CONFIGURATION-VALUED EXPERT IMPORTANCE FOR MIXED-PRECISION QUANTIZATION OF MULTIMODAL MIXTURE-OF-EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse activation does not eliminate the expert-weight storage cost of multimodal Mixture-of-Experts (MoE) models, motivating mixed-precision quantization under memory constraints. However, existing importance-based approaches often rely on isolated expert or projection scores, overlooking the context dependence of expert importance, the natural hierarchy between experts and their internal projections, and the coupled quantization effects within each expert. We propose Config-Owen, a hierarchical game-theoretic framework for mixed-precision MoE quantization. Config-Owen formulates expert-level precision allocation as a cooperative game, uses Owen-style attribution to preserve the expert–projection hierarchy, and incorporates within-expert Harsanyi interactions to capture coupled quantization effects, yielding configuration-valued utilities for precision allocation. We efficiently estimate these utilities and optimize the resulting precision assignment under a global memory budget. Across four multimodal MoE models and Mixtral-8×7B, Config-Owen consistently improves low-bit performance preservation; on DeepSeek-VL2-S at 2.58 bits, it retains 98.39% of the FP16 five-task average score. Code will be released upon acceptance.

1 INTRODUCTION

Multimodal Mixture-of-Experts (MoE) models (Shazeer et al., 2017) expand model capacity while activating only a subset of routing experts for each input token (Lepikhin et al., 2021; Fedus et al., 2022; Zoph et al., 2022). However, sparse activation does not by itself reduce the storage required for the expert pool (Rajbhandari et al., 2022). In multimodal systems, expert weights share a memory budget with the visual encoder, cross-modal projector, and other language-model components. Post-training quantization offers a practical way to reduce this footprint without retraining (Yao et al., 2022; Dettmers et al., 2022; Xiao et al., 2023); however, aggressive compression makes the allocation of limited precision particularly important.

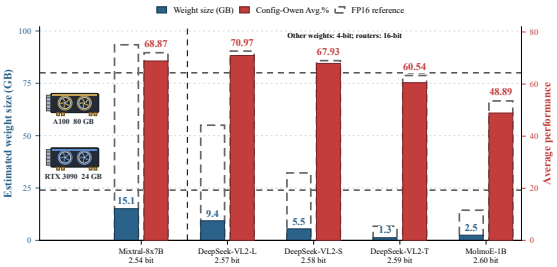


Figure 1: Weight storage and performance across MoE models.

Existing MoE mixed-precision methods generally exploit heterogeneous quantization sensitivity to allocate higher precision to more important components. PMQ (Huang et al., 2026) performs importance-driven allocation at the expert level, while finer-grained methods such as Auto Fine-Grained (Xie et al., 2025) further assign precision below the expert level using sensitivity information. Despite this progress, important structural dependencies remain insufficiently captured. First, expert quantization perturbs the hidden states propagated through subsequent MoE layers, coupling its effect with the precision states of other experts. Such cross-expert contextual dependence is not explicitly captured by isolated expert-wise importance. Second, expert-level importance is too coarse for fine-grained precision allocation, whereas directly moving to projection-level scoring overlooks the natural hierarchical organization between experts and their internal projections. Third, the gated computation within each MoE expert intrinsically couples its

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.

internal projections, such that quantization perturbations from different projections can interact rather than contribute independently. Consequently, projection-wise importance may not fully reflect the utility of their joint precision choice.

To address these limitations, we propose **Config-Owen**, a hierarchical game-theoretic framework for mixed-precision MoE quantization. We first formulate expert-level precision allocation as a cooperative game, evaluating each expert through its marginal contribution under different expert precision contexts. We then model experts and their internal projections as a natural hierarchical coalition structure and employ Owen-style attribution to connect expert-level context with projection-level precision decisions. To capture the interactions induced by the coupled expert computation, we further incorporate within-expert Harsanyi interactions, yielding a configuration-valued utility for each candidate precision choice. Using the Harsanyi reconstruction identity, this utility admits an equivalent endpoint representation that avoids explicitly constructing individual interaction terms. We estimate the resulting utilities over sampled expert-ordering contexts and optimize the final precision assignment through a multiple-choice knapsack formulation under a global memory budget. Figure 1 summarizes the resulting weight storage and benchmark performance across the evaluated MoE models.

Our contributions are threefold. **(1)** We formulate expert-level mixed-precision quantization as a cooperative game, enabling expert importance to account for cross-expert precision context rather than relying on isolated scalar scores. **(2)** We identify the expert–projection organization of MoE blocks as a natural hierarchical coalition structure and introduce Owen-style attribution to connect expert-level context with projection-level precision decisions. **(3)** We incorporate within-expert Harsanyi interactions to capture coupled quantization effects, yielding a configuration-valued utility for precision allocation. We evaluate Config-Owen on four multimodal MoE models and Mixtral-8×7B, demonstrating consistent performance preservation under low-bit memory budgets.

2 Related Work

Post-training and mixed-precision quantization. Post-training quantization for large transformers includes second-order weight reconstruction (Frantar et al., 2023), activation-aware scaling (Lin et al., 2024), outlier-aware representations (Dettmers et al., 2024; Kim et al., 2024), and calibrated quantization parameters (Shao et al., 2024). Recent ICLR work improves low-bit quantizability through adaptive channel reassembly, affine or orthogonal transformations, and learned rotations (Liu et al., 2024a; Ma et al., 2024; Liu et al., 2025; Hu et al., 2025), reconstructs coupled low-bit errors across blocks (Ding et al., 2025), learns loss-aware quantization grids (Zhang & Shrivastava, 2025), and compensates quantization error with analytically derived low-rank terms (Zhang et al., 2025). Mixed-precision methods allocate bits according to curvature, salience, or hardware constraints (Dong et al., 2020; Yao et al., 2021; Chen et al., 2021; Huang et al., 2025c).

Quantization for MoE models. MoE-specific work further exploits expert heterogeneity: QMoE targets extreme expert compression (Frantar & Alistarh, 2024); MC introduces importance-driven pre-loading mixed-precision quantization for language MoEs (Huang et al., 2025b); MxMoE co-designs allocation and kernels (Duanmu et al., 2025); MoEQuant addresses expert-dependent calibration imbalance (Chen et al., 2025); and MiLo augments aggressively quantized experts with low-rank compensators (Huang et al., 2025a). Recent analyses identify a small subset of compression-sensitive super experts (Su et al., 2026), while KBVQ-MoE removes cross-expert redundancy before vector quantization (Xu et al., 2026). Other methods use router statistics, theoretical generalization criteria, channel-level search, or Hessian sensitivity for precision assignment (Chowdhury et al., 2026; Xie et al., 2025; Chitty-Venkata et al., 2025). MC# (Huang et al., 2026) extends the MC protocol to multimodal MoEs; its Pre-Loading Mixed-Precision Quantization (PMQ) stage provides the closest experimental baseline here, evaluating comparable average-bit budgets and assigning precision per expert through an importance-driven scalar surrogate. In contrast, Config-Owen values each feasible within-expert precision configuration.

Game-theoretic attribution with coalition structure. Cooperative game theory models a set of players that jointly produce a utility and distributes credit according to their contributions. The Shapley value assigns each player its average marginal contribution over all possible joining orders and has been used for data valuation and MoE expert attribution (Ghorbani & Zou, 2019; Huang et al., 2025d). The Owen value generalizes the Shapley value to predefined coalitions by accounting for

both the ordering among coalitions and the ordering of players within each coalition (Owen, 1977), matching the expert–projection hierarchy of MoE models.

3 METHOD

Method overview. Figure 2 summarizes the proposed configuration-valued utility computation and allocation pipeline. We formulate MoE mixed-precision quantization as a two-level cooperative game, where routing experts act as outer players and the up, gate, and down projections inside each expert form an a priori union induced by the gated FFN structure. Under this hierarchy, Owen-style attribution captures expert-level ordering contexts, while Harsanyi dividends represent within-expert projection interactions. Rather than computing isolated projection scores, our method evaluates each candidate local precision configuration by comparing it with the baseline state under the same sampled outer context. The resulting configuration-utility table is estimated using CASP to discard collapsed suffix contexts and RPOS to prioritize informative expert orders, followed by budget-constrained selection through a multiple-choice knapsack formulation.

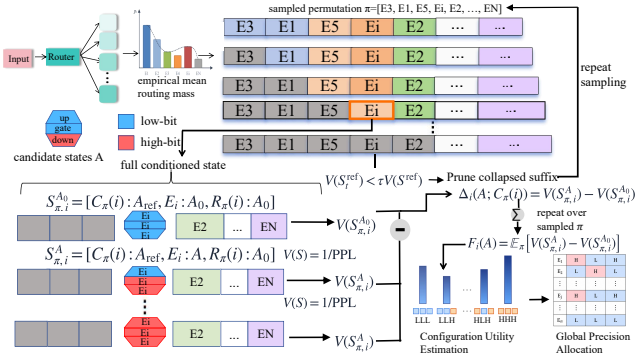


Figure 2: Config-Owen utility computation and mixed-precision allocation pipeline.

3.1 PRELIMINARIES

MoE expert structure. Consider a multimodal MoE model with L expert layers. For input $x \in \mathbb{R}^d$, the ℓ-th expert layer computes $\sum_{k=1}^{n_\ell} G_k^{(\ell)}(x)\, E_k^{(\ell)}(x)$, where $E_k^{(\ell)}$ is the k-th routing expert and $G_k^{(\ell)}(x)$ is its router weight. Each routing expert is a gated FFN; thus, we have:

$$E_i(h) = W_i^{\text{down}} \left(\sigma(W_i^{\text{gate}} h) \odot (W_i^{\text{up}} h)\right). \tag{1}$$

Mixed-precision configuration space. Let $\mathcal{E} = \{1, \ldots, N\}$ index the routing experts, and let $\mathcal{J} = \{u, g, d\}$ denote the up, gate, and down projections inside each expert. Each projection $j \in \mathcal{J}$ of expert $i \in \mathcal{E}$ takes one of two precision states, $\mathcal{B} = \{b_{\text{low}}, b_{\text{high}}\}$. The local configuration of expert i is $A_i = (a_i^j)_{j \in \mathcal{J}} \in \mathcal{B}^{|\mathcal{J}|}$, and the global configuration is $S = (A_i)_{i \in \mathcal{E}}$. We write A^0 and A^{ref} for the local baseline and high-bit anchor configurations, where all projections are assigned b_{low} in A^0 and b_{high} in A^{ref}; their global counterparts are S^0 and S^{ref}. In our experiments, $b_{\text{low}} = 2$ and $b_{\text{high}} = 4$; "FP16 reference" denotes the separate full-precision evaluation baseline. The calibration value function is $V_{\mathcal{C}}(S) := 1/\text{PPL}_{\mathcal{C}}(S)$, where $\text{PPL}_{\mathcal{C}}(S)$ denotes perplexity over the calibration examples. The term $m_i(A_i)$ denotes the memory cost of expert i under local configuration A_i. Since the main setting has two precision states, we sometimes identify a local configuration $A \in \mathcal{B}^{|\mathcal{J}|}$ with the subset of projections assigned b_{high}; under this notation, $A = \emptyset$ denotes A^0 and $A = \mathcal{J}$ denotes A^{ref}.

3.2 CONFIGURATION-VALUED LOCAL UTILITY

Mixed-precision allocation must assign a joint local state to the three projections inside each expert, so the utility should evaluate coupled components rather than isolated projections. Cooperative games with a priori unions provide a direct model for this structure: routing experts are outer units, while the up, gate, and down projections inside one expert form an inner union induced by the gated product.

Owen value under expert-projection unions. The Owen value is a value allocation rule for cooperative games whose players are partitioned into pre-defined unions. It evaluates a player's

contribution over two nested orders: the order in which unions enter the coalition, and the order in which players inside the selected union enter. For expert i, the first-order Owen attribution of projection $j \in \mathcal{J}$ is defined as:

$$\phi_{i,j}^{\mathrm{Owen}} = \frac{1}{|\mathcal{E}|!\,|\mathcal{J}|!} \sum_{\pi \in \Pi(\mathcal{E})} \sum_{\sigma \in \Pi(\mathcal{J})} \left[V_{\mathcal{C}}\left(S_{+}^{(\pi,\sigma,j)}\right) - V_{\mathcal{C}}\left(S_{-}^{(\pi,\sigma,j)}\right) \right], \tag{2}$$

where π is an outer permutation of experts and σ is an inner permutation of projections. The states $S_{\pm}^{(\pi,\sigma,j)}$ differ only at projection j of expert i: in the plus state, this projection is upgraded; in the minus state, it remains at the baseline precision. Experts following i under π and projections preceding j under σ are placed at the upgraded precision. Reversing the uniformly averaged outer order leaves the Owen value unchanged, while this suffix convention aligns the computation with the full-to-small traversal used below. Thus, $\phi_{i,j}^{\mathrm{Owen}}$ measures the calibration gain of upgrading projection j under nested expert-projection orders.

