rate-distortion characterization and precision-allocation principle.** We derive lower and upper quantization rate-distortion bounds for co-inference and characterize source contribution through a PID-inspired Shapley attribution. The resulting characterization reveals how source contribution and parameter statistics jointly determine precision requirements and yields an explicit continuous precision-allocation law, which is further translated into a lightweight algorithm for hardware-supported bit-widths.
- **Validation from theoretical characterization to practical deployment.** We systematically validate the proposed framework from source-wise distortion propagation and rate-distortion characterization to precision allocation and edge deployment. Experiments show that the theoretical characterization captures practical quantization trends, while the resulting allocation consistently outperforms the baselines under both bit-budget and latency/energy constraints, and remains close to the performance upper bound from exhaustive search at substantially lower computational cost.

2 Related Work

Post-training and mixed-precision quantization. PTQ reduces model memory and inference cost without requiring full retraining. Representative methods include GPTQ, which uses approximate second-order information for weight quantization (Frantar et al., 2023); SmoothQuant, which transfers activation quantization difficulty to weights through mathematically equivalent scaling (Xiao et al., 2023); AWQ, which protects salient weights through activation-aware scaling (Lin et al., 2024); and OmniQuant, which optimizes quantization parameters through lightweight calibration (Shao et al., 2024). Mixed-precision methods further recognize that different model components may require different bit-widths. For example, HAWQ use Hessian-based sensitivity to allocate precision across layers (Dong et al., 2020), while recent work extends heterogeneous precision allocation to individual experts in mixture-of-experts models (Chowdhury et al., 2026). These methods primarily study mixed-precision quantization within a single model, determining component-wise precision according to quantization sensitivity. They do not explicitly consider multi-source co-inference. In contrast, we study this setting, where the precision requirement of each branch depends not only on its quantization behavior but also on its task contribution.

Theory-guided model compression. A growing body of work uses analytical principles to characterize compression behavior and guide practical model compression. Rate-distortion and information-theoretic formulations have been used to characterize fundamental compression-distortion tradeoffs and derive principled objectives for model compression and pruning (e.g., (Gao et al., 2019; Isik et al., 2022; Young, 2025)). Sensitivity-based approaches instead exploit parameter/gradient information to guide mixed-precision quantization (Dong et al., 2020). More recent studies further provide theoretical explanations for the interaction between sparsity and quantization and for the geometric structure underlying modern PTQ (Harma et al., 2025; Chen et al., 2026). Our work follows this theory-to-design philosophy, but focuses on a different setting: we characterize how source-wise parameter quantization propagates through multi-source co-inference and combine this distortion characterization with heterogeneous task contributions to guide source-wise precision allocation.

Multi-source collaboration. Multi-source AI broadly spans multi-source learning and co-inference (Baltrušaitis et al., 2019). A common challenge is to effectively fuse or aggregate information from heterogeneous sources (e.g., via tensor fusion or cross-modal attention (Zadeh et al., 2017; Tsai et al., 2019)). For multi-source learning, Deep Sets and Set Transformer develop permutation-invariant architectures for aggregating and modeling interactions among set-structured inputs (Zaheer et al., 2017; Lee et al., 2019). Multi-source co-inference further enables distributed agents to exchange complementary information during collaborative prediction. Representative approaches optimize communication graphs and timing, selectively transmit task-relevant spatial features, or characterize the communication rate-distortion tradeoff (Liu et al., 2020; Hu et al., 2022; Liu et al., 2026). These inference methods primarily focus on the transmission of intermediate information. In contrast, our work studies joint model compression across source-specific branches: we investigate how heterogeneous source contributions should guide source-wise model compression.

3 Problem Setup and Local-to-Global Distortion

We consider K information sources indexed by $\mathcal{K} = \{1, \dots, K\}$. Source k observes $\boldsymbol{x}_k$ and is processed by a source-specific branch $f_k(\cdot)$ with parameters $\boldsymbol{W}_k \in \mathbb{R}^{q_k}$, yielding $\boldsymbol{o}_k = f_k(\boldsymbol{x}_k; \boldsymbol{W}_k)$. The source outputs are fused as $\boldsymbol{u} = g(\boldsymbol{o}_1, \dots, \boldsymbol{o}_K)$, where $g(\cdot)$ denotes the fusion module. Finally, the task-specific head produces $\boldsymbol{z} = h(\boldsymbol{u}; \boldsymbol{V})$, where $h(\cdot; \boldsymbol{V})$ is the shared task executor with parameters $\boldsymbol{V}$. We quantize only the source-specific parameters as $\hat{\boldsymbol{W}}_k = Q_k(\boldsymbol{W}_k; R_k)$, where R_k denotes the bit-width assigned to each parameter of source k. Accordingly, the quantized source output, fusion result, and final output are $\hat{\boldsymbol{o}}_k = f_k(\boldsymbol{x}_k; \hat{\boldsymbol{W}}_k)$, $\hat{\boldsymbol{u}} = g(\hat{\boldsymbol{o}}_1, \dots, \hat{\boldsymbol{o}}_K)$, and $\hat{\boldsymbol{z}} = h(\hat{\boldsymbol{u}}; \boldsymbol{V})$.

We next characterize how source-wise parameter distortion propagates to the collaborative output. We first introduce the following assumptions.

Assumption 1 (Normalized input)**.** For every source k, $\|\boldsymbol{x}_k\|_1 \leq 1$.

Assumption 2 (Activation smoothness)**.** The activation function $\sigma(\cdot)$ is 1-Lipschitz, i.e., $\|\sigma(\boldsymbol{x}) - \sigma(\boldsymbol{y})\|_1 \leq \|\boldsymbol{x} - \boldsymbol{y}\|_1$, and satisfies $\sigma(\boldsymbol{0}) = \boldsymbol{0}$. This condition holds for commonly used activation functions such as ReLU, LeakyReLU, and tanh.

Assumption 3 (Bounded layer-wise quantization error)**.** For the l-th layer of source branch k, $\|\boldsymbol{W}_k^{(l)} - \hat{\boldsymbol{W}}_k^{(l)}\|_1 \le \tau_k^{(l)}$, where $\tau_k^{(l)} \ge 0$ depends on the quantization resolution.

Assumption 4 (Collaborative-model smoothness)**.** The fusion function is block-Lipschitz such that

$$\|g(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_K) - g(\hat{\boldsymbol{o}}_1, \ldots, \hat{\boldsymbol{o}}_K)\|_1 \le \sum\nolimits_{k=1}^{K} L_k \|\boldsymbol{o}_k - \hat{\boldsymbol{o}}_k\|_1. \tag{1}$$

For example, concatenation satisfies $L_k = 1$, whereas averaging satisfies $L_k = 1/K$. The shared task executor is also C_h-Lipschitz continuous, i.e., $\|h(\boldsymbol{u}; \boldsymbol{V}) - h(\hat{\boldsymbol{u}}; \boldsymbol{V})\|_1 \le C_h \|\boldsymbol{u} - \hat{\boldsymbol{u}}\|_1$.

Assumption 5 (Layer-wise gradient bound)**.** For the l-th parameter block of source branch k, the corresponding Jacobian has bounded element-wise ℓ_1 norm, i.e., $\|\nabla_{\boldsymbol{W}_k^{(l)}} f_k(\boldsymbol{x}_k; \boldsymbol{W}_k)\|_1 \le A_k^{(l)}$.