This first-order value respects the hierarchy; however, mixed-precision allocation selects a complete local configuration rather than one projection score at a time. Because the gated FFN multiplies the gate and up branches before the down projection, upgrading one projection may depend on the precision states of the others. We therefore augment the Owen-style attribution with within-union interaction terms.

Within-union Harsanyi interactions. Fix an outer permutation π, and let $C_\pi(i) = \{k \in \mathcal{E} : k \text{ follows } i \text{ in } \pi\}$. For any local state $A \subseteq \mathcal{J}$, define $v_\pi(A) := V_{\mathcal{C}}(S_{C_\pi(i)}^{i \leftarrow A})$, where $S_{C_\pi(i)}^{i \leftarrow A}$ is the global configuration in which experts in $C_\pi(i)$ use the reference state, expert i uses high precision on the projection subset A and low precision on $\mathcal{J} \setminus A$, and all remaining experts use the baseline state.

For this fixed outer context, $v_\pi : 2^{\mathcal{J}} \to \mathbb{R}$ is a three-player local cooperative game over u, g, d. Its Harsanyi expansion decomposes the configuration value into first-order, pairwise, and third-order projection effects. We define the interaction-complete configuration utility as:

$$F_i(A) = \frac{1}{|\mathcal{E}|!} \sum_{\pi \in \Pi(\mathcal{E})} \Bigg[\underbrace{\sum_{j \in A} \left(v_\pi(\{j\}) - v_\pi(\emptyset)\right)}_{\text{first-order effects}} + \underbrace{\sum_{\substack{j,k \in A \\ j<k}} \left(v_\pi(\{j,k\}) - v_\pi(\{j\}) - v_\pi(\{k\}) + v_\pi(\emptyset)\right)}_{\text{pairwise interactions}}$$
$$\underbrace{+ \mathbf{1}[A = \mathcal{J}] \left(v_\pi(\mathcal{J}) - \sum_{|B|=2} v_\pi(B) + \sum_{j \in \mathcal{J}} v_\pi(\{j\}) - v_\pi(\emptyset) \right)}_{\text{third-order interaction}} \Bigg], \tag{3}$$

where $j < k$ follows the fixed ordering $u < g < d$.

Endpoint representation of configuration utility. The Harsanyi expansion in Eq. 3 identifies the interaction structure activated by a local configuration. We now show that the same quantity admits a simpler endpoint representation. For a fixed outer state $C_\pi(i)$, the two values $v_\pi(A)$ and $v_\pi(\emptyset)$ differ only in the local precision state of expert i; their difference is the gain of assigning configuration A to expert i. Averaging over all outer expert orders gives:

$$\widetilde{F}_i(A) = \frac{1}{|\mathcal{E}|!} \sum_{\pi \in \Pi(\mathcal{E})} \left[V_{\mathcal{C}}\left(S_{C_\pi(i)}^{i \leftarrow A}\right) - V_{\mathcal{C}}\left(S_{C_\pi(i)}^{i \leftarrow \emptyset}\right) \right]. \tag{4}$$

The same outer-context set $C_\pi(i)$ appears on both sides, isolating the local configuration change of expert i.

Theorem 1 (Harsanyi expansion and endpoint equivalence). *For every expert $i \in \mathcal{E}$ and local configuration $A \subseteq \mathcal{J}$, the interaction expansion in Eq. 3 and the endpoint evaluation in Eq. 4 are identical, $F_i(A) = \widetilde{F}_i(A)$.*

Proof sketch. Fix i, A, and an outer permutation π. For any non-empty $B \subseteq \mathcal{J}$, let $\mathcal{H}_\pi(B) = \sum_{U \subseteq B} (-1)^{|B|-|U|} v_\pi(U)$ denote the Harsanyi dividend of the local game v_π. The key step is the

following Möbius cancellation:

$$\begin{aligned}\sum_{\emptyset\neq B\subseteq A}\mathcal{H}_\pi(B) &= \sum_{\emptyset\neq B\subseteq A}\sum_{U\subseteq B}(-1)^{|B|-|U|}v_\pi(U)\\ &= \sum_{U\subseteq A}v_\pi(U)\sum_{\substack{B:\,U\subseteq B\subseteq A\\ B\neq\emptyset}}(-1)^{|B|-|U|} = v_\pi(A)-v_\pi(\emptyset).\end{aligned} \tag{5}$$

The inner coefficient vanishes for every intermediate subset U, while it remains $+1$ for $U = A$ and -1 for $U = \emptyset$. For $\mathcal{J} = \{u, g, d\}$, the left-hand side is exactly the first-, second-, and third-order interaction expansion in Eq. 3. Averaging over $\pi \in \Pi(\mathcal{E})$ gives $F_i(A) = \widetilde{F}_i(A)$. The complete derivation is given in Appendix H.

We therefore use Eq. 4 as the operational form of the configuration-valued utility: the endpoint form of the full Harsanyi interaction expansion.

3.3 Sampling and Optimization

Sampling. We estimate the outer-context expectation in Eq. 4 using a finite set of sampled expert orderings, avoiding explicit enumeration of the permutation space. Given uniformly sampled permutations $\pi^{(1)}, \ldots, \pi^{(M)} \sim \mathrm{Unif}(\Pi(\mathcal{E}))$, the configuration utility is estimated as:

$$\widehat{F}_i^{\mathrm{MC}}(A) = \frac{1}{M}\sum_{m=1}^{M}\left[V_{\mathcal{C}_{\mathrm{est}}}\left(S^{i\leftarrow A}_{C_{\pi^{(m)}}(i)}\right) - V_{\mathcal{C}_{\mathrm{est}}}\left(S^{i\leftarrow \emptyset}_{C_{\pi^{(m)}}(i)}\right)\right]. \tag{6}$$

To use the finite sampling budget more effectively, we combine collapse-aware suffix pruning (CASP) with router-prior order sampling (RPOS): CASP removes collapsed suffix contexts from further evaluation, while RPOS prioritizes informative expert orderings.

Collapse-aware suffix pruning (CASP). Following Shapley-MoE (Huang et al., 2025d), define $R_t^{(m)}$ as the state in which the suffix $\{\pi^{(m)}(t), \ldots, \pi^{(m)}(|\mathcal{E}|)\}$ uses the high-bit anchor state and all earlier experts use the low-bit baseline. We process $t = 1, \ldots, |\mathcal{E}|$, starting from $R_1^{(m)} = S^{\mathrm{ref}}$ and progressively shrinking the high-bit coalition. Expert $i = \pi^{(m)}(t)$ is evaluated only if no state encountered before reaching position t has collapsed, i.e., on the event

$$\Omega_{\pi^{(m)}(t)} = \left\{\min_{s=1,\ldots,t} V_{\mathcal{C}_{\mathrm{est}}}\left(R_s^{(m)}\right) \geq \tau\, V_{\mathcal{C}_{\mathrm{est}}}\left(S^{\mathrm{ref}}\right)\right\}. \tag{7}$$

Once the current chain state falls below the retention threshold, CASP prunes the remaining suffix of the sampled order, and all subsequent marginals for that permutation are set to zero. This avoids evaluating progressively smaller contexts that have already entered a collapsed regime. The population target therefore changes from $F_i(A)$ in Eq. 4 to the suffix-pruned target $F_i^{\mathrm{prune}}(A) = \mathbb{E}_q[w(\pi)\Delta_i^\pi(A)\mathbf{1}_{\Omega_i}(\pi)]$, introducing the explicit suffix-pruning bias bounded in Theorem 2 and Appendix I.

Router-prior order sampling (RPOS). Uniform permutation sampling ignores router statistics, which indicate how strongly experts are weighted on the calibration distribution. We use this signal to construct a router-prior proposal over expert orders, while correcting the proposal mismatch by standard importance weighting. Let a_i be the empirical mean routing mass and $\epsilon_r > 0$ a small smoothing constant. The smoothed routing prior, Plackett–Luce proposal, and importance weight are defined as:

$$\begin{aligned} a_i &= \frac{1}{|\mathcal{C}_{\mathrm{est}}|}\sum_{x\in\mathcal{C}_{\mathrm{est}}}\mathbb{E}[G_i(x)], & p_i &= \frac{a_i+\epsilon_r}{\sum_{j\in\mathcal{E}}(a_j+\epsilon_r)},\\ q(\pi) &= \prod_{t=1}^{|\mathcal{E}|}\frac{p_{\pi(t)}}{\sum_{s\geq t}p_{\pi(s)}}, & w(\pi) &= \frac{1}{|\mathcal{E}|!\,q(\pi)}. \end{aligned} \tag{8}$$

where smoothing gives every expert positive proposal mass, $q(\pi)$ is the router-prior order proposal, and $w(\pi)$ converts the sampled average back to the uniform permutation measure. Thus, before suffix pruning, RPOS preserves the target in Eq. 4 through importance correction; when combined

with CASP, it preserves the suffix-pruned population target $F_i^{\mathrm{prune}}(A)$. The importance-correction identity and the variance motivation of the router-prior order proposal are provided in Appendix J.

Combining the endpoint marginal in Eq. 6, the CASP pruning event in Eq. 7, and the proposal correction in Eq. 8 gives the final Monte Carlo approximation; thus we have:

$$\widehat{F}_i(A) = \frac{1}{M} \sum_{m=1}^{M} w(\pi^{(m)})\, \mathbf{1}_{\Omega_i}(\pi^{(m)}) \left[V_{\mathcal{C}_{\mathrm{est}}}\left(S^{i\leftarrow A}_{C_{\pi^{(m)}}(i)}\right) - V_{\mathcal{C}_{\mathrm{est}}}\left(S^{i\leftarrow \emptyset}_{C_{\pi^{(m)}}(i)}\right)\right]. \tag{9}$$

Define the aggregate CASP bias as $\varepsilon_{\mathrm{prune}} := \max_{i,A} |F_i(A) - F_i^{\mathrm{prune}}(A)|$.

Theorem 2 (Uniform recovery). *Suppose each omitted configuration marginal satisfies $|\Delta_i^{\pi}(A)| \le \varepsilon_{\mathrm{omit}}$ whenever its suffix position is pruned by CASP, and the importance-weighted evaluated marginal satisfies $|w(\pi)\mathbf{1}_{\Omega_i}(\pi)\Delta_i^{\pi}(A)| \le G_{\max}$ uniformly over i, A, and π. Then $\varepsilon_{\mathrm{prune}} \le \varepsilon_{\mathrm{omit}} \max_i \Pr_{\pi}[\Omega_i^c]$, and for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\max_{i\in\mathcal{E}} \sup_{A\subseteq\mathcal{J}} \left|\widehat{F}_i(A) - F_i(A)\right| \le \varepsilon_{\mathrm{prune}} + G_{\max}\sqrt{\frac{2}{M}\log\left(\frac{2|\mathcal{E}|\,2^{|\mathcal{J}|}}{\delta}\right)}. \tag{10}$$

The suffix-pruning bias component is proved in Appendix I, and the uniform concentration argument is given in Appendix K.

Budgeted selection. After computing the utility table $\{\widehat{F}_i(A)\}$, the final mixed-precision assignment is a multiple-choice knapsack problem:

$$\max_{(A_i)_{i\in\mathcal{E}}} \sum_{i\in\mathcal{E}} \widehat{F}_i(A_i) \quad \text{s.t.} \quad \sum_{i\in\mathcal{E}} m_i(A_i) \le B_{\mathrm{mem}}, \qquad A_i \subseteq \mathcal{J}. \tag{11}$$

Each expert is one group, and each candidate configuration $A \subseteq \mathcal{J}$ is one item with value $\widehat{F}_i(A)$ and cost $m_i(A)$. We solve Eq. 11 exactly by dynamic programming:

$$D_i(b) = \max_{A\subseteq\mathcal{J}} \left\{ D_{i-1}\big(b - m_i(A)\big) + \widehat{F}_i(A) \right\}, \qquad D_0(b) = 0. \tag{12}$$

Backtracking yields $\widehat{S} = (\widehat{A}_i)_{i\in\mathcal{E}}$. We further reduce redundant computation by reusing precomputed quantization perturbations, performing in-place expert-state switching, caching prefix activations, and batching candidate local configurations. These implementation optimizations enable efficient configuration-utility estimation in practice, with detailed utility-construction times reported in Appendix B.

4 Experiments

4.1 Experimental Setup

Models, quantization, and implementation. We evaluate MolmoE-1B (Deitke et al., 2025), DeepSeek-VL2-L/S/T (Wu et al., 2024), and Mixtral-8×7B (Jiang et al., 2024), covering multimodal and language-only MoE settings. Following the PMQ stage of MC# (Huang et al., 2026), routing-expert weights use 4 or 2 bits; attention, shared-expert, vision-encoder, cross-modal-projector, and other non-router weights use 4 bits, while routers remain at 16 bits. The mixed-precision budgets are the model-dependent PMQ reference (2.57–2.59 bits) and 2.30 bits; uniform 3- and 2-bit quantization serve as fixed-precision references. All methods use GPTQ (Frantar et al., 2023). Calibration uses 128 VQA-style examples pooled from DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), ScienceQA (Lu et al., 2022), TextVQA (Singh et al., 2019), and ChartQA (Masry et al., 2022) for the multimodal models, and 128 random 2048-token C4 (Raffel et al., 2020) sequences for Mixtral, following PMQ. These samples are used only for configuration utilities. Unless stated otherwise, attribution methods share the calibration set, candidate local configurations, and budgeted MCKP solver, with $M = 20$ and CASP threshold $\tau = 0.5$. Appendix E specifies the full GPTQ setup, parameter-weighted bit accounting, and storage definition.

Evaluation and baselines. The four multimodal models are evaluated on MMBench (Liu et al., 2024b), MMStar (Chen et al., 2024), MME (Fu et al., 2025), MMMU (Yue et al., 2024), AI2D (Kembhavi et al., 2017), and OCRBench (Liu et al., 2024c). Avg.% averages all but the differently scaled MME, and Drop↓ is its relative decrease from the same model's FP16 reference; Section 4.7 instead uses BF16 execution. Config-Owen results are three-seed means, and Size denotes stored model size in decimal GB. At matched average-bit budgets, we compare Uniform (one bit-width for all expert projections), Hessian (second-order sensitivity), and PMQ (the static mixed-precision component of MC#).

We prioritize baselines that allow direct comparison under matched weight-only budgets. MxMoE (Duanmu et al., 2025) is not included in the main matched-budget table because its reported deployment-oriented results use a different quantization and kernel setup, including activation quantization and customized kernels. MoPEQ (Chitty-Venkata et al., 2025) uses a clustering-based precision allocation scheme, whose operating points are tied to the resulting cluster structure rather than directly specified by arbitrary average-bit targets, and it also uses a different quantization backend. For MC# (Huang et al., 2026), we compare against PMQ alone, since OTP is a separate dynamic expert-pruning module outside our weight-allocation setting. For Mixtral-8×7B, we additionally include RouterNorm+MaxVar (Chowdhury et al., 2026) and Auto Fine-Grained (Xie et al., 2025), following the model scope and evaluation settings of their original studies, where a directly corresponding VLM evaluation setup is not considered.

Table 1: Main comparison across four multimodal MoE models. Best quantized results at each matched budget are in **bold**.

Model	Method	Bits	Size	MMB	MMS	MMMU	AI2D	OCRB	MME	Avg.%↑	Drop↓
MolmoE-1B	FP16	16.00	14.4	60.76	43.13	35.33	73.45	55.00	1360.67	53.53	0.00%
	Uni	3.00	2.8	58.68	44.20	30.67	69.18	53.10	1329.38	51.17	4.42%
	Uni	2.00	2.0	37.00	37.13	28.56	53.92	43.50	1128.64	40.02	25.24%
	Hessian	2.60	2.5	49.38	41.23	26.83	60.74	47.90	1196.43	45.22	15.54%
	Hessian	2.30	2.2	47.63	40.69	25.48	58.16	46.20	1181.34	43.63	18.50%
	PMQ	2.60	2.5	51.89	41.07	29.89	66.88	48.60	1284.52	47.67	10.96%
	PMQ	2.30	2.2	49.63	**42.23**	27.46	**62.48**	**48.80**	1259.33	46.12	13.85%
	Ours	2.60	2.5	**53.33**	**42.33**	**30.22**	**68.59**	**50.00**	**1300.81**	**48.89**	**8.67%**
	Ours	2.30	2.2	**50.96**	40.97	**28.11**	62.39	48.60	**1261.09**	**46.21**	**13.69%**
DeepSeek-VL2-L	FP16	16.00	55.0	82.99	62.07	54.67	82.32	81.30	1644.93	72.67	0.00%
	Uni	3.00	10.7	82.64	60.47	48.67	82.90	90.80	1637.85	73.10	-0.59%
	Uni	2.00	7.6	60.91	49.00	40.22	68.36	74.50	1480.39	58.60	19.36%
	Hessian	2.57	9.4	80.71	60.00	45.38	78.25	74.61	1617.23	67.79	6.72%
	Hessian	2.30	8.5	77.04	56.14	46.16	77.60	73.30	1593.24	66.05	9.11%
	PMQ	2.57	9.4	**82.56**	60.00	47.33	82.29	80.80	1628.42	70.60	2.85%
	PMQ	2.30	8.5	**81.57**	58.74	47.85	**81.70**	81.40	1601.27	70.25	3.33%
	Ours	2.57	9.4	81.42	**60.12**	**48.56**	**82.87**	**81.90**	**1633.78**	**70.97**	**2.33%**
	Ours	2.30	8.5	80.57	**60.04**	**48.03**	80.98	**82.30**	**1609.52**	**70.38**	**3.15%**
DeepSeek-VL2-S	FP16	16.00	32.2	78.61	56.73	47.33	79.05	83.50	1654.30	69.04	0.00%
	Uni	3.00	6.3	78.69	54.00	42.00	77.42	81.80	1669.77	66.78	3.28%
	Uni	2.00	4.5	34.62	33.47	23.33	48.77	57.00	1247.61	39.44	42.88%
	Hessian	2.58	5.5	72.88	52.27	40.67	67.84	74.31	1613.41	61.59	10.79%
	Hessian	2.30	5.0	71.63	51.98	38.83	68.30	73.50	1601.33	60.85	11.87%
	PMQ	2.58	5.5	74.05	54.33	43.67	69.53	76.70	1674.45	63.66	7.80%
	PMQ	2.30	5.0	73.86	52.43	41.42	70.32	75.60	1653.41	62.73	9.15%
	Ours	2.58	5.5	**77.24**	**56.27**	**46.67**	**78.27**	**81.20**	**1697.07**	**67.93**	**1.61%**
	Ours	2.30	5.0	**78.17**	**56.20**	**43.67**	**78.11**	**80.90**	**1683.75**	**67.41**	**2.37%**
DeepSeek-VL2-T	FP16	16.00	6.7	72.08	49.47	39.89	74.77	80.30	1554.01	63.30	0.00%
	Uni	3.00	1.4	67.10	46.53	38.33	70.85	77.40	1560.91	60.04	5.15%
	Uni	2.00	1.1	1.12	1.00	0.00	4.93	0.00	448.38	1.41	97.77%
	Hessian	2.59	1.3	66.68	42.36	29.79	69.20	70.02	1566.27	55.61	12.15%
	Hessian	2.30	1.2	64.22	43.08	31.24	68.57	68.50	1548.66	55.12	12.92%
	PMQ	2.59	1.3	**70.10**	46.93	34.00	**72.15**	74.10	1564.56	59.46	6.08%
	PMQ	2.30	1.2	**68.42**	45.58	33.23	**72.54**	73.10	1557.46	58.57	7.47%
	Ours	2.59	1.3	67.41	**47.40**	**37.44**	71.76	**78.70**	**1572.60**	**60.54**	**4.36%**
	Ours	2.30	1.2	66.71	**46.60**	**36.56**	72.44	**78.50**	**1568.27**	**60.16**	**4.96%**

4.2 Main Results

Table 1 reports the full comparison across all four multimodal MoE models. Config-Owen achieves the best quantized Avg.% in all eight matched model–budget settings. The gains are largest on DeepSeek-VL2-L/S and become more pronounced at 2.30 bits, indicating that configuration-level allocation is most useful when the precision budget leaves less room for suboptimal local choices.

4.3 Interaction Evidence

For each expert, we compute context-averaged Harsanyi coefficients $H_i(R)$ over nonempty subsets R of its three projections and measure the higher-order share as $\rho_i = \sum_{|R|\geq 2} |H_i(R)| / \sum_{R\neq\emptyset} |H_i(R)|$. Figure 3 shows that 81.8–94.0% of experts across the four models have $\rho_i > 0.3$, indicating that non-additive projection effects are widespread rather than confined to a few experts. The order-wise breakdown appears in Appendix F.

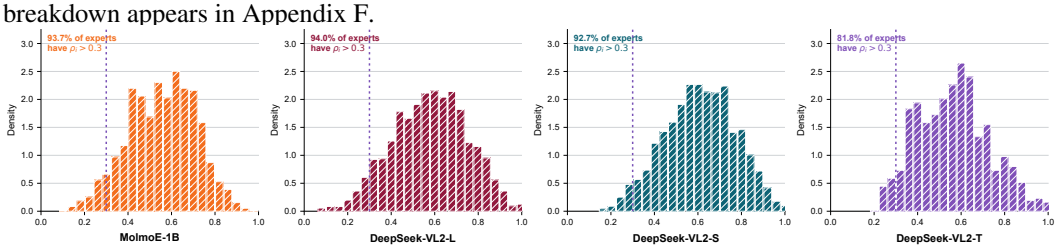
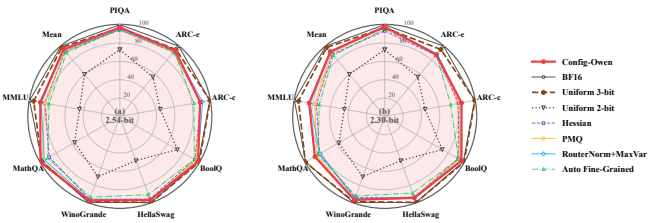


Figure 3: Interaction-energy ratio distributions across routing experts.

4.4 Task-Wise Results on Mixtral-8×7B

RouterNorm+MaxVar (Chowdhury et al., 2026) orders pretrained experts by ascending router ℓ_2 norm and promotes experts with unusually large maximum intra-neuron variance before assigning precision. Auto Fine-Grained (Xie et al., 2025) derives channel-wise bit allocations from Fisher-information and condition-number sensitivity.

Figure 4 compares five mixed-precision methods and three reference settings across the eight Mixtral tasks at the two matched budgets. The radar includes the eight tasks and an additional Mean axis. Scores on each axis are normalized by the best of all eight curves, so 100 denotes the strongest method on that task or on the mean. This relative view removes task-scale differences. Config-Owen attains the highest mean among the matched-budget methods at both budgets and remains competitive across the task axes, showing that its aggregate gain is not driven by a single benchmark. Its advantage over Hessian also widens at the tighter budget, consistent with joint configuration choices becoming more consequential under stronger compression. The original scores are reported in Appendix C.



(a) 2.54-bit (b) 2.30-bit

Figure 4: Relative task performance on Mixtral-8×7B.

4.5 Bit-Budget Ablation

Figure 5 shows that Config-Owen's advantage on DeepSeek-VL2-S persists across the 2.1–3.0 average-bit range rather than only at the two tabulated budgets. It remains the strongest quantized method throughout the sweep and approaches FP16 at the upper end. Under tight budgets, the solver must choose a joint local bit-width configuration, which a single expert- or projection-level importance score cannot fully represent. The failure of uniform 2-bit quantization under aggressive compression further illustrates the need for structured precision assignment.

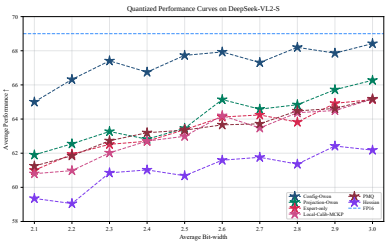


Figure 5: DeepSeek-VL2-S performance across average-bit budgets.

4.6 Ablation Studies

The four attribution variants use the same calibration data, candidate local configurations, and MCKP solver. Expert-only models the game only at the expert level and assigns one marginal-contribution score per expert, omitting the projection hierarchy. Projection-Owen retains the two-level structure and expert-level context but allocates from additive projection scores rather than a configuration-valued utility, thereby omitting higher-order projection terms. Local-Calib-MCKP directly evaluates joint local configurations under one fixed outer-expert context, while Config-Owen averages the same configuration utilities over expert-ordering contexts. Figure 6 visualizes these variants at the reference budgets: the five tasks are ordered by Config-Owen's score in each panel, with Mean as a sixth dimension. Full values are reported in Appendix D.