Proposition 1 (Local-to-global distortion)**.** Under the first-order parameter perturbation approximation and Assumptions 4 and 5, the collaborative output distortion is upper bounded by

$$\|\boldsymbol{z} - \hat{\boldsymbol{z}}\|_1 \le C_h \sum\nolimits_{k=1}^{K} L_k \sum\nolimits_{l=1}^{L} A_k^{(l)} \|\boldsymbol{W}_k^{(l)} - \hat{\boldsymbol{W}}_k^{(l)}\|_1. \tag{2}$$

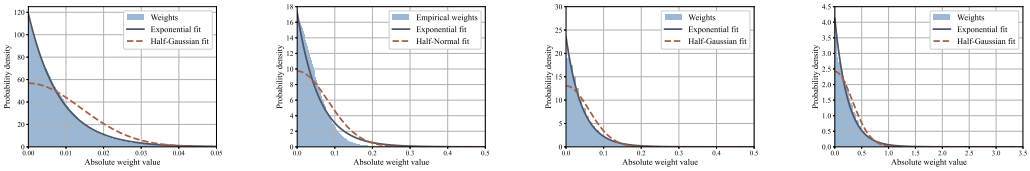
Proof. Please see Appendix A. □

Corollary 1. For an L-layer fully connected (FC) source branch and an N-layer FC shared task executor, under Assumptions 1–3, $A_k^{(l)}$ and C_h admit the forms

$$A_k^{(l)} = \prod\nolimits_{j=1}^{l-1} \|\boldsymbol{W}_k^{(j)}\|_1 \prod\nolimits_{r=l+1}^{L} \left(\|\boldsymbol{W}_k^{(r)}\|_1 + \tau_k^{(r)}\right), C_h = \prod\nolimits_{n=1}^{N} \|\boldsymbol{V}^{(n)}\|_1. \tag{3}$$

Proof. The detailed derivation is provided in Appendix A. □

Insight 1. Proposition 1 shows that local quantization distortions from different sources enter the upper bound on the collaborative output distortion additively. Hence, source-wise parameter distortion provides a tractable surrogate for analyzing quantization-induced task degradation.



(a) ResNet152 He et al. (2016) (b) ViT Dosovitskiy et al. (2021) (c) BERT Devlin et al. (2019) (d) GPT-3 EleutherAI (2021)

Figure 2: Empirical distributions of parameter magnitudes and their exponential and Gaussian-type fits. Compared with the commonly adopted Gaussian-type assumption, the exponential model more accurately captures the empirical magnitude distributions.

4 Task-Aware Rate–Distortion Characterization

1) Source-wise Rate-Distortion Bounds. We next characterize the inference distortion as a function of quantization bit-width. Directly analyzing downstream task performance is generally intractable due to the high nonlinearity of modern AI models. Motivated by Proposition 1, we instead use source-wise parameter distortion as a tractable surrogate.

For source k, let Θ_k denote the magnitude of a model parameter. We model Θ_k by an exponential distribution with parameter λ_k, i.e., $p_k(\theta) = \lambda_k e^{-\lambda_k \theta}, \theta \ge 0$. As shown in Fig. 2, the exponential model provides a tractable approximation to the empirical parameter-magnitude distributions of representative pretrained models. Motivated by Proposition 1, we adopt the absolute-error distortion $d(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$.

Proposition 2 (Multi-source quantization rate-distortion region)**.** Let $\mathcal{D}_{\boldsymbol{\alpha}}(\boldsymbol{R})$ denote the task-aware multi-source quantization distortion under source-wise quantization rates $\boldsymbol{R} = (R_1, \ldots, R_K)$ and task contributions $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$. For the considered exponential parameter model,

$$\sum_{k=1}^{K} \frac{\alpha_k q_k}{\lambda_k 2^{R_k+1}} \le \mathcal{D}_{\boldsymbol{\alpha}}(\boldsymbol{R}) \le \sum_{k=1}^{K} \frac{\alpha_k q_k}{2\lambda_k} \left(\sqrt{1 + \frac{4}{2^{R_k} - 1}} - 1\right). \tag{4}$$

Proof. Please see Appendix B. □

Insight 2. Proposition 2 characterizes how quantization precision, model statistics, and source contribution jointly determine the task-aware distortion. A larger R_k reduces the distortion, while the marginal distortion reduction gradually decreases as precision increases. Moreover, λ_k reflects the parameter-distribution characteristics of source branch k: a smaller λ_k corresponds to a higher-entropy and heavier-tailed distribution, which is more difficult to compress and thus leads to larger distortion under the same bit-width. Finally, a larger α_k places greater weight on the distortion of source k, indicating that task-important sources should generally be quantized more conservatively.

2) Task Contribution α_k via PID and Shapley Attribution. Different sources may provide redundant, unique, or synergistic information for the same task, which can be interpreted through PID Williams & Beer (2010). However, exact PID becomes combinatorially expensive as the number of sources increases. We therefore use the Shapley value as a tractable multi-source attribution mechanism, which measures the average marginal contribution of each source over all possible source coalitions Shapley (1953). Specifically, for a subset $S \subseteq \mathcal{K}$, let $v(S)$ denote its task utility gain relative to the empty-source baseline. The Shapley contribution of source k is

$$\phi_k = \sum\nolimits_{S \subseteq \mathcal{K} \setminus \{k\}} \frac{|S|!(K - |S| - 1)!}{K!} \big[v(S \cup \{k\}) - v(S)\big]. \tag{5}$$

We obtain the task contribution coefficient by nonnegative normalization, i.e., $\alpha_k = [\phi_k]^+ / \sum_j [\phi_j]^+$. For large K, (5) can be efficiently approximated through Monte Carlo sampling over random source permutations, as detailed in Appendix C.

5 Theory-Guided Precision Allocation

The rate-distortion characterization in Section 4 provides both lower and upper distortion bounds. We first use the lower bound to derive a source-wise precision allocation principle[1] Specifically, we have the following proposition.

Proposition 3 (Task-aware precision allocation)**.** Under a total parameter-bit budget B, the source-wise precision allocation that minimizes the aggregate lower-envelope distortion is

$$R_k^\star = \left[\log_2\left(\frac{\alpha_k}{\nu \lambda_k}\right)\right]^+, \tag{6}$$

where $\nu > 0$ is selected such that $\sum_{k=1}^K q_k R_k^\star = B$.

Proof. Please see Appendix D. □

Insight 3 (Who needs more bits?) Proposition 3 provides an explicit ordering of source-wise precision requirements. For any two active sources i and j, $R_i^\star > R_j^\star \leftrightarrow \frac{\alpha_i}{\lambda_i} > \frac{\alpha_j}{\lambda_j}$. Hence, a source should receive more bits when it contributes more to the collaborative task (higher α_k), or when its parameters incur larger distortion at a given rate. In particular, for the same task contribution, a smaller λ_k corresponds to a more dispersed magnitude distribution and thus a larger marginal benefit from additional precision. Uniform per-parameter precision is optimal only when α_k/λ_k is identical across all active sources. Thus, the appropriate precision is jointly determined by task contribution and source-specific parameter statistics.

Practical discrete precision allocation. In practice, each source must select its precision from a finite set of hardware-supported bit-widths $\mathcal{B} = \{\beta_1, \ldots, \beta_M\}$, resulting in a discrete allocation. We therefore use the continuous solution in Proposition 3 to guide the practical bit allocation.

Specifically, as shown in Algorithm 1, after obtaining $R_k^\star$ for each source, we construct a small candidate set $\mathcal{C}_k$ by retaining the J nearest supported bit-widths below and above $R_k^\star$. We then jointly search over $\{\mathcal{C}_k\}_{k=1}^K$ using dynamic programming (DP) to find the source-wise bit allocation that minimizes the task-aware theoretical distortion while satisfying the total parameter-bit budget. Algorithm 1 has a complexity of $\mathcal{O}(KJ\bar{B})$ with $\bar{B}$ budget states, which is much lower than exhaustive search with $\mathcal{O}(M^K)$ complexity.

[1]The exponential form of the lower bound leads to an explicit allocation rule, whereas the upper bound does not yield a comparably simple analytical solution.

Algorithm 1 Theory-Guided Discrete Precision Allocation

Require: $\{\alpha_k\}$, $\{\lambda_k\}$, $\{q_k\}$, supported bit-widths $\mathcal{B}$, total budget B, neighborhood size J

1: Compute continuous allocations $\{R_k^\star\}$ from Proposition 3.
2: **for** $k = 1, \ldots, K$ **do**
3: Construct $\mathcal{C}_k$ using the J nearest supported bit-widths below and above $R_k^\star$.
4: **end for**
5: Use DP over $\{\mathcal{C}_k\}_{k=1}^K$ to minimize the task-aware distortion under the total bit budget B.
6: **return** the discrete allocation $\{b_k\}_{k=1}^K$.