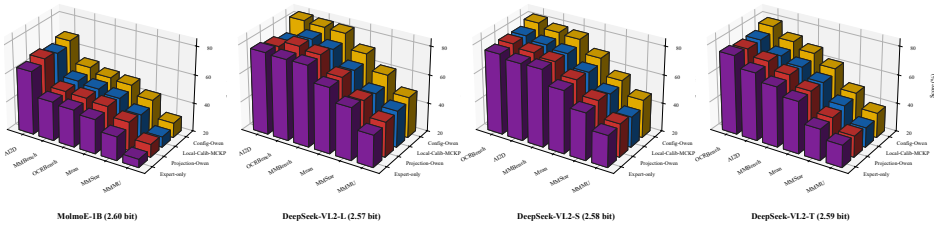


Figure 6: Task-wise attribution-granularity ablation at the reference mixed-precision budgets.

Config-Owen yields the highest five-task mean in all eight settings, showing that both within-expert configuration structure and outer-context averaging contribute beyond expert- or projection-level scalar attribution. The larger separation from Projection-Owen under tighter compression further indicates that higher-order local structure matters more as the budget tightens.

Table 2 complements this analysis at the sampling level: Uniform MC removes both CASP and RPOS, the two intermediate variants retain only one component, and Full retains both. CASP reduces utility-construction time, RPOS improves allocation under the fixed sample budget, and their combination gives the best efficiency–performance trade-off.

Table 2: CASP/RPOS component ablation.

Variant	CASP	RPOS	Utility Time↓	Avg.%↑
Uniform MC	×	×	63 min	65.86
+ CASP	✓	×	47 min	66.92
+ RPOS	×	✓	64 min	67.18
Full	✓	✓	**43 min**	**67.93**

Table 3: Deployment efficiency.

Model	GPU	Peak mem.	Tok./s
Mixtral-8×7B (BF16)	2×A800	90.8 GiB	11
Mixtral-8×7B (BF16)	1×3090	**OOM**	–
Ours 2.54-bit	1×3090	**18.1 GiB**	**18**
DeepSeek-VL2-L (BF16)	1×A800	54.3 GiB	15
DeepSeek-VL2-L (BF16)	1×3090	**OOM**	–
Ours 2.57-bit	1×3090	**10.9 GiB**	**23**

4.7 Deployment Efficiency

We evaluate peak GPU memory and decoding throughput at batch size 1 with a 2,048-token input and 512 generated tokens; peak memory sums the peak allocations across the listed GPUs. Table 3 shows that both mixed-precision models fit on one RTX 3090, whereas their BF16 counterparts do not. The reported throughputs are end-to-end rates for the listed configurations; compact low-bit weights reduce memory footprint and weight-transfer demand during autoregressive decoding.

5 Conclusion

This paper presents Config-Owen, a hierarchical game-theoretic framework for mixed-precision quantization of MoE models. By combining cooperative expert-level valuation, Owen-style hierarchical attribution, and within-expert Harsanyi interactions, Config-Owen derives configuration-valued utilities that better reflect structured precision decisions. Experiments across multimodal and language MoE models demonstrate consistent improvements under low-bit budgets; on DeepSeek-VL2-S, Config-Owen achieves average scores of 67.93 at 2.58 bits and 67.41 at 2.30 bits, compared with 69.04 in FP16. These results highlight the importance of modeling expert context, hierarchical structure, and within-expert interactions in MoE precision allocation. Future work will explore task-aware calibration, cross-distribution utility transfer, and joint modeling of routing and precision allocation.

AI USE STATEMENT

The authors utilized OpenAI's ChatGPT to refine the language and enhance the readability of this paper. It is important to state that ChatGPT was not used for generating ideas or the conceptual framework.

REPRODUCIBILITY STATEMENT

We provide detailed quantization settings, calibration protocols, bit-accounting rules, hyperparameters, utility-construction costs, complete ablations, and theoretical derivations in the appendix. All main Config-Owen results are evaluated over three random seeds where specified. Code will be released upon acceptance.

REFERENCES

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087, 2024. doi: 10.52202/079017-0850.

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*, pp. 5350–5359, October 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Chen_Towards_Mixed-Precision_Quantization_of_Neural_Networks_via_Constrained_Optimization_ICCV_2021_paper.html.

Zhixuan Chen, Xing Hu, Dawei Yang, Zukang Xu, Xu Chen, Zhihang Yuan, Sifan Zhou, and Jiangyong Yu. MoEQuant: Enhancing quantization for mixture-of-experts large language models via expert-balanced sampling and affinity guidance. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 8245–8260, 2025.

Krishna Teja Chitty-Venkata, Jie Ye, and Murali Emani. MoPEQ: Mixture of mixed precision quantized experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 4082–4091, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025W/BiVision/html/Chitty-Venkata_MoPEQ_Mixture_of_Mixed_Precision_Quantized_Experts_ICCVW_2025_paper.html.

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 *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=yiMlVBAoQi.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 91–104, 2025.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, pp. 18089–18102, 2022.

Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. SpQR: A sparse-quantized representation for near-lossless LLM weight compression. In *International Conference on Learning Representations*, 2024.

Xin Ding, Xiaoyu Liu, Zhijun Tu, Yun Zhang, Wei Li, Jie Hu, Hanting Chen, Yehui Tang, Zhiwei Xiong, Baoqun Yin, and Yunhe Wang. CBQ: Cross-block quantization for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/15212bd2265c4a3ab0dbc1b1982c1b69-Abstract-Conference.html.

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*, pp. 18518–18529, 2020.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201, 2024.

Haojie Duanmu, Xiuhong Li, Zhihang Yuan, Size Zheng, Jiangfei Duan, Xingcheng Zhang, and Dahua Lin. MxMoE: Mixed-precision quantization for MoE with accuracy and performance co-design. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 14793–14806, 2025.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Elias Frantar and Dan Alistarh. QMoE: Sub-1-bit compression of trillion parameter models. In *Proceedings of Machine Learning and Systems*, volume 6, pp. 439–451, 2024.

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.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 162549–162567, 2025. doi: 10.52202/085713-4899.

Amirata Ghorbani and James Zou. Data Shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, pp. 2242–2251, 2019.

Xing Hu, Yuan Cheng, Zhixuan Chen, Dawei Yang, Jiangyong Yu, Xu Chen, Zhihang Yuan, Zhe Jiang, and Sifan Zhou. OSTQuant: Refining large language model quantization with orthogonal and scaling transformations for better distribution fitting. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5cebc89b113920dbff7c79854ba765a3-Abstract-Conference.html.

Beichen Huang, Yueming Yuan, Zelei Shao, and Minjia Zhang. MiLo: Efficient quantized MoE inference with mixture of low-rank compensators. In *Proceedings of Machine Learning and Systems*, volume 7, 2025a.

Wei Huang, Yue Liao, Jianhui Liu, Ruifei He, Haoru Tan, Shiming Zhang, Hongsheng Li, Si Liu, and Xiaojuan Qi. Mixture compressor for mixture-of-experts LLMs gains more. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=hheFYjOsWO.

Wei Huang, Haotong Qin, Yangdong Liu, Yawei Li, Qinshuo Liu, Xianglong Liu, Luca Benini, Michele Magno, Shiming Zhang, and Xiaojuan Qi. SliM-LLM: Salience-driven mixed-precision quantization for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 25672–25692, 2025c.

Wei Huang, Yue Liao, Yukang Chen, Jianhui Liu, Haoru Tan, Si Liu, Shiming Zhang, Shuicheng Yan, and Xiaojuan Qi. MC#: Mixture compressor for mixture-of-experts large models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(7):7607–7621, 2026. doi: 10.1109/TPAMI.2026.3664873.

Weizhong Huang, Yuxin Zhang, Xiawu Zheng, Fei Chao, Rongrong Ji, and Liujuan Cao. Discovering important experts for mixture-of-experts models pruning through a theoretical perspective. In *Advances in Neural Information Processing Systems*, volume 38, 2025d. doi: 10.52202/085713-4538.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. URL https://arxiv.org/abs/2401.04088.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Conference on Computer Vision and Pattern Recognition*, pp. 5376–5384, 2017.

Sehoon Kim, Coleman Richard Charles Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W. Mahoney, and Kurt Keutzer. SqueezeLLM: Dense-and-sparse quantization. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 23901–23923, 2024.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021.

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.

Jing Liu, Ruihao Gong, Xiuying Wei, Zhiwei Dong, Jianfei Cai, and Bohan Zhuang. QLLM: Accurate and efficient low-bitwidth quantization for large language models. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/437bc4ccafd3fc6d4289bd10940be42b-Abstract-Conference.html.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, 2024b.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: On the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024c. doi: 10.1007/s11432-024-4235-6.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. SpinQuant: LLM quantization with learned rotations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e5b1c0d4866f72393c522c8a00eed4eb-Abstract-Conference.html.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2507–2521, 2022. doi: 10.52202/068431-0182. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html.

Yuexiao Ma, Huixia Li, Xiawu Zheng, Feng Ling, Xuefeng Xiao, Rui Wang, Shilei Wen, Fei Chao, and Rongrong Ji. AffineQuant: Affine transformation quantization for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/ddfb58b61e7213dfb9e06c695475b2bd-Abstract-Conference.html.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2200–2209, January 2021. URL https://openaccess.thecvf.com/content/WACV2021/html/Mathew_DocVQA_A_Dataset_for_VQA_on_Document_Images_WACV_2021_paper.html.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. InfographicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706, 2022.

Guillermo Owen. Values of games with a priori unions. In *Mathematical Economics and Game Theory*, pp. 76–88. Springer, 1977.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, and Yuxiong He. DeepSpeed-MoE: Advancing mixture-of-experts inference and training to power next-generation AI scale. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 18332–18346, 2022.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8Wuvhh0LYW.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019.

Zunhai Su, Qingyuan Li, Hao Zhang, Weihao Ye, Qibo Xue, Yulei Qian, Ngai Wong, and Kehong Yuan. Unveiling super experts in mixture-of-experts large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c01b90b643d0e877cff0f34cbe27410f-Abstract-Conference.html.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302, 2024.

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*, pp. 38087–38099, 2023.

Zhanhao Xie, Yuexiao Ma, Xiawu Zheng, Fei Chao, Wanchen Sui, Yong Li, Shen Li, and Rongrong Ji. Automated fine-grained mixture-of-experts quantization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 27024–27037, 2025. doi: 10.18653/v1/2025.findings-acl.1386.

Zukang Xu, Zhixiong Zhao, Xing Hu, Zhixuan Chen, and Dawei Yang. KBVQ-MoE: KLT-guided SVD with bias-corrected vector quantization for MoE large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/5f868c3e0050f13ec82d3694df531de1-Abstract-Conference.html.

Zhewei Yao, Zhen Dong, Zhangcheng Zheng, Amir Gholami, Jiali Yu, Eric Tan, Leyuan Wang, Qijing Huang, Yida Wang, Michael W. Mahoney, and Kurt Keutzer. HAWQ-V3: Dyadic neural network quantization. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 11875–11886, 2021.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. ZeroQuant: Efficient and affordable post-training quantization for large-scale transformers. In *Advances in Neural Information Processing Systems*, pp. 27168–27183, 2022.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Cheng Zhang, Jeffrey T. H. Wong, Can Xiao, George A. Constantinides, and Yiren Zhao. QERA: An analytical framework for quantization error reconstruction. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/21718991f6acf19a42376b5c7a8668c5-Abstract-Conference.html.

Tianyi Zhang and Anshumali Shrivastava. LeanQuant: Accurate and scalable large language model quantization with loss-error-aware grid. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/57ccc284de6f060c8dcde8f9352f70a5-Abstract-Conference.html.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. ST-MoE: Designing stable and transferable sparse expert models. arXiv preprint arXiv:2202.08906, 2022. URL https://arxiv.org/abs/2202.08906.

APPENDIX

A STABILITY ACROSS RANDOM SEEDS

To quantify the variability of the quantization procedure, we repeat Config-Owen on DeepSeek-VL2-S with three random seeds. The random seed affects the sampled calibration examples and the Monte Carlo expert permutations, while the benchmark evaluation protocol is kept fixed. Table 4 compares these runs with the corresponding published PMQ results (Huang et al., 2026).

Table 4: Stability across random seeds on DeepSeek-VL2-S. Config-Owen is reported as mean ± standard deviation over three seeds; PMQ values are the published reference results (Huang et al., 2026). Size is stored model size in decimal GB.

Method	Bits	Size	Avg.%↑	Drop↓
PMQ	2.58	5.5	63.66	7.80%
Ours	2.58	5.5	$\mathbf{67.93 \pm 0.28}$	$\mathbf{1.61 \pm 0.41}\%$
PMQ	2.30	5.0	62.73	9.15%
Ours	2.30	5.0	$\mathbf{67.41 \pm 0.31}$	$\mathbf{2.37 \pm 0.45}\%$

B Utility Computation Efficiency

Table 5 reports the end-to-end offline time for constructing the configuration-utility table on NVIDIA A800 80GB GPUs. The measurement includes routing-mass profiling, expert-order sampling, endpoint marginal evaluation, and utility aggregation, while excluding downstream benchmark evaluation. Across the evaluated multimodal MoE models, the complete utility-construction stage finishes within 22–91 minutes, demonstrating efficient offline estimation at the evaluated model scales.

Table 5: End-to-end configuration-utility construction time on NVIDIA A800 80GB GPUs.

Model	MolmoE-1B	DeepSeek-VL2-T	DeepSeek-VL2-S	DeepSeek-VL2-L
Utility construction time (min)	26	22	43	91

C Extended Comparison on Mixtral-8×7B

Table 6 reports an extended comparison on Mixtral-8×7B, including the BF16 reference, uniform quantization, and two average-bit mixed-precision budgets. The mean is the arithmetic average over PIQA, ARC-e, ARC-c, BoolQ, HellaSwag, WinoGrande, MathQA, and MMLU.

Table 6: Extended comparison on Mixtral-8×7B. Mean is computed over all eight benchmarks. Size is stored model size in decimal GB. Bold indicates the best result within each average-bit budget.

Method	Bits	Size	PIQA	ARCe	ARCc	Bool	HSwg	Wino	MthQ	MMLU	Mean↑
BF16 reference	16.00	93.4	85.20	84.01	57.17	85.35	81.48	75.93	39.29	67.88	72.04
Uniform	3.00	17.7	82.32	80.05	57.42	86.09	81.51	75.14	39.43	64.84	70.85
Uniform	2.00	12.1	61.98	47.20	25.71	62.39	41.91	53.22	22.79	30.36	43.20
Hessian	2.54	15.1	80.34	76.71	51.46	81.38	78.22	73.06	35.62	56.48	66.66
PMQ	2.54	15.1	80.52	77.10	51.28	82.54	79.03	73.95	39.18	56.37	67.50
RouterNorm+MaxVar	2.54	15.1	80.59	78.17	**52.63**	84.51	78.75	73.52	38.11	59.22	68.19
Auto Fine-Grained	2.54	15.1	81.88	77.45	47.10	80.82	74.07	71.27	36.38	53.26	65.28
Config-Owen	2.54	15.1	**82.67**	**78.92**	51.13	**84.83**	**79.26**	**74.22**	**40.02**	**59.87**	**68.87**
Hessian	2.30	13.8	79.48	73.06	47.18	79.62	76.74	71.53	32.41	51.37	63.92
PMQ	2.30	13.8	83.11	73.59	47.78	80.83	76.48	73.14	33.84	52.54	65.16
RouterNorm+MaxVar	2.30	13.8	83.16	**74.87**	48.72	82.56	76.44	72.65	32.11	56.35	65.86
Auto Fine-Grained	2.30	13.8	81.28	73.66	41.98	79.93	72.59	70.17	33.03	49.28	62.74
Config-Owen	2.30	13.8	**83.33**	73.82	**48.91**	**83.34**	**77.21**	**73.74**	**34.69**	**56.87**	**66.49**

D Full Attribution-Granularity Ablation Across Benchmarks

Table 7: Full attribution-granularity ablation across benchmarks. Avg.% is computed over MMBench, MMStar, MMMU, AI2D, and OCRBench. MME is reported separately because its score scale differs from the other benchmarks. Size is stored model size in decimal GB. Drop↓ is the relative decrease in Avg.% from the FP16 reference of the same model.

Model	Method	Bits	Size	MMB	MMS	MMMU	AI2D	OCRB	MME	Avg.%↑	Drop↓
MolmoE-1B	FP16	16.00	14.4	60.76	43.13	35.33	73.45	55.00	1360.67	53.53	0.00%
	Expert-only	2.60	2.5	47.49	36.66	25.76	64.53	46.30	1229.43	44.15	17.53%
	Projection-Owen	2.60	2.5	48.58	40.12	28.97	67.46	48.10	1277.89	46.65	12.87%
	Local-Calib-MCKP	2.60	2.5	49.72	38.54	27.26	65.48	48.20	1276.74	45.84	14.37%
	Config-Owen	2.60	2.5	**53.33**	**42.33**	**30.22**	**68.59**	**50.00**	**1300.81**	**48.89**	**8.67%**
	Expert-only	2.30	2.2	45.20	35.54	22.01	**63.08**	44.90	1240.03	42.15	21.27%
	Projection-Owen	2.30	2.2	46.06	36.18	21.64	60.08	45.80	1222.67	41.95	21.63%
	Local-Calib-MCKP	2.30	2.2	46.18	34.41	25.92	61.76	44.50	1247.58	42.55	20.51%
	Config-Owen	2.30	2.2	**50.96**	**40.97**	**28.11**	62.39	**48.60**	**1261.09**	**46.21**	**13.69%**
DeepSeek-VL2-L	FP16	16.00	55.0	82.99	62.07	54.67	82.32	81.30	1644.93	72.67	0.00%
	Expert-only	2.57	9.4	77.02	56.84	42.57	77.64	76.90	1599.45	66.19	8.91%
	Projection-Owen	2.57	9.4	78.19	55.18	44.64	76.04	80.10	1608.04	66.83	8.04%
	Local-Calib-MCKP	2.57	9.4	76.05	53.46	45.63	73.22	79.20	1556.92	65.51	9.85%
	Config-Owen	2.57	9.4	**81.42**	**60.12**	**48.56**	**82.87**	**81.90**	**1633.78**	**70.97**	**2.33%**
	Expert-only	2.30	8.5	74.64	53.98	42.79	76.65	75.20	1573.62	64.65	11.03%
	Projection-Owen	2.30	8.5	77.89	56.22	43.26	78.64	77.20	1583.48	66.64	8.30%
	Local-Calib-MCKP	2.30	8.5	75.88	54.31	43.11	74.06	76.30	**1610.47**	64.73	10.92%
	Config-Owen	2.30	8.5	**80.57**	**60.04**	**48.03**	**80.98**	**82.30**	1609.52	**70.38**	**3.15%**
DeepSeek-VL2-S	FP16	16.00	32.2	78.61	56.73	47.33	79.05	83.50	1654.30	69.04	0.00%
	Expert-only	2.58	5.5	74.26	53.27	42.38	74.09	76.60	1593.55	64.12	7.13%
	Projection-Owen	2.58	5.5	73.56	55.17	43.43	75.62	77.90	1617.58	65.14	5.66%
	Local-Calib-MCKP	2.58	5.5	73.14	53.28	41.46	75.31	77.80	1637.22	64.20	7.02%
	Config-Owen	2.58	5.5	**77.24**	**56.27**	**46.67**	**78.27**	**81.20**	**1697.07**	**67.93**	**1.61%**
	Expert-only	2.30	5.0	72.91	50.21	39.28	75.59	74.60	1600.56	62.52	9.45%
	Projection-Owen	2.30	5.0	74.71	51.38	38.97	73.38	77.90	1594.48	63.27	8.37%
	Local-Calib-MCKP	2.30	5.0	72.37	49.47	40.19	70.46	77.60	1612.36	62.02	10.18%
	Config-Owen	2.30	5.0	**78.17**	**56.20**	**43.67**	**78.11**	**80.90**	**1683.75**	**67.41**	**2.37%**
DeepSeek-VL2-T	FP16	16.00	6.7	72.08	49.47	39.89	74.77	80.30	1554.01	63.30	0.00%
	Expert-only	2.59	1.3	62.38	42.12	35.59	67.78	75.60	**1596.76**	56.69	10.44%
	Projection-Owen	2.59	1.3	64.44	43.26	34.92	69.78	74.40	1538.87	57.36	9.39%
	Local-Calib-MCKP	2.59	1.3	63.71	44.92	33.81	67.55	74.90	1557.13	56.98	9.99%
	Config-Owen	2.59	1.3	**67.41**	**47.40**	**37.44**	**71.76**	**78.70**	1572.60	**60.54**	**4.36%**
	Expert-only	2.30	1.2	61.22	43.38	32.34	68.62	75.40	1537.65	56.19	11.23%
	Projection-Owen	2.30	1.2	63.38	45.67	34.57	67.68	75.60	**1571.56**	57.38	9.36%
	Local-Calib-MCKP	2.30	1.2	62.04	45.11	33.28	65.73	74.90	1549.88	56.21	11.20%
	Config-Owen	2.30	1.2	**66.71**	**46.60**	**36.56**	**72.44**	**78.50**	1568.27	**60.16**	**4.96%**

E Quantization Configuration and Bit Accounting

We use asymmetric group-wise GPTQ with group and internal block sizes of 128, per-output-channel scales and zero points, Hessian damping of 1% of the mean diagonal value, and no activation ordering. Transformer blocks are processed sequentially using 128 calibration sequences of at most 2048 tokens.

Mixed-precision allocation applies to the up, gate, and down projections of routing experts, with $b_{ij} \in \{2, 4\}$ denoting the bit-width of projection j in expert i. Attention, shared experts, and other non-router language-backbone weights are fixed at 4 bits, while router weights remain at 16 bits. Let P_{ij} be the parameter count of routing-expert projection (i, j), $P_{\text{fix}}^{\text{LM}}$ the number of fixed 4-bit language-backbone parameters, and P_{router} the router parameter count. We first define the

language-backbone bit payload and parameter count as

$$B_{\mathrm{LM}}(S) = \sum_{i\in\mathcal{E}} \sum_{j\in\mathcal{J}} P_{ij} b_{ij} + 4P_{\mathrm{fix}}^{\mathrm{LM}} + 16P_{\mathrm{router}}, \tag{E.1}$$

$$P_{\mathrm{LM}} = \sum_{i\in\mathcal{E}} \sum_{j\in\mathcal{J}} P_{ij} + P_{\mathrm{fix}}^{\mathrm{LM}} + P_{\mathrm{router}}. \tag{E.2}$$

The Bits column reports the parameter-weighted average over this language backbone,

$$\bar{b}_{\mathrm{LM}}(S) = \frac{B_{\mathrm{LM}}(S)}{P_{\mathrm{LM}}}. \tag{E.3}$$

The Size column instead reports the complete stored model. With P_{vis} denoting the 4-bit vision-encoder and cross-modal-projector parameters and $M_{\mathrm{meta}}(S)$ the stored scale and zero-point metadata in bits, its accounting is

$$\mathrm{Size}(S) = \frac{B_{\mathrm{LM}}(S) + 4P_{\mathrm{vis}} + M_{\mathrm{meta}}(S)}{8 \times 10^9} \quad \mathrm{GB}. \tag{E.4}$$

The table reports Size to one decimal place. FP16 reference sizes use 16 bits per model parameter and contain no quantization metadata.

F Additional Interaction Evidence

We aggregate absolute Harsanyi dividends over all routing experts and normalize them by interaction order to obtain the model-level shares of first-, second-, and third-order energy in Figure 7.

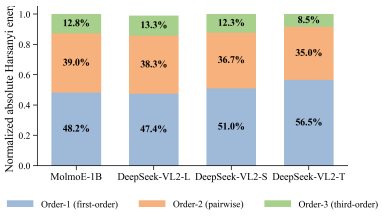


Figure 7: Harsanyi interaction energy by order.

First-order terms remain dominant at 47.4%–56.5%, but pairwise terms still contribute 35.0%–39.0%, and third-order terms add another 8.5%–13.3%. In total, non-additive effects explain 43.5%–51.8% of the absolute energy across models. Higher-order coupling therefore remains substantial after aggregation over experts and can alter allocation decisions under a tight bit budget.

Taken together, Figure 3 and Figure 7 show that interaction terms form a systematic part of the local utility landscape. This observation motivates configuration-level attribution, which evaluates a joint local precision state rather than three isolated projection upgrades.

G Detailed Algorithm

The complete procedure is summarized below. The algorithm follows the permutation-based sampling pipeline used in Shapley-style MoE compression, but replaces expert keep/remove marginals with configuration-valued precision marginals over the local state space $\mathcal{B}^{|\mathcal{J}|}$. In the main two-state setting, the local configuration space is fixed to $K = |\mathcal{B}|^{|\mathcal{J}|} = 8$. With this compact local state space, utility construction scales linearly with the number of sampled expert orders and routing experts.

Algorithm 1 Configuration-Valued Mixed-Precision Quantization for MoE Models

Require: MoE model $\mathcal{M}$ with routing expert set $\mathcal{E}$; projection set $\mathcal{J} = \{u, g, d\}$; calibration set $\mathcal{C}_{\text{est}}$; precision states $\mathcal{B} = \{b_{\text{low}}, b_{\text{high}}\}$; Monte Carlo samples M_s; CASP retention threshold τ; RPOS router-prior smoothing $\epsilon_r > 0$; memory budget B_{mem}.