6 EXPERIMENTS

We evaluate the proposed characterization from four aspects: **Q1:** whether source-wise parameter distortion can characterize collaborative output distortion; **Q2:** whether the derived rate-distortion bounds capture empirical quantization distortion under practical quantizers; **Q3:** whether the theoretical characterization can guide source-wise precision allocation; and **Q4:** whether the resulting allocation principle remains effective under practical latency and energy constraints.

Experimental setup. We evaluate a four-source collaborative model using ViT-Tiny Steiner et al. (2022) as the source backbone and Qwen3-0.6B-Base Yang et al. (2025) as the task head. The model is fine-tuned and tested on Mini-ImageNet Vinyals et al. (2016). The four sources observe horizontal image regions $[0, 0.1]$, $[0.1, 0.45]$, $[0.4, 0.8]$, and $[0.8, 1]$, respectively. After fine-tuning, we apply two PTQ schemes without any further retraining: uniform weight quantization and non-uniform K-means-based weight quantization.

For further evaluation, we additionally consider a co-inference model based on ViT-Small Steiner et al. (2022). The model is fine-tuned and tested on Caltech-256 Griffin et al. (2007). The early ViT layers serve as source-specific feature extractors, after which the source embeddings are fused and processed by the remaining ViT layers as the downstream task head. The corresponding results are reported in Appendix E.

6.1 Q1: SOURCE CONTRIBUTION AND LOCAL-TO-GLOBAL DISTORTION

Fig. 3 shows the Shapley-based source contribution across the four source devices. Sources 2 and 3 provide the largest contributions, 0.34 and 0.45, respectively, whereas source 1 contributes little on average and has a slightly negative estimated contribution of -0.01. This is consistent with our view-splitting setup, where sources 2 and 3 observe relatively larger central image regions. This clear source heterogeneity motivates source-dependent precision allocation: treating all source branches equally can spend precision on sources that provide limited marginal utility while under-protecting more task-relevant sources.

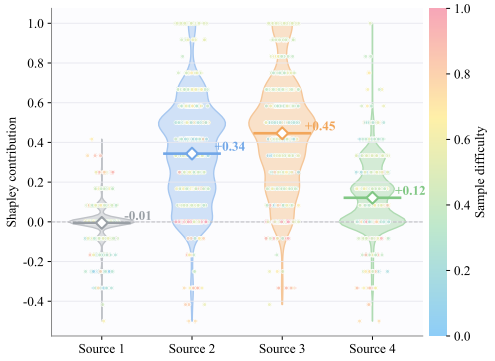


Figure 3: Shapley-based source contributions across the four source devices. The distributions illustrate sample-wise contribution variations, while the annotated values denote the estimated source-level contributions.

Figure 4 further compares source-wise parameter distortion with the resulting collaborative output distortion when varying the source bit-widths. For both quantizers, the output distortion closely follows the variation of parameter distortion, supporting the local-to-global structure in Proposition 1. Moreover, the dashed trajectories reveal different sensitivities across sources. When the bit-widths of the other sources are fixed, varying the bit-width of source 1 causes only a relatively small change in output distortion, whereas changing the bit-width of source 2 leads to a much more pronounced variation. This indicates that the collaborative output exhibits different sensitivities to different sources, which is consistent with their unequal task contributions observed in Fig. 3.

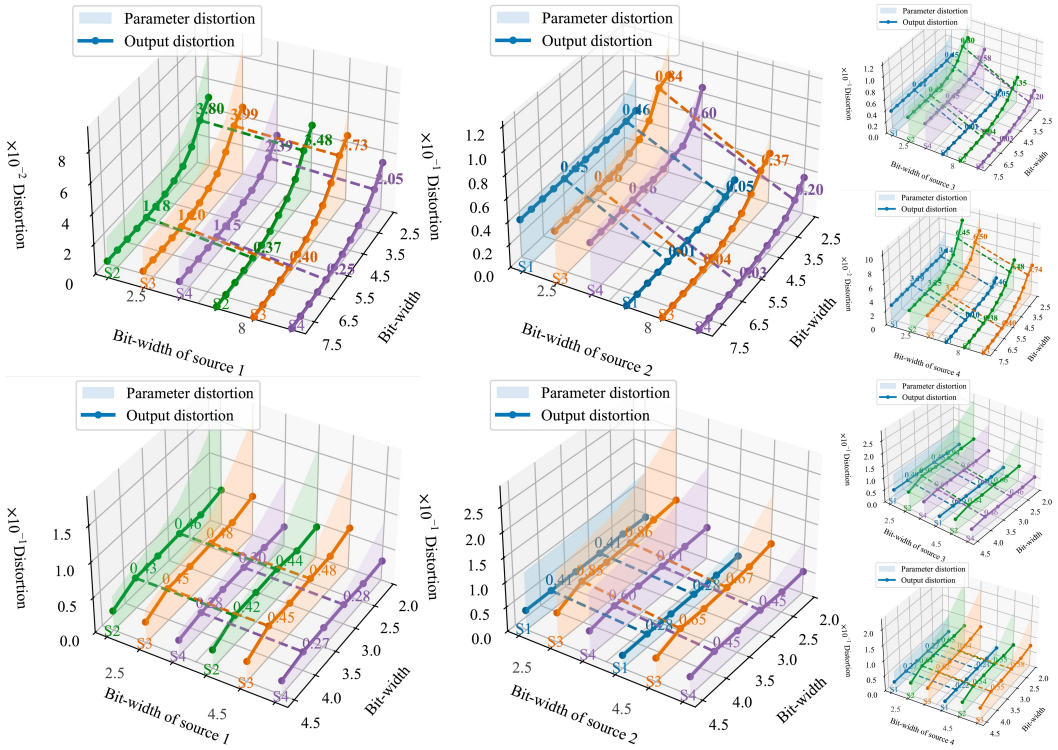
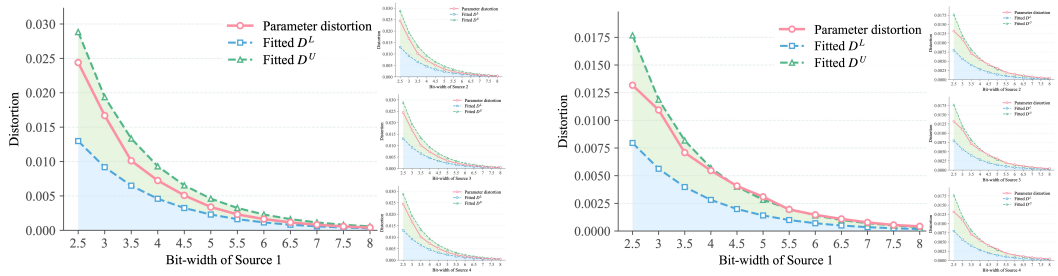


Figure 4: Source-wise parameter distortion and the resulting collaborative output distortion under varying source bit-widths. Top: uniform quantization. Bottom: non-uniform K-means-based quantization.



(a) Uniform quantization

(b) Non-uniform K-means quantization

Figure 5: Empirical source-wise parameter distortion and the analytical lower and achievable upper rate–distortion bounds under uniform and non-uniform K-means quantization.

6.2 Q2: Rate-Distortion Characterization

Figure 5 compares the empirical parameter distortion with the analytical lower and upper rate-distortion bounds in Proposition 2. For both uniform and non-uniform quantization, the empirical distortion decreases consistently with the bit-width, and the derived bounds capture the corresponding trends. Also, the gaps between the two analytical bounds decreases as the precision increases, indicating a tighter characterization in the higher-precision regime. The four sources exhibit similar rate-distortion trends because they share the same ViT backbone and have similar parameter statistics despite observing different views.

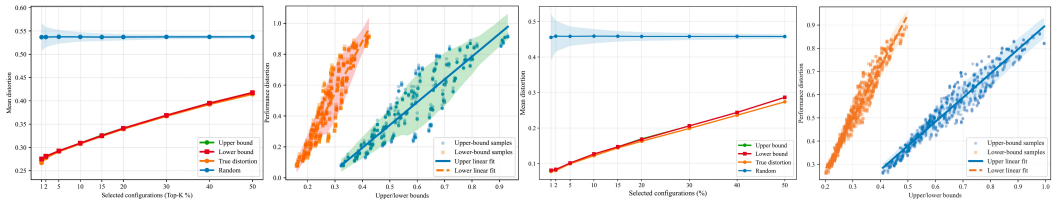


Figure 6: Theory-guided configuration selection and correspondence between theoretical bounds and measured task-performance degradation. From left to right: configuration selection and bound-performance correspondence under uniform and non-uniform quantization, respectively.