Ensure: Quantized mixed-precision configuration $\widehat{S}$.

Compute the high-bit-anchor value $V_{\text{ref}} \leftarrow V_{\mathcal{C}_{\text{est}}}(S^{\text{ref}})$.
Profile mean routing masses $\{a_i\}_{i\in\mathcal{E}}$ and set $p_i \leftarrow (a_i + \epsilon_r)/\sum_j (a_j + \epsilon_r)$. ▷ Construct the RPOS router prior
for all $i \in \mathcal{E}$ **do**
  for all $A \in \mathcal{B}^{|\mathcal{J}|}$ **do**
    Initialize $\widehat{F}_i(A) \leftarrow 0$.
  end for
end for
for $m = 1, \ldots, M_s$ **do**
  Sample a permutation $\pi^{(m)}$ from the RPOS Plackett–Luce order proposal.
  Compute the RPOS importance-correction weight $w^{(m)} \leftarrow 1/\big(|\mathcal{E}|!\, q(\pi^{(m)})\big)$.
  for $t = 1, \ldots, |\mathcal{E}|$ **do**
    Set $i \leftarrow \pi^{(m)}(t)$.
    Set $C_{\pi^{(m)}}(i) \leftarrow \{\pi^{(m)}(t+1), \ldots, \pi^{(m)}(|\mathcal{E}|)\}$.
    Construct $R_t^{(m)}$ with $\{\pi^{(m)}(t), \ldots, \pi^{(m)}(|\mathcal{E}|)\}$ in the high-bit anchor state.
    if $V_{\mathcal{C}_{\text{est}}}(R_t^{(m)}) < \tau V_{\text{ref}}$ **then**
      break ▷ CASP: prune the remaining collapsed suffix
    end if
    for all $A \in \mathcal{B}^{|\mathcal{J}|}$ **do**
      Evaluate $\Delta_i^{\pi^{(m)}}(A)$ by Eq. (G.1).
      Update
$$\widehat{F}_i(A) \leftarrow \widehat{F}_i(A) + M_s^{-1} w^{(m)} \Delta_i^{\pi^{(m)}}(A).$$
    end for
  end for
end for
Solve the multiple-choice knapsack problem in Eq. (G.3).
Recover $\widehat{S} = (\widehat{A}_i)_{i\in\mathcal{E}}$ by dynamic-programming backtracking.
Apply $\widehat{A}_i$ to the up, gate, and down projections of each expert i.
return the quantized MoE model.

The configuration marginal, utility accumulation, and final budgeted selection are defined as

$$\Delta_i^{\pi^{(m)}}(A) \leftarrow V_{\mathcal{C}_{\text{est}}}(S^{(i\leftarrow A)}_{C_{\pi^{(m)}}(i)}) - V_{\mathcal{C}_{\text{est}}}(S^{(i\leftarrow A^0)}_{C_{\pi^{(m)}}(i)}). \tag{G.1}$$

$$\widehat{F}_i(A) \leftarrow \widehat{F}_i(A) + \frac{1}{M_s} w^{(m)} \Delta_i^{\pi^{(m)}}(A). \tag{G.2}$$

$$\max \sum_{i\in\mathcal{E}} \widehat{F}_i(A_i) \quad \text{s.t.} \quad \sum_{i\in\mathcal{E}} m_i(A_i) \le B_{\text{mem}}, \quad A_i \in \mathcal{B}^{|\mathcal{J}|}. \tag{G.3}$$

The dynamic-programming recursion used in the final MCKP step is

$$D_i(b) = \max_{A\in\mathcal{B}^{|\mathcal{J}|}} \Big\{ D_{i-1}(b - m_i(A)) + \widehat{F}_i(A) \Big\}, \qquad D_0(b) = 0. \tag{G.4}$$

The algorithm computes the full configuration-valued utility table $\big\{\widehat{F}_i(A) : i \in \mathcal{E},\ A \in \mathcal{B}^{|\mathcal{J}|}\big\}$, rather than a scalar importance score per expert.

H Proof of Theorem 1: Möbius Endpoint Equivalence

We prove Theorem 1 by first establishing the result under a fixed outer expert context and then averaging over outer permutations. The proof proceeds by observing that the precision states of the projections define a complete local cooperative game over $\mathcal{P}(\mathcal{J})$, whose Harsanyi dividends form the Möbius coordinates of the corresponding set function.

H.1 Local Precision Game and Möbius Coordinates

Fix expert i, local configuration A, and outer permutation $\pi \in \Pi(\mathcal{E})$. Let $C_\pi(i)$ denote the experts following i. For each projection subset $R \subseteq \mathcal{J}$, define the local state A_R and the induced characteristic function jointly as

$$A_R^j = \begin{cases} b_{\text{high}}, & j \in R, \\ b_{\text{low}}, & j \notin R, \end{cases} \qquad V_i^\pi(R) = V_{\mathcal{C}}\left(S_{C_\pi(i)}^{(i \leftarrow A_R)}\right). \tag{H.1}$$

The vector

$$\mathbf{v}_i^\pi = \left(V_i^\pi(R)\right)_{R \subseteq \mathcal{J}} \in \mathbb{R}^{2^{|\mathcal{J}|}} \tag{H.2}$$

contains all within-expert precision states and completely specifies the local game under the fixed outer context $C_\pi(i)$. Let $\mathbf{h}_i^\pi = (\mathcal{H}_i^\pi(R))_{R \subseteq \mathcal{J}}$ denote the Harsanyi dividend vector. On the Boolean lattice $2^{\mathcal{J}}$, the zeta and Möbius representations are

$$\mathbf{v}_i^\pi = Z\,\mathbf{h}_i^\pi, \qquad \mathbf{h}_i^\pi = Z^{-1}\,\mathbf{v}_i^\pi, \qquad Z_{T,R} = \mathbf{1}[R \subseteq T], \qquad (Z^{-1})_{R,R'} = \mathbf{1}[R' \subseteq R]\,(-1)^{|R|-|R'|}. \tag{H.3}$$

Equivalently, for every $T, R \subseteq \mathcal{J}$,

$$V_i^\pi(T) = \sum_{R \subseteq T} \mathcal{H}_i^\pi(R), \qquad \mathcal{H}_i^\pi(R) = \sum_{R' \subseteq R} (-1)^{|R|-|R'|}\, V_i^\pi(R'). \tag{H.4}$$

The second identity is exactly the inner signed sum in Eq. (3) of the main paper. Therefore, Eq. (3) is the Harsanyi interaction expansion of the local precision game under the outer context $C_\pi(i)$.

H.2 Activated-Dividend Reconstruction

For a configuration A, define the activated projection subset

$$T(A) = \{j \in \mathcal{J} : a^j = b_{\text{high}}\}. \tag{H.5}$$

The interaction-aware local gain under context π is the sum of all non-empty Harsanyi dividends activated by A. Expanding this sum and collecting the coefficient of each state value $V_i^\pi(R')$ gives

$$\begin{aligned} G_i^\pi(A) &:= \sum_{\emptyset \neq R \subseteq T(A)} \mathcal{H}_i^\pi(R) \\ &= \sum_{\emptyset \neq R \subseteq T(A)} \sum_{R' \subseteq R} (-1)^{|R|-|R'|} V_i^\pi(R') \\ &= \sum_{R' \subseteq T(A)} V_i^\pi(R') \underbrace{\sum_{\substack{R:\, R' \subseteq R \subseteq T(A) \\ R \neq \emptyset}} (-1)^{|R|-|R'|}}_{\kappa_{R'}^{T(A)}}. \end{aligned} \tag{H.6}$$

The coefficient $\kappa_{R'}^{T(A)}$ admits a closed form. Setting $U = R \setminus R'$ yields

$$\kappa_{R'}^{T(A)} = \begin{cases} \displaystyle\sum_{U \subseteq T(A) \setminus R'} (-1)^{|U|}, & R' \neq \emptyset, \\ \displaystyle\sum_{\substack{U \subseteq T(A) \\ U \neq \emptyset}} (-1)^{|U|}, & R' = \emptyset, \end{cases} = \mathbf{1}[R' = T(A)] - \mathbf{1}[R' = \emptyset]. \tag{H.7}$$

Substituting this coefficient back yields

$$\begin{aligned} G_i^\pi(A) &= \sum_{R' \subseteq T(A)} V_i^\pi(R') \left(\mathbf{1}[R' = T(A)] - \mathbf{1}[R' = \emptyset]\right) \\ &= V_i^\pi(T(A)) - V_i^\pi(\emptyset). \end{aligned} \tag{H.8}$$

Thus, all intermediate precision states cancel through Möbius coefficient cancellation, and the full activated Harsanyi expansion reduces exactly to an endpoint difference. The endpoint evaluation is therefore not an interaction-free surrogate; it is the collapsed representation of the complete Harsanyi interaction expansion associated with A.

Remark (instantiation under three projections). Under $\mathcal{J} = \{u, g, d\}$, the local game decomposes into first-order projection effects, three pairwise Harsanyi dividends, and one third-order dividend; the eight local state values form a complete coordinate system for these dividends. The endpoint difference $V_i^\pi(T(A)) - V_i^\pi(\emptyset)$ recovers exactly the sum of the dividends activated by the selected local precision state.

H.3 Outer Averaging

By definition of the local characteristic function,

$$V_i^\pi(T(A)) = V_{\mathcal{C}}\left(S_{C_\pi(i)}^{(i\leftarrow A)}\right), \qquad V_i^\pi(\emptyset) = V_{\mathcal{C}}\left(S_{C_\pi(i)}^{(i\leftarrow A^0)}\right). \tag{H.9}$$

Averaging the endpoint reconstruction over uniformly sampled outer expert permutations yields

$$\begin{aligned} F_i(A) &= \mathbb{E}_{\pi\sim\mathrm{Unif}(\Pi(\mathcal{E}))}\left[\sum_{\emptyset\neq R\subseteq T(A)} \mathcal{H}_i^\pi(R)\right] \\ &= \mathbb{E}_{\pi\sim\mathrm{Unif}(\Pi(\mathcal{E}))}\left[V_{\mathcal{C}}\left(S_{C_\pi(i)}^{(i\leftarrow A)}\right) - V_{\mathcal{C}}\left(S_{C_\pi(i)}^{(i\leftarrow A^0)}\right)\right] = \widetilde{F}_i(A). \end{aligned} \tag{H.10}$$

This proves Theorem 1.

I Bias Analysis of Collapse-Aware Suffix Pruning

This section establishes the suffix-pruning bias bound for the configuration-indexed utility table. The omitted term is the configuration marginal utility, rather than an expert keep/remove marginal.