Table 1: Task performance (%) and allocation time under different average bit budgets. The full-precision performance is 92.8%.

Method	4 bit		6 bit		8 bit		Allocation time
	Uniform	Non-uniform	Uniform	Non-uniform	Uniform	Non-uniform	
Equal-bit	66.8	73.5	81.6	85.4	89.4	90.3	0.00 ms
Random	48.6	56.8	67.2	73.0	80.2	82.3	0.02 ms
Theory + rounding	72.9	78.8	85.9	88.3	91.4	91.8	0.04 ms
Ours	**75.8**	**81.3**	**87.9**	**89.8**	**91.9**	**92.2**	1.02 ms
Exhaustive search	76.4	82.0	88.5	90.4	92.1	92.4	31.70 min

6.3 Q3: Theory-Guided Precision Allocation

1) Theory-guided configuration selection. We first examine whether the theoretical bounds can identify effective source-wise bit-width configurations. We sample candidate bit-width combinations and measure their task distortion together with the corresponding theoretical lower- and upper-bound scores. The candidates are then ranked by each score, and the average measured distortion of the selected top-$K\%$ configurations is compared with random selection.

As shown in Figure 6, configurations selected by the theoretical bounds consistently achieve lower measured distortion than random selection and closely track the selection based on measured task distortion. Moreover, the approximately monotonic and linear relationship between the theoretical scores and measured task-performance degradation further shows that the proposed characterization provides effective guidance for understanding model behavior under quantization.

2) Practical precision allocation. We next evaluate whether the above theoretical results translates into effective source-wise precision allocation schemes. We compare our method with four baselines: *1) Equal-bit* assigns the same bit-width to all source branches. *2) Random* samples 1000 feasible source-wise allocations under the same budget. *3) Theory + rounding* directly rounds the continuous solution in Proposition 3 to supported bit-widths. *4) Exhaustive search* evaluates all feasible configurations using their empirical task performance and serves as an empirical upper bound.

Table 1 reports the resulting task performance. Our method consistently outperforms equal-bit allocation, with more pronounced gains in the low-precision regime, where the limited bit budget makes effective allocation across heterogeneous sources particularly important. Direct rounding already provides clear gains over equal-bit allocation, confirming that the continuous solution in Proposition 3 captures a useful source-wise allocation principle. Compared with direct rounding, the subsequent discrete refinement further improves task performance and remains close to exhaustive search across all settings. Meanwhile, our method requires

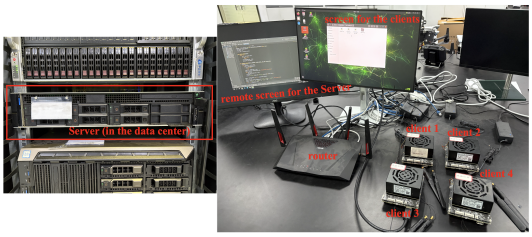


Figure 7: Hardware testbed for evaluating source-wise precision allocation under practical latency and energy constraints.

Table 2: Task performance (%) under practical latency and energy constraints on the edge testbed.

Method	Delay T_0 (s)						Energy E_0 (J)					
	1.7		1.8		1.9		38		40		42	
	Uniform	Non-uniform	Uniform	Non-uniform	Uniform	Non-uniform	Uniform	Non-uniform	Uniform	Non-uniform	Uniform	Non-uniform
Equal-bit	75.7	80.6	83.4	86.6	88.4	89.8	74.2	79.4	82.5	86.0	87.4	89.2
Random	59.8	66.5	70.2	75.3	78.5	81.4	57.9	64.8	68.7	74.1	76.9	80.2
Theory + rounding	80.7	84.5	87.0	89.0	90.0	90.9	79.4	83.6	86.4	88.6	89.5	90.6
Ours	**83.1**	**86.4**	**88.7**	**90.3**	**90.9**	**91.6**	**81.8**	**85.6**	**88.3**	**90.0**	**90.5**	**91.4**
Exhaustive search	83.7	87.0	89.2	90.8	91.2	91.9	82.4	86.2	88.9	90.6	90.8	91.7

substantially less allocation time than exhaustive search. These results demonstrate that the theoretical characterization provides effective guidance for bit allocation while avoiding computationally expensive exhaustive search.

6.4 Q4: Deployment under Practical Constraints

We finally evaluate the proposed allocation under practical latency and energy constraints. The source-specific computation is performed on four NVIDIA Jetson AGX Orin 64GB devices, while the shared downstream task computation is executed on a Dell PowerEdge R740 server equipped with dual Intel Xeon Gold 6246R CPUs and two NVIDIA RTX 3090 GPUs. The detailed experimental setup, including the system modeling and problem formulation, is provided in Appendix F.

Table 2 reports the task performance under different resource constraints. As the latency or energy budget increases, more high-precision configurations become feasible, resulting in consistently improved task performance. Across all constraints, the proposed allocation outperforms equal-bit, random, and direct-rounding baselines under both quantization schemes. The performance gain is more pronounced under tighter resource constraints, which is consistent with the observations in Table 1. These results further demonstrate that the proposed theoretical characterization provides effective guidance for source-wise precision allocation under practical edge-resource constraints.

7 Discussion and Limitations

First, our local-to-global analysis relies on several simplifying assumptions, including a first-order parameter perturbation approximation and bounded layer-wise gradients, to relate parameter distortion to collaborative output distortion. While this provides a tractable source-wise characterization, more general setups, such as incorporating higher-order Taylor terms, may further improve the accuracy of this relationship. Second, the exponential parameter-magnitude model provides analytical tractability but may not perfectly fit all model parameters. More refined modeling, such as layer-wise distributions and mixture models that capture heterogeneous parameter statistics, could provide a more accurate characterization of practical models. Third, the derived rate-distortion bounds depend on both the adopted distortion measure and the assumed parameter distribution. With more accurate distortion measures and refined distributional models, it would be valuable to derive tighter lower and upper bounds. Reducing the gap between these bounds could provide a more precise characterization of practical quantization behavior and further improve theory-guided precision allocation.

8 Conclusion

We studied source-wise model quantization for multi-source co-inference. A local-to-global perturbation analysis establishes an additive source-wise distortion structure, while rate-distortion bounds characterize how source precision affects parameter distortion. By incorporating Shapley-based task contributions, we further derive an explicit precision-allocation law and a practical discrete allocation rule for hardware-supported bit-widths. Experiments validate the theoretical analysis and demonstrate the effectiveness of the proposed method in source-wise precision allocation.

REFERENCES

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

Jiale Chen, Yalda Shabanzadeh, Elvir Crnčević, Torsten Hoefler, and Dan Alistarh. The geometry of LLM quantization: GPTQ as babai's nearest plane algorithm. In *International Conference on Learning Representations*, 2026.

Mohammed Nowaz Rabbani Chowdhury, Kaoutar El Maghraoui, Hsinyu Tsai, Naigang Wang, Geoffrey W. Burr, Liu Liu, and Meng Wang. Efficient quantization of mixture-of-experts with theoretical generalization guarantees. In *International Conference on Learning Representations*, 2026.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186. Association for Computational Linguistics, 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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

EleutherAI. GPT-Neo 125M, 2021. URL `https://huggingface.co/EleutherAI/gpt-neo-125m`. Hugging Face model repository.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023.

Weihao Gao, Yu-Han Liu, Chong Wang, and Sewoong Oh. Rate distortion for model compression: From theory to practice. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2102–2111. PMLR, 2019.

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. Technical Report CNS-TR-2007-001, California Institute of Technology, 2007.

Simla Burcu Harma, Ayan Chakraborty, Elizaveta Kostenok, Danila Mishin, Dongho Ha, Babak Falsafi, Martin Jaggi, Ming Liu, Yunho Oh, Suvinay Subramanian, and Amir Yazdanbakhsh. Effective interplay between sparsity and quantization: From theory to practice. In *International Conference on Learning Representations*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Berivan Isik, Tsachy Weissman, and Albert No. An information-theoretic justification for model pruning. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3821–3846. PMLR, 2022.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, 2024.

Genjia Liu, Anning Hu, Yue Hu, Wenjun Zhang, and Siheng Chen. Rate-distortion optimized pragmatic communication for collaborative perception. In *International Conference on Learning Representations*, 2026.

Yen-Cheng Liu, Junjiao Tian, Nathaniel Glaser, and Zsolt Kira. When2com: Multi-agent perception via communication graph grouping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4106–4115, 2020.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. OmniQuant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations*, 2024.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games II*, pp. 307–317. Princeton University Press, 1953.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1656.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099. PMLR, 2023.

Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *European Conference on Computer Vision*, pp. 107–124. Springer, 2022. doi: 10.1007/978-3-031-19842-7_7.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Sean I. Young. Radio: Rate–distortion optimization for large language model compression. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 72819–72836. PMLR, 2025.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1115.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3391–3401, 2017.

A Proof of the Local-to-Global Distortion Analysis

We first derive the explicit distortion bound for fully connected (FC) networks and then extend the same source-wise structure to general differentiable models through first-order parameter perturbation analysis.

A.1 Fully Connected Networks

Consider an L-layer FC source branch

$$f_k(\boldsymbol{x}_k; \boldsymbol{W}_k) = \boldsymbol{W}_k^{(L)} \sigma\Big(\boldsymbol{W}_k^{(L-1)} \sigma\Big(\cdots \sigma(\boldsymbol{W}_k^{(1)} \boldsymbol{x}_k)\Big)\Big). \tag{7}$$

Let $\boldsymbol{o}_k^{(l)}$ and $\hat{\boldsymbol{o}}_k^{(l)}$ denote the outputs of the l-th layer before and after quantization, respectively. Under Assumptions 1 and 2,

$$\|\boldsymbol{o}_k^{(l-1)}\|_1 \le \prod_{j=1}^{l-1} \|\boldsymbol{W}_k^{(j)}\|_1. \tag{8}$$

Using the triangle inequality and the 1-Lipschitz property of $\sigma(\cdot)$, the distortion at layer l satisfies

$$\|\boldsymbol{o}_k^{(l)} - \hat{\boldsymbol{o}}_k^{(l)}\|_1 \le \|\boldsymbol{W}_k^{(l)} - \hat{\boldsymbol{W}}_k^{(l)}\|_1 \|\boldsymbol{o}_k^{(l-1)}\|_1 + \|\hat{\boldsymbol{W}}_k^{(l)}\|_1 \|\boldsymbol{o}_k^{(l-1)} - \hat{\boldsymbol{o}}_k^{(l-1)}\|_1. \tag{9}$$

From Assumption 3,

$$\|\hat{\boldsymbol{W}}_k^{(l)}\|_1 \le \|\boldsymbol{W}_k^{(l)}\|_1 + \tau_k^{(l)}. \tag{10}$$

Recursively applying (8)–(10) from $l = 1$ to L gives

$$\|\boldsymbol{o}_k - \hat{\boldsymbol{o}}_k\|_1 \le \sum_{l=1}^{L} A_k^{(l)} \|\boldsymbol{W}_k^{(l)} - \hat{\boldsymbol{W}}_k^{(l)}\|_1, \tag{11}$$

where

$$A_k^{(l)} = \prod_{j=1}^{l-1} \|\boldsymbol{W}_k^{(j)}\|_1 \prod_{r=l+1}^{L} \Big(\|\boldsymbol{W}_k^{(r)}\|_1 + \tau_k^{(r)}\Big). \tag{12}$$

We next consider an N-layer FC shared task executor

$$h(\boldsymbol{u}; \boldsymbol{V}) = \boldsymbol{V}^{(N)} \sigma\Big(\boldsymbol{V}^{(N-1)} \sigma\Big(\cdots \sigma(\boldsymbol{V}^{(1)} \boldsymbol{u})\Big)\Big). \tag{13}$$

By the same Lipschitz argument,

$$\|h(\boldsymbol{u}; \boldsymbol{V}) - h(\hat{\boldsymbol{u}}; \boldsymbol{V})\|_1 \le C_h \|\boldsymbol{u} - \hat{\boldsymbol{u}}\|_1, \qquad C_h = \prod_{n=1}^{N} \|\boldsymbol{V}^{(n)}\|_1. \tag{14}$$

From Assumption 4,

$$\|\boldsymbol{u} - \hat{\boldsymbol{u}}\|_1 \le \sum_{k=1}^{K} L_k \|\boldsymbol{o}_k - \hat{\boldsymbol{o}}_k\|_1. \tag{15}$$

Combining (11)–(15) yields

$$\|\boldsymbol{z} - \hat{\boldsymbol{z}}\|_1 \le C_h \sum_{k=1}^{K} L_k \sum_{l=1}^{L} A_k^{(l)} \|\boldsymbol{W}_k^{(l)} - \hat{\boldsymbol{W}}_k^{(l)}\|_1. \tag{16}$$

This establishes Corollary 1 and the local-to-global distortion structure for FC networks.

A.2 Extension to General Models

We next extend the same structure to general differentiable source branches. For source k, a first-order Taylor expansion around $\boldsymbol{W}_k$ gives

$$f_k(\boldsymbol{x}_k;\hat{\boldsymbol{W}}_k) - f_k(\boldsymbol{x}_k;\boldsymbol{W}_k) \approx \sum_{l=1}^{L} \nabla_{\boldsymbol{W}_k^{(l)}} f_k(\boldsymbol{x}_k;\boldsymbol{W}_k) : \left(\hat{\boldsymbol{W}}_k^{(l)} - \boldsymbol{W}_k^{(l)}\right), \tag{17}$$

where : denotes tensor contraction after vectorization. Under Assumption 5, the resulting source-output distortion is approximately upper bounded by

$$\|\boldsymbol{o}_k - \hat{\boldsymbol{o}}_k\|_1 \le \sum_{l=1}^{L} A_k^{(l)} \|\boldsymbol{W}_k^{(l)} - \hat{\boldsymbol{W}}_k^{(l)}\|_1. \tag{18}$$

Finally, applying Assumption 4 gives

$$\|\boldsymbol{z} - \hat{\boldsymbol{z}}\|_1 \le C_h \sum_{k=1}^{K} L_k \|\boldsymbol{o}_k - \hat{\boldsymbol{o}}_k\|_1 \le C_h \sum_{k=1}^{K} L_k \sum_{l=1}^{L} A_k^{(l)} \|\boldsymbol{W}_k^{(l)} - \hat{\boldsymbol{W}}_k^{(l)}\|_1. \tag{19}$$

This proves Proposition 1.

B Derivation of the Multi-Source Rate-Distortion Bounds

We provid the derivation of Proposition 2. First, we derive the rate-distortion bounds for a single source and then extend the results to the multi-source setting.

For source k, let $\Theta_k \sim \mathrm{Exp}(\lambda_k)$ denote its parameter magnitude. Under the absolute-error distortion $d(\theta,\hat{\theta}) = |\theta - \hat{\theta}|$, the rate–distortion function is defined as

$$R_k(D) = \inf_{p(\hat{\theta}|\theta):\mathbb{E}[|\Theta_k - \hat{\Theta}_k|]\le D} I(\Theta_k;\hat{\Theta}_k). \tag{20}$$

Correspondingly, $D_k(R)$ denotes the minimum achievable expected distortion at rate R. The task-aware multi-source distortion in Proposition 2 is formally defined as

$$\mathcal{D}_{\boldsymbol{\alpha}}(\boldsymbol{R}) = \sum_{k=1}^{K} \alpha_k q_k D_k(R_k). \tag{21}$$

B.1 Single-Source Lower Bound

For notational simplicity, we omit the source index and consider $\Theta \sim \mathrm{Exp}(\lambda)$. Let $Z = \Theta - \hat{\Theta}$ denote the reconstruction error. For any feasible $p(\hat{\theta}|\theta)$,

$$I(\Theta;\hat{\Theta}) = h(\Theta) - h(\Theta|\hat{\Theta}) = h(\Theta) - h(Z|\hat{\Theta}) \ge h(\Theta) - h(Z), \tag{22}$$

where the last inequality follows because conditioning does not increase differential entropy.