For each expert i, local configuration A, and outer permutation π, define

$$\Delta_i^\pi(A) = V_{\mathcal{C}_{\mathrm{est}}}\left(S_{C_\pi(i)}^{(i\leftarrow A)}\right) - V_{\mathcal{C}_{\mathrm{est}}}\left(S_{C_\pi(i)}^{(i\leftarrow A^0)}\right). \tag{I.1}$$

Let q denote the Plackett–Luce order proposal distribution used by RPOS, and let

$$w(\pi) = \frac{p_{\mathrm{unif}}(\pi)}{q(\pi)} = \frac{1}{|\mathcal{E}|!\, q(\pi)} \tag{I.2}$$

be the importance weight. The smoothed prior in Eq. (J.2) ensures full support:

$$q(\pi) > 0, \qquad \forall\, \pi \in \Pi(\mathcal{E}). \tag{I.3}$$

For a permutation π, let $R_s^\pi = S^{\mathrm{ref}}_{\{\pi(s),\ldots,\pi(|\mathcal{E}|)\}}$. Let $t_\pi(i)$ denote the position of expert i. The event that i is reached before the full-to-small traversal terminates is

$$\Omega_i = \left\{\pi : \min_{s=1,\ldots,t_\pi(i)} V_{\mathcal{C}_{\mathrm{est}}}(R_s^\pi) \geq \tau\, V_{\mathcal{C}_{\mathrm{est}}}\left(S^{\mathrm{ref}}\right)\right\}, \tag{I.4}$$

and Ω_i^c its complement. The target before suffix pruning can be written under the proposal distribution as

$$\begin{aligned} F_i(A) &= \mathbb{E}_{\pi\sim\mathrm{Unif}}\left[\Delta_i^\pi(A)\right] \\ &= \mathbb{E}_{\pi\sim q}\left[w(\pi)\,\Delta_i^\pi(A)\right] \\ &= \mathbb{E}_{\pi\sim q}\left[w(\pi)\,\Delta_i^\pi(A)\mathbb{1}_{\Omega_i}(\pi)\right] + \mathbb{E}_{\pi\sim q}\left[w(\pi)\,\Delta_i^\pi(A)\mathbb{1}_{\Omega_i^c}(\pi)\right]. \end{aligned} \tag{I.5}$$

The population target of the CASP-pruned approximation is

$$F_i^{\mathrm{prune}}(A) = \mathbb{E}_{\pi\sim q}\left[w(\pi)\,\Delta_i^\pi(A)\mathbb{1}_{\Omega_i}(\pi)\right]. \tag{I.6}$$

Subtracting equation I.6 from equation I.5,

$$F_i(A) - F_i^{\mathrm{prune}}(A) = \mathbb{E}_{\pi\sim q}\left[w(\pi)\,\Delta_i^\pi(A)\mathbb{1}_{\Omega_i^c}(\pi)\right]. \tag{I.7}$$

Assume that the omitted configuration marginal is bounded when the suffix containing expert i is pruned:

$$|\Delta_i^\pi(A)| \le \varepsilon_{\text{omit}}, \qquad \forall\, \pi \in \Omega_i^c. \tag{I.8}$$

Then

$$\begin{aligned} |F_i(A) - F_i^{\text{prune}}(A)| &\le \mathbb{E}_{\pi\sim q}\left[w(\pi)\,|\Delta_i^\pi(A)|\,\mathbb{1}_{\Omega_i^c}(\pi)\right] \\ &\le \varepsilon_{\text{omit}}\,\mathbb{E}_{\pi\sim q}\left[w(\pi)\,\mathbb{1}_{\Omega_i^c}(\pi)\right]. \end{aligned} \tag{I.9}$$

Expanding the weighted missing mass,

$$\begin{aligned} \mathbb{E}_{\pi\sim q}\left[w(\pi)\mathbb{1}_{\Omega_i^c}(\pi)\right] &= \sum_{\pi\in\Pi(\mathcal{E})} q(\pi)\,\frac{1}{|\mathcal{E}|!\,q(\pi)}\,\mathbb{1}_{\Omega_i^c}(\pi) \\ &= \frac{1}{|\mathcal{E}|!}\sum_{\pi\in\Pi(\mathcal{E})}\mathbb{1}_{\Omega_i^c}(\pi) \\ &= \Pr_{\pi\sim\text{Unif}}\left[\Omega_i^c\right]. \end{aligned} \tag{I.10}$$

Combining equation I.9 and equation I.10,

$$|F_i(A) - F_i^{\text{prune}}(A)| \le \varepsilon_{\text{omit}}\,\Pr_{\pi\sim\text{Unif}}\left[\min_{s=1,\dots,t_\pi(i)} V_{\mathcal{C}_{\text{est}}}(R_s^\pi) < \tau\, V_{\mathcal{C}_{\text{est}}}\left(S^{\text{ref}}\right)\right]. \tag{I.11}$$

Defining the aggregate pruning bias by $\varepsilon_{\text{prune}} := \max_{i,A}|F_i(A) - F_i^{\text{prune}}(A)|$, Eq. (I.11) gives $\varepsilon_{\text{prune}} \le \varepsilon_{\text{omit}}\max_i \Pr_{\pi\sim\text{Unif}}[\Omega_i^c]$. This establishes the suffix-pruning bias bound for configuration marginal utilities. □

J Importance Correction and Variance Analysis of RPOS

This section restates the importance-sampling identity that justifies the RPOS proposal, instantiated for the configuration marginal utility $\Delta_i^\pi(A)$.

J.1 Plackett–Luce Proposal and Importance Weights

Let $\Pi(\mathcal{E})$ denote the set of all expert permutations. The uniform target distribution is

$$p_{\text{unif}}(\pi) = \frac{1}{|\mathcal{E}|!}. \tag{J.1}$$

For expert i, let a_i denote its empirical mean routing mass and smooth the normalized proposal mass with $\epsilon_r > 0$:

$$a_i = \frac{1}{|\mathcal{C}_{\text{est}}|}\sum_{x\in\mathcal{C}_{\text{est}}}\mathbb{E}[G_i(x)], \qquad p_i = \frac{a_i + \epsilon_r}{\sum_{j\in\mathcal{E}}(a_j + \epsilon_r)}. \tag{J.2}$$

The Plackett–Luce order proposal used by RPOS is

$$q(\pi) = \prod_{t=1}^{|\mathcal{E}|}\frac{p_{\pi(t)}}{\sum_{s\ge t} p_{\pi(s)}}. \tag{J.3}$$

The corresponding importance weight is

$$w(\pi) = \frac{p_{\text{unif}}(\pi)}{q(\pi)} = \frac{1}{|\mathcal{E}|!}\prod_{t=1}^{|\mathcal{E}|}\frac{\sum_{s\ge t} p_{\pi(s)}}{p_{\pi(t)}}. \tag{J.4}$$

The expanded form in equation J.4 makes explicit how router-prior order sampling is corrected back to the uniform outer-permutation target.

J.2 Unbiasedness Before Suffix Pruning

For fixed i and A, the target utility is

$$F_i(A) = \frac{1}{|\mathcal{E}|!} \sum_{\pi \in \Pi(\mathcal{E})} \Delta_i^\pi(A). \tag{J.5}$$

The RPOS sample average before suffix pruning is

$$\widehat{F}_i^{\mathrm{RPOS}}(A) = \frac{1}{M} \sum_{m=1}^{M} w(\pi^{(m)})\, \Delta_i^{\pi^{(m)}}(A), \qquad \pi^{(m)} \sim q. \tag{J.6}$$

Taking expectation under q,

$$\begin{aligned} \mathbb{E}_q\left[\widehat{F}_i^{\mathrm{RPOS}}(A)\right] &= \mathbb{E}_q\left[w(\pi)\, \Delta_i^\pi(A)\right] \\ &= \sum_{\pi \in \Pi(\mathcal{E})} q(\pi)\, \frac{p_{\mathrm{unif}}(\pi)}{q(\pi)}\, \Delta_i^\pi(A) \\ &= \sum_{\pi \in \Pi(\mathcal{E})} p_{\mathrm{unif}}(\pi)\, \Delta_i^\pi(A) \\ &= F_i(A). \end{aligned} \tag{J.7}$$

Thus, before suffix pruning, RPOS changes only the order-sampling distribution and not the target expectation. When combined with CASP, the sample average becomes

$$\widehat{F}_i^{\mathrm{RPOS+CASP}}(A) = \frac{1}{M} \sum_{m=1}^{M} w(\pi^{(m)})\, \Delta_i^{\pi^{(m)}}(A) \mathbb{1}_{\not\vdash \Omega_i}(\pi^{(m)}), \tag{J.8}$$

whose expectation is $F_i^{\mathrm{prune}}(A)$. The difference between $F_i^{\mathrm{prune}}(A)$ and $F_i(A)$ is precisely the suffix-pruning bias bound in Appendix I.

J.3 Second-Moment Form of the RPOS Variance

For the sample average before suffix pruning, define

$$Z_{i,A}(\pi) = w(\pi)\, \Delta_i^\pi(A). \tag{J.9}$$

Then

$$\mathrm{Var}\left[\widehat{F}_i^{\mathrm{RPOS}}(A)\right] = \frac{1}{M} \mathrm{Var}_{\pi \sim q}\left[Z_{i,A}(\pi)\right]. \tag{J.10}$$

Expanding the variance,

$$\begin{aligned} \mathrm{Var}_q\left[Z_{i,A}(\pi)\right] &= \mathbb{E}_q\left[w(\pi)^2 \left(\Delta_i^\pi(A)\right)^2\right] - F_i(A)^2 \\ &= \sum_{\pi \in \Pi(\mathcal{E})} q(\pi) \left(\frac{p_{\mathrm{unif}}(\pi)}{q(\pi)}\right)^2 \left(\Delta_i^\pi(A)\right)^2 - F_i(A)^2 \\ &= \sum_{\pi \in \Pi(\mathcal{E})} \frac{p_{\mathrm{unif}}(\pi)^2}{q(\pi)} \left(\Delta_i^\pi(A)\right)^2 - F_i(A)^2. \end{aligned} \tag{J.11}$$

For comparison, the uniform Monte Carlo average has variance

$$\mathrm{Var}\left[\widehat{F}_i^{\mathrm{MC}}(A)\right] = \frac{1}{M} \left(\sum_{\pi \in \Pi(\mathcal{E})} p_{\mathrm{unif}}(\pi) \left(\Delta_i^\pi(A)\right)^2 - F_i(A)^2 \right). \tag{J.12}$$

Therefore, RPOS has no larger variance than uniform sampling for the pair (i, A) whenever

$$\sum_{\pi \in \Pi(\mathcal{E})} \frac{p_{\mathrm{unif}}(\pi)^2}{q(\pi)} \left(\Delta_i^\pi(A)\right)^2 \le \sum_{\pi \in \Pi(\mathcal{E})} p_{\mathrm{unif}}(\pi) \left(\Delta_i^\pi(A)\right)^2. \tag{J.13}$$

This condition is the standard second-moment criterion for variance reduction in importance sampling.

J.4 Optimal Proposal and Router Alignment

The variance-minimizing proposal for computing $F_i(A)$ is proportional to the absolute target contribution:

$$q^{\star}_{i,A}(\pi) = \frac{p_{\mathrm{unif}}(\pi)\,|\Delta_i^{\pi}(A)|}{\sum_{\pi' \in \Pi(\mathcal{E})} p_{\mathrm{unif}}(\pi')\,|\Delta_i^{\pi'}(A)|}. \tag{J.14}$$

Substituting $q^{\star}_{i,A}$ into equation J.11 yields the minimum possible second moment

$$\left(\sum_{\pi \in \Pi(\mathcal{E})} p_{\mathrm{unif}}(\pi)\,|\Delta_i^{\pi}(A)| \right)^2. \tag{J.15}$$

RPOS uses the readily available mean routing mass as a lightweight surrogate for informative expert contexts. When mean router masses are aligned with the magnitude of configuration marginal utilities, the Plackett–Luce proposal in equation J.3 moves probability mass toward permutations that better approximate the optimal proposal in equation J.14, thereby reducing the second moment in equation J.11.

K Uniform Concentration of the Configuration Utility Table

This section proves the uniform concentration bound for the RPOS sample average around its suffix-pruned population mean. The smoothed prior in equation J.2 satisfies $p_i \geq \epsilon_r / \sum_j (a_j + \epsilon_r) > 0$. Hence, the Plackett–Luce proposal density admits a uniform lower bound $q(\pi) \geq q_{\min}$ over $\Pi(\mathcal{E})$, so the importance weight $w(\pi) \leq 1/(|\mathcal{E}|!\, q_{\min}) =: w_{\max}$ is bounded. Combined with the bounded marginal $|\Delta_i^{\pi}(A)| \leq G_{\Delta}$, the importance-weighted suffix-pruned random variable is bounded, with effective constant $G_{\max} := w_{\max}\, G_{\Delta}$ absorbing the importance-weight bound.

For each table entry (i, A), define the importance-weighted suffix-pruned random variable

$$Z_m(i, A) = w(\pi^{(m)})\, \mathbb{1}_{\Omega_i}(\pi^{(m)})\, \Delta_i^{\pi^{(m)}}(A), \qquad \pi^{(m)} \sim q. \tag{K.1}$$

Then

$$\widehat{F}_i(A) = \frac{1}{M} \sum_{m=1}^{M} Z_m(i, A), \qquad F_i^{\mathrm{prune}}(A) = \mathbb{E}_q[Z_m(i, A)]. \tag{K.2}$$

By the bounds above, the importance-weighted suffix-pruned marginal is uniformly bounded:

$$|Z_m(i, A)| \leq G_{\max}, \qquad \forall\, i \in \mathcal{E},\ A \in \mathcal{B}^{|\mathcal{J}|},\ m \in \{1, \ldots, M\}. \tag{K.3}$$

For fixed (i, A), the variables $Z_1(i, A), \ldots, Z_M(i, A)$ are independent and bounded in $[-G_{\max}, G_{\max}]$. By Hoeffding's inequality,

$$\Pr\left[\left| \widehat{F}_i(A) - F_i^{\mathrm{prune}}(A) \right| > \epsilon \right] \leq 2 \exp\left(-\frac{M\,\epsilon^2}{2\,G_{\max}^2} \right). \tag{K.4}$$

The utility table contains $|\mathcal{E}| \cdot |\mathcal{B}|^{|\mathcal{J}|}$ entries. Applying the union bound over all entries,

$$\begin{aligned} &\Pr\left[\max_{i \in \mathcal{E}} \sup_{A \in \mathcal{B}^{|\mathcal{J}|}} \left| \widehat{F}_i(A) - F_i^{\mathrm{prune}}(A) \right| > \epsilon \right] \\ &\quad \leq \sum_{i \in \mathcal{E}} \sum_{A \in \mathcal{B}^{|\mathcal{J}|}} \Pr\left[\left| \widehat{F}_i(A) - F_i^{\mathrm{prune}}(A) \right| > \epsilon \right] \\ &\quad \leq 2\,|\mathcal{E}|\,|\mathcal{B}|^{|\mathcal{J}|} \exp\left(-\frac{M\,\epsilon^2}{2\,G_{\max}^2} \right). \end{aligned} \tag{K.5}$$

Setting the right-hand side to δ and solving for ϵ gives

$$\epsilon = G_{\max} \sqrt{\frac{2}{M} \log\left(\frac{2\,|\mathcal{E}|\,|\mathcal{B}|^{|\mathcal{J}|}}{\delta} \right)}. \tag{K.6}$$

Therefore, with probability at least $1-\delta$,

$$\max_{i\in\mathcal{E}} \sup_{A\in\mathcal{B}^{|\mathcal{J}|}} \left|\widehat{F}_i(A) - F_i^{\text{prune}}(A)\right| \le G_{\max}\sqrt{\frac{2}{M}\log\left(\frac{2\,|\mathcal{E}|\,|\mathcal{B}|^{|\mathcal{J}|}}{\delta}\right)}. \tag{K.7}$$

This establishes uniform concentration for the entire expert-configuration utility table. □

L Broader Impact

This work studies mixed-precision quantization for multimodal MoE models. Its potential positive societal impact is to reduce weight-storage requirements and, when efficient low-bit kernels are available, may lower deployment cost and energy use while making multimodal systems more accessible under limited hardware budgets. At the same time, more efficient deployment may also make powerful multimodal models easier to use in harmful applications, such as generating misleading content, automated surveillance, or other misuse enabled by the underlying base models. Our method does not introduce new model capabilities or release new datasets, and responsible deployment should follow the safety policies, usage restrictions, and safeguards associated with the corresponding base models and benchmarks.

M Existing Assets and Licenses

This work uses existing public models, software tools, evaluation benchmarks, and calibration datasets, including DeepSeek-VL2 (Wu et al., 2024), MolmoE-1B (Deitke et al., 2025), Mixtral-8×7B (Jiang et al., 2024), VLMEvalKit (Duan et al., 2024), GPTQ (Frantar et al., 2023), C4 (Raffel et al., 2020), MMBench (Liu et al., 2024b), MMStar (Chen et al., 2024), MME (Fu et al., 2025), MMMU (Yue et al., 2024), AI2D (Kembhavi et al., 2017), OCRBench (Liu et al., 2024c), DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), ScienceQA (Lu et al., 2022), TextVQA (Singh et al., 2019), and ChartQA (Masry et al., 2022). We cite the original papers or repositories of these assets where they are introduced or used in our experimental protocol. All assets are used only for research evaluation and post-training quantization experiments, and we do not redistribute the original models or datasets. We follow the licenses and terms of use specified by the corresponding model, dataset, benchmark, and software-tool providers. Mixtral-8×7B is obtained from the Mistral AI model repository and used under the Apache 2.0 license.

N STRUCTURAL INTERACTION GEOMETRY OF GATED-EXPERT QUANTIZATION

This section establishes that configuration-level interactions arise directly from the algebraic structure of the gated feed-forward network, rather than from the particular cooperative-game representation introduced in the main text. In particular, jointly quantizing the up, gate, and down projections induces mixed perturbation terms that cannot, in general, be represented by independent scalar projection sensitivities.

N.1 Exact perturbation decomposition

Consider routing expert i with gated feed-forward mapping

$$\mathcal{E}_i(h) = W_i^d\left[\sigma(W_i^g h)\odot(W_i^u h)\right], \tag{N.1}$$

where W_i^u, W_i^g, W_i^d denote the up, gate, and down projections, respectively. We omit the expert index whenever no ambiguity arises.

Let a local quantization configuration induce perturbations

$$\widehat{W}^u = W^u + \Delta^u, \qquad \widehat{W}^g = W^g + \Delta^g, \qquad \widehat{W}^d = W^d + \Delta^d. \tag{N.2}$$

Define $u = W^u h, \qquad a = \sigma(W^g h)$, together with

$$\delta u = \Delta^u h, \qquad \delta a = \sigma((W^g + \Delta^g)h) - \sigma(W^g h). \tag{N.4}$$

Lemma N.1. Exact seven-term decomposition The quantization perturbation of the expert output satisfies

$$\begin{aligned}\widehat{\mathcal{E}}(h) - \mathcal{E}(h) = &\underbrace{W^d(a \odot \delta u)}_{\mathcal{T}_u} + \underbrace{W^d(\delta a \odot u)}_{\mathcal{T}_g} + \underbrace{\Delta^d(a \odot u)}_{\mathcal{T}_d} \\ &+ \underbrace{W^d(\delta a \odot \delta u)}_{\mathcal{T}_{ug}} + \underbrace{\Delta^d(a \odot \delta u)}_{\mathcal{T}_{ud}} \\ &+ \underbrace{\Delta^d(\delta a \odot u)}_{\mathcal{T}_{gd}} + \underbrace{\Delta^d(\delta a \odot \delta u)}_{\mathcal{T}_{ugd}}.\end{aligned} \tag{N.5}$$

The first three terms correspond to isolated perturbations of individual projections, the next three terms are pairwise cross-projection interactions, and the last term is a third-order interaction.

Proof Substituting Eq. (N.2),

$$\widehat{\mathcal{E}}(h) = (W^d + \Delta^d)\left[(a + \delta a) \odot (u + \delta u)\right].$$

Expanding the Hadamard product,

$$(a + \delta a) \odot (u + \delta u) = a \odot u + a \odot \delta u + \delta a \odot u + \delta a \odot \delta u.$$

Left-multiplication by $W^d + \Delta^d$ gives eight terms. Removing the original unquantized contribution $W^d(a \odot u)$ leaves exactly the seven terms in Eq. (N.5). □

N.2 Mixed Fréchet geometry

The same phenomenon can be characterized differentially. Consider

$$\mathfrak{E} : (W^u, W^g, W^d) \mapsto W^d\left[\sigma(W^g h) \odot W^u h\right]. \tag{N.6}$$

Assume that σ is differentiable in a neighborhood of $W^g h$. The mixed Fréchet derivatives across different projections satisfy

$$D^2_{u,g}\mathfrak{E}[\Delta^u, \Delta^g] = W^d\left[D\sigma_{W^g h}[\Delta^g h] \odot (\Delta^u h)\right], \tag{N.7}$$

$$D^2_{u,d}\mathfrak{E}[\Delta^u, \Delta^d] = \Delta^d\left[\sigma(W^g h) \odot (\Delta^u h)\right], \tag{N.8}$$

The remaining mixed derivative is

$$D^2_{g,d}\mathfrak{E}[\Delta^g, \Delta^d] = \Delta^d\left[D\sigma_{W^g h}[\Delta^g h] \odot W^u h\right]. \tag{N.9}$$

Moreover,

$$D^3_{u,g,d}\mathfrak{E}[\Delta^u, \Delta^g, \Delta^d] = \Delta^d\left[D\sigma_{W^g h}[\Delta^g h] \odot (\Delta^u h)\right]. \tag{N.10}$$

Proposition N.2. Structural non-separability Unless all cross-projection mixed derivatives in Eqs. (N.7)–(N.10) vanish over the relevant perturbation directions, there do not exist functions f_u, f_g, f_d such that the local quantization response is represented exactly by

$$\mathcal{R}(b_u, b_g, b_d) = f_u(b_u) + f_g(b_g) + f_d(b_d) + c \tag{N.11}$$

for every admissible joint precision configuration.

Proof Every function of the form in Eq. (N.11) has identically zero mixed derivatives between distinct coordinates. Therefore, any non-zero mixed derivative of the true response contradicts separability. □

This result provides a structural explanation for configuration-level attribution: cross-projection interactions are induced by the gated computation itself before any cooperative-game construction is introduced.

N.3 Interaction-aware perturbation bounds

Assumption N.3 Assume $\|h\|_2 \le R, \qquad \|W^u\|_2 \le M_u, \qquad \|W^d\|_2 \le M_d$, and suppose σ is L_σ-Lipschitz with $\|\sigma(W^g h)\|_\infty \le M_\sigma$. Let

$$\epsilon_u = \|\Delta^u\|_2, \qquad \epsilon_g = \|\Delta^g\|_2, \qquad \epsilon_d = \|\Delta^d\|_2. \tag{N.14}$$

Lemma N.4 Under Assumption N.3,

$$\begin{aligned}\|\widehat{\mathcal{E}}(h) - \mathcal{E}(h)\|_2 \le\ & C_u\epsilon_u + C_g\epsilon_g + C_d\epsilon_d \\ & + C_{ug}\epsilon_u\epsilon_g + C_{ud}\epsilon_u\epsilon_d + C_{gd}\epsilon_g\epsilon_d \\ & + C_{ugd}\epsilon_u\epsilon_g\epsilon_d,\end{aligned} \tag{N.15}$$

where one valid choice is

$$C_u = M_d M_\sigma R, \qquad C_g = M_d L_\sigma M_u R^2, \qquad C_d = M_\sigma M_u R, \tag{N.16}$$

$$C_{ug} = M_d L_\sigma R^2, \qquad C_{ud} = M_\sigma R, \qquad C_{gd} = L_\sigma M_u R^2, \tag{N.17}$$

The third-order coefficient is

$$C_{ugd} = L_\sigma R^2. \tag{N.18}$$

Proof By Lipschitz continuity, $\|\delta a\|_2 \le L_\sigma \|\Delta^g h\|_2 \le L_\sigma \epsilon_g R$, while $\|\delta u\|_2 \le \epsilon_u R, \qquad \|u\|_2 \le M_u R$. Using $\|x \odot y\|_2 \le \|x\|_\infty \|y\|_2$ for the seven terms in Eq. (N.5) yields Eqs. (N.16)–(N.18). Summing these bounds completes the proof. □

Assumption N.5. Quantization-resolution scaling Suppose the quantizer satisfies $\|\Delta^j(b_j)\|_2 \le \kappa_j 2^{-b_j}, \qquad j \in \{u, g, d\}$. Then Eq. (N.15) implies

$$\begin{aligned}\|\widehat{\mathcal{E}} - \mathcal{E}\|_2 \le\ & \alpha_u 2^{-b_u} + \alpha_g 2^{-b_g} + \alpha_d 2^{-b_d} \\ & + \alpha_{ug} 2^{-(b_u+b_g)} + \alpha_{ud} 2^{-(b_u+b_d)} + \alpha_{gd} 2^{-(b_g+b_d)} \\ & + \alpha_{ugd} 2^{-(b_u+b_g+b_d)},\end{aligned} \tag{N.22}$$

for constants $\alpha_\cdot$ determined by the unquantized expert and quantizer.

Equation (N.22) makes explicit that low-bit distortion generally contains precision-coupled terms rather than only independent one-dimensional sensitivities.

O CONTINUOUS–DISCRETE INTERACTION CORRESPONDENCE

We next connect the structural mixed perturbations above to the Harsanyi interaction terms used by the proposed configuration-valued utility.

O.1 CONTINUOUS PRECISION INTERPOLATION

Fix expert i and outer permutation π. For projection $j \in \mathcal{J}$, interpolate between its low- and high-precision realizations by

$$W_j(t_j) = W_j^{\text{low}} + t_j \left(W_j^{\text{high}} - W_j^{\text{low}} \right), \qquad t_j \in [0,1]. \tag{O.1}$$

Let $\mathbf{t} = (t_j)_{j \in \mathcal{J}} \in [0,1]^{|\mathcal{J}|}$. Define the continuously extended local calibration response

$$f_i^\pi(\mathbf{t}) = V_{\mathcal{C}} \left(S_{C_\pi(i)}^{i \leftarrow \mathbf{t}} \right). \tag{O.2}$$

For $R \subseteq \mathcal{J}$, let $\mathbf{1}_R$ denote the Boolean vertex whose coordinates in R equal one and all remaining coordinates equal zero. Then the Harsanyi coefficient is

$$H_i^\pi(R) = \sum_{T \subseteq R} (-1)^{|R|-|T|} f_i^\pi(\mathbf{1}_T). \tag{O.3}$$

Theorem O.1. Integrated mixed-derivative representation Suppose f_i^π is $C^{|R|}$ in the coordinates indexed by R. Then

$$H_i^\pi