For the exponential distribution,

$$h(\Theta) = \log_2 \frac{\mathrm{e}}{\lambda}. \tag{23}$$

Moreover, among all real-valued random variables satisfying $\mathbb{E}[|Z|] \le D$, the differential entropy is maximized by the zero-mean Laplacian distribution

$$p_Z(z) = \frac{1}{2D} e^{-|z|/D}, \tag{24}$$

which yields

$$h(Z) \le \log_2(2\mathrm{e}D). \tag{25}$$

Combining (22)–(25), we obtain

$$R(D) \ge [-\log_2(2\lambda D)]^+ \triangleq R^{\mathrm{L}}(D). \tag{26}$$

Equivalently, the corresponding distortion–rate lower bound is

$$D(R) \ge \frac{1}{\lambda 2^{R+1}} \triangleq D^{\mathrm{L}}(R). \tag{27}$$

B.2 Single-Source Upper Bound

We next construct a feasible test channel

$$\hat{\Theta} = \Theta + Z, \tag{28}$$

where Z is independent of Θ and follows a zero-mean Laplacian distribution with $\mathbb{E}[|Z|] = D$. The resulting expected distortion is therefore D. By the independence of Θ and Z,

$$R(D) \le I(\Theta; \hat{\Theta}) = h(\Theta + Z) - h(Z). \tag{29}$$

Using the maximum-entropy property under a first absolute-moment constraint, we have

$$h(\Theta + Z) \le \log_2 \left(2\mathrm{e}\mathbb{E}[|\Theta + Z|]\right). \tag{30}$$

For any fixed $\theta \ge 0$,

$$\mathbb{E}_Z[|\theta + Z|] = \theta + De^{-\theta/D}. \tag{31}$$

Averaging over $\Theta \sim \mathrm{Exp}(\lambda)$ gives

$$\mathbb{E}[|\Theta + Z|] = \int_0^\infty \left(\theta + De^{-\theta/D}\right) \lambda e^{-\lambda\theta} d\theta = \frac{1}{\lambda} + D\frac{\lambda D}{1 + \lambda D}. \tag{32}$$

Substituting (32) into (29), together with $h(Z) = \log_2(2\mathrm{e}D)$, yields

$$R(D) \le \log_2 \left(\frac{1}{\lambda D} + \frac{\lambda D}{1 + \lambda D}\right) \triangleq R^{\mathrm{U}}(D). \tag{33}$$

Then the achievable distortion–rate upper bound is

$$D(R) \le \frac{1}{2\lambda} \left(\sqrt{1 + \frac{4}{2^R - 1}} - 1\right) \triangleq D^{\mathrm{U}}(R), \quad R > 0. \tag{34}$$

B.3 Extension to Multiple Sources

Applying (27) and (34) to source k gives

$$\frac{1}{\lambda_k 2^{R_k+1}} \le D_k(R_k) \le \frac{1}{2\lambda_k} \left(\sqrt{1 + \frac{4}{2^{R_k} - 1}} - 1\right). \tag{35}$$

Since $\alpha_k q_k \ge 0$, multiplying (35) by $\alpha_k q_k$ and summing over all sources yields

$$\sum_{k=1}^{K} \frac{\alpha_k q_k}{\lambda_k 2^{R_k+1}} \le \mathcal{D}_{\boldsymbol{\alpha}}(\boldsymbol{R}) \le \sum_{k=1}^{K} \frac{\alpha_k q_k}{2\lambda_k} \left(\sqrt{1 + \frac{4}{2^{R_k} - 1}} - 1\right), \tag{36}$$

this thus completes the proof of Proposition 2.

C Shapley-Based Source Contribution

For a subset of sources $S \subseteq \mathcal{K}$, given an evaluation dataset with N samples, we use the downstream task performance as the coalition performance $\ell(S)$. The corresponding task utility gain relative to the empty-source baseline is defined as $v(S) = \ell(S) - \ell(\emptyset)$.

Exact Shapley evaluation requires considering all possible source coalitions and becomes expensive as K increases. We therefore use Monte Carlo permutation sampling. Let $\pi^{(m)}$ denote the m-th uniformly sampled permutation of the K sources, and let $P_k^{(m)}$ denote the set of sources appearing before source k in $\pi^{(m)}$. The Shapley contribution is estimated as

$$\hat{\phi}_k = \frac{1}{M} \sum\nolimits_{m=1}^{M} \left[v\left(P_k^{(m)} \cup \{k\}\right) - v\left(P_k^{(m)}\right)\right], \tag{37}$$

where M denotes the number of sampled permutations. Under uniform permutation sampling, (37) is an unbiased estimator of the exact Shapley value and requires $O(MK)$ marginal-utility evaluations.

The task contribution coefficient is then obtained through nonnegative normalization as $\hat{\alpha}_k = [\hat{\phi}_k]^+ / \sum_{j=1}^{K} [\hat{\phi}_j]^+$. From a PID perspective, the coalition-based marginal contribution naturally captures unique, redundant, and synergistic information among sources.

D Proof of Proposition 3

Using the lower rate–distortion bound, the continuous precision-allocation problem is

$$\min_{\{R_k \geq 0\}} \quad \sum_{k=1}^{K} \frac{\alpha_k q_k}{2\lambda_k} 2^{-R_k}$$
$$\text{s.t.} \quad \sum_{k=1}^{K} q_k R_k \leq B. \tag{38}$$

The objective is convex in $\{R_k\}$ and the constraint is linear, so the KKT conditions characterize the global optimum.

Introducing the dual variable $\mu \geq 0$ for the bit-budget constraint, for an active source with $R_k > 0$, the stationarity condition gives

$$-\frac{\alpha_k q_k \ln 2}{2\lambda_k} 2^{-R_k} + \mu q_k = 0. \tag{39}$$

Therefore,

$$R_k = \log_2 \left(\frac{\alpha_k}{\nu \lambda_k} \right), \qquad \nu \triangleq \frac{2\mu}{\ln 2}, \tag{40}$$

where q_k cancels because it appears in both the distortion objective and the bit-budget constraint. Accounting for the nonnegativity constraint gives

$$R_k^\star = \left[\log_2 \left(\frac{\alpha_k}{\nu \lambda_k} \right) \right]^+. \tag{41}$$

Since the objective decreases with R_k for $\alpha_k > 0$, the available bit budget is fully used at the optimum. Hence, $\nu > 0$ is selected such that

$$\sum_{k=1}^{K} q_k \left[\log_2 \left(\frac{\alpha_k}{\nu \lambda_k} \right) \right]^+ = B. \tag{42}$$

This completes the proof.

E Additional Experimental Results

To further examine whether the observations in Q1 and Q2 generalize beyond the main ViT-Tiny+Qwen setting, we repeat the corresponding experiments using the ViT-Small co-inference model on Caltech-256 described in Sec. 6. We consider six sources observing the normalized horizontal image regions $[0.00, 0.18]$, $[0.12, 0.38]$, $[0.30, 0.55]$, $[0.45, 0.70]$, $[0.62, 0.88]$, and $[0.82, 1.00]$, respectively. The model consists of six source-specific branches, whose representations are fused and subsequently processed by the shared downstream ViT layers.

Source contribution and local-to-global distortion. Figure 8 shows the Shapley-based contributions of the six sources. Similar to the main experiment, the sources exhibit clearly heterogeneous task contributions, further motivating source-dependent precision allocation.

Figures 9 and 10 compare source-wise parameter distortion with the resulting collaborative output distortion under uniform and non-uniform K-means quantization, respectively. In both cases, the output distortion follows the variation of parameter distortion as the source bit-width changes, exhibiting the same local-to-global distortion trend observed in the main ViT-Tiny+Qwen experiment.

Rate–distortion characterization. Figures 11 and **??** further compare the empirical source-wise parameter distortion with the analytical lower and achievable upper rate-distortion bounds. Under both quantization schemes, the empirical distortion decreases consistently with bit-width and follows the trend characterized by the theoretical bounds. These observations are consistent with the results in Sec. 6.2 and further support the proposed rate-distortion characterization in a different model and dataset setting.

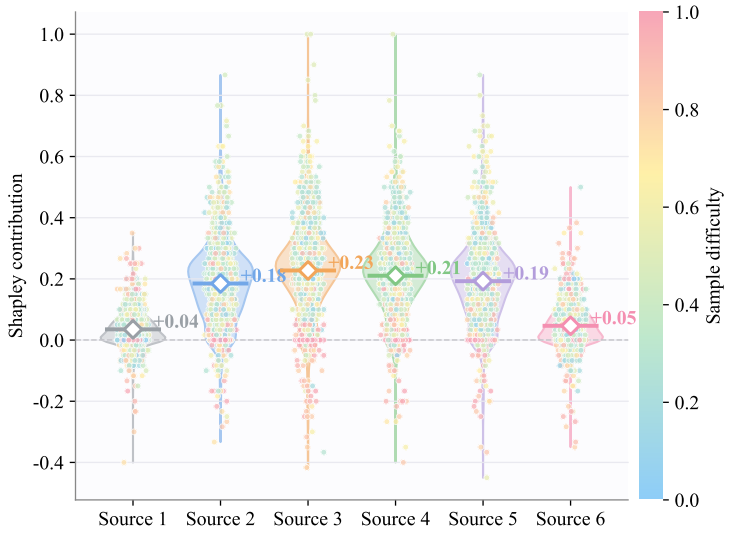


Figure 8: Shapley-based source contributions across the six sources in the ViT-Small co-inference model.

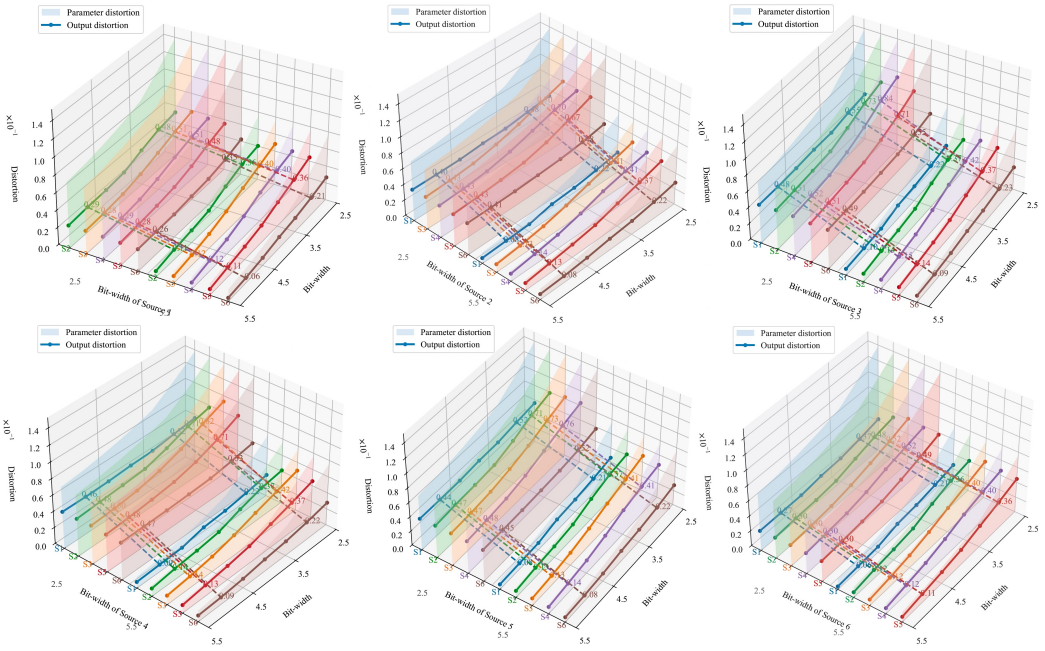


Figure 9: Source-wise parameter distortion and the resulting collaborative output distortion under uniform quantization.

Theory-guided configuration selection. We further repeat the configuration-selection experiment in Q3 using the ViT-Small co-inference model. Figure 13 shows the results under uniform and non-uniform K-means quantization. Consistent with the main ViT-Tiny+Qwen experiment, configurations selected by the theoretical bounds exhibit lower measured task distortion than random selection and closely follow the selection based on measured distortion. Moreover, the theoretical scores maintain an approximately monotonic relationship with the measured task-performance degradation. These results further show that the proposed theoretical characterization provides consistent guidance for source-wise bit allocation.

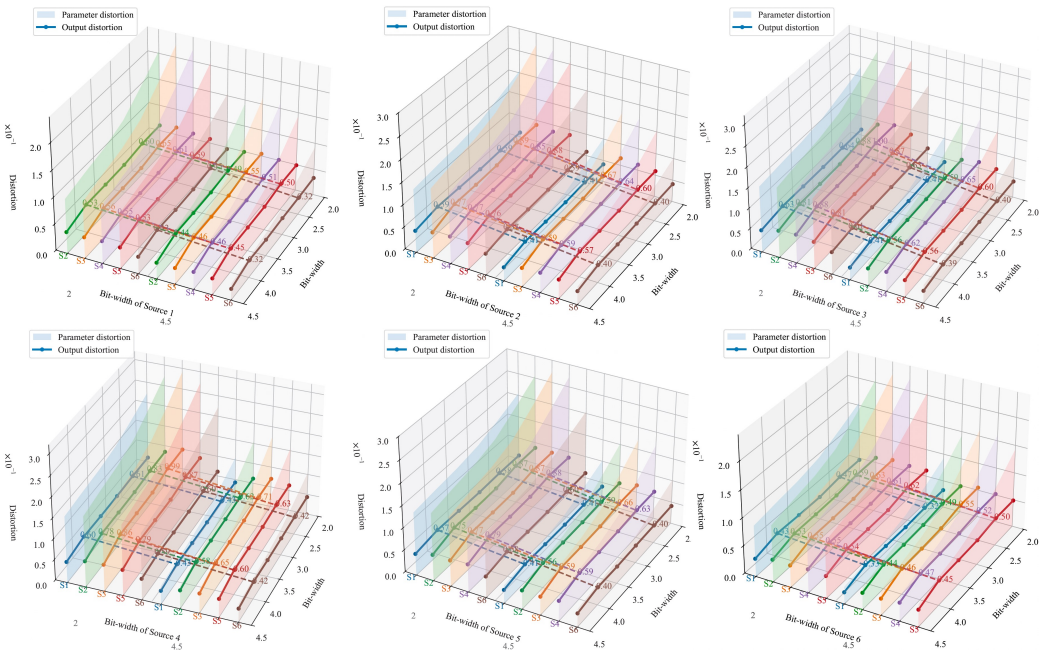


Figure 10: Source-wise parameter distortion and the resulting collaborative output distortion under non-uniform K-means quantization.

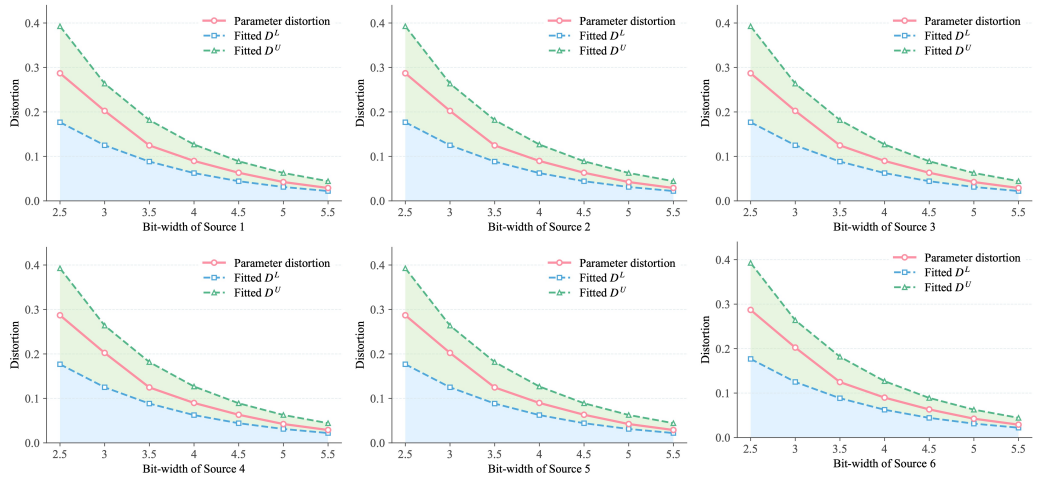


Figure 11: Empirical source-wise parameter distortion and the analytical lower and achievable upper rate–distortion bounds under uniform quantization.

F Edge Deployment under Practical Resource Constraints

We extend the proposed source-wise precision allocation to the practical latency and energy constraints considered in Sec. 6.4. We first model the computation latency and energy consumption, and then formulate the resource-constrained precision-allocation problem and discuss its solution.

1) Computation latency and energy. Let b_k denote the quantization bit-width assigned to source k, and let f_k denote its fixed processor frequency. Let N_k^{FLOP} denote the full-precision workload, b_{fp} the reference full-precision bit-width, and c_k the processor throughput in FLOPs per cycle. The

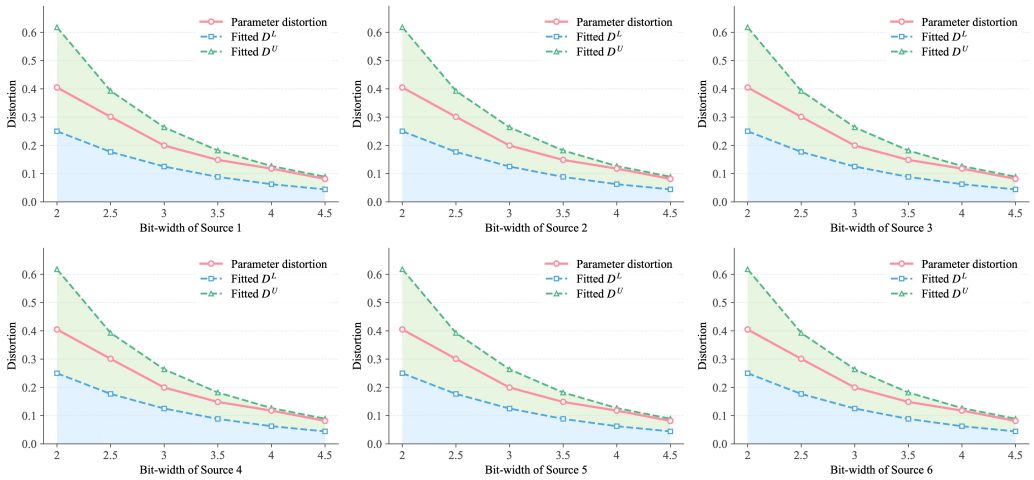


Figure 12: Empirical source-wise parameter distortion and the analytical lower and achievable upper rate–distortion bounds under non-uniform K-means quantization.

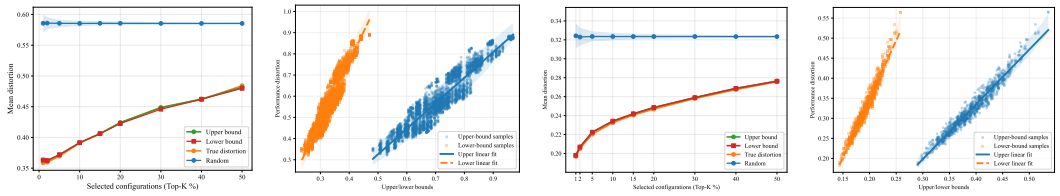


Figure 13: Theory-guided configuration selection and correspondence between theoretical bounds and measured task-performance degradation for the ViT-Small co-inference model. From left to right: configuration selection and bound-performance correspondence under uniform and non-uniform K-means quantization, respectively.

computation latency of source k is modeled as

$$t_k^{\mathrm{cmp}}(b_k) = \frac{b_k N_k^{\mathrm{FLOP}}}{b_{\mathrm{fp}} f_k c_k}. \tag{43}$$

With power coefficient ψ_k and PUE factor η_k, the corresponding computation energy is

$$e_k^{\mathrm{cmp}}(b_k) = \eta_k \frac{b_k N_k^{\mathrm{FLOP}}}{b_{\mathrm{fp}} c_k} \psi_k f_k^2. \tag{44}$$

The shared downstream task is executed at the server with fixed frequency $\widetilde{f}$. Let $\widetilde{N}^{\mathrm{FLOP}}$, $\widetilde{c}$, $\widetilde{\eta}$, and $\widetilde{\psi}$ denote the server-side workload, processor throughput, PUE factor, and power coefficient, respectively. The corresponding latency and energy are

$$\widetilde{t}^{\mathrm{cmp}} = \frac{\widetilde{N}^{\mathrm{FLOP}}}{\widetilde{f}\widetilde{c}}, \qquad \widetilde{e}^{\mathrm{cmp}} = \widetilde{\eta}\frac{\widetilde{N}^{\mathrm{FLOP}}}{\widetilde{c}} \widetilde{\psi}\widetilde{f}^2. \tag{45}$$

Since the source-specific computations are performed in parallel, the end-to-end computation latency and total energy consumption are

$$T(\boldsymbol{b}) = \max_k t_k^{\mathrm{cmp}}(b_k) + \widetilde{t}^{\mathrm{cmp}}, \qquad E(\boldsymbol{b}) = \sum_{k=1}^{K} e_k^{\mathrm{cmp}}(b_k) + \widetilde{e}^{\mathrm{cmp}}. \tag{46}$$

Fixed preprocessing and idle-power costs can be incorporated into these expressions as constant terms without changing the bit-allocation structure.

2) Problem formulation and solution. Proposition 2 provides the lower and upper rate–distortion bounds $D_k^{\mathrm{L}}(b_k)$ and $D_k^{\mathrm{U}}(b_k)$ for source k. We jointly minimize the two distortion bounds by optimizing the source-wise bit allocation $\boldsymbol{b}$ under practical latency and energy constraints. The resulting problem is

$$\begin{aligned} \min_{\boldsymbol{b}} \quad & \sum_{k=1}^{K} \alpha_k q_k \left[D_k^{\mathrm{L}}(b_k) + D_k^{\mathrm{U}}(b_k) \right] \\ \text{s.t.} \quad & T(\boldsymbol{b}) \le T_0, \qquad (47) \\ & E(\boldsymbol{b}) \le E_0. \qquad (48) \end{aligned}$$

Since the practical bit-widths $\boldsymbol{b}$ are discrete, we first relax the discrete bit-allocation variables to continuous ones. The resulting relaxed problem is convex and can be efficiently solved using CVX, yielding the continuous solution $\{b_k^{\mathrm{c}}\}_{k=1}^{K}$. We then recover practical discrete bit-widths using the refinement procedure in Algorithm 1. Specifically, a small set of supported bit-widths around each b_k^{c} is retained, and the feasible combination minimizing $\sum_k \alpha_k q_k [D_k^{\mathrm{L}}(b_k) + D_k^{\mathrm{U}}(b_k)]$ under the latency and energy constraints is selected.

The proposed allocation only evaluates the analytical distortion characterization and does not require repeated model inference. In contrast, exhaustive search evaluates the empirical task performance of all feasible bit-width configurations and therefore incurs substantially higher configuration-search overhead.

Reproducibility Statement

To facilitate reproducibility, we state the modeling assumptions and distortion definitions in Sections 3 and 4, and provide the derivations of the local-to-global distortion characterization, rate–distortion bounds, and continuous precision-allocation law in Appendices A, B, and D, respectively. The coalition utility, Monte Carlo Shapley estimator, and nonnegative normalization used to obtain source contributions are described in Section 4 and Appendix C. The theory-guided discrete allocation procedure is specified in Section 5 and Algorithm 1. For the empirical evaluation, Section 6 identifies the datasets, model backbones, and post-training quantization schemes, describes the four-source image partitioning, and defines the allocation baselines and evaluation budgets. Additional source-contribution results are presented in Appendix E. The edge hardware is specified in Section 6.4, while the latency and energy models, resource constraints, and continuous-to-discrete optimization procedure are given in Appendix F.

AI Use Statement

Generative AI tools were used only for language polishing, prose revision, and the formatting of tables and references. The methodology, theoretical development, proofs, and experimental design were developed and conducted by the authors without substantive AI assistance. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the paper's text, technical claims, proofs, and experimental results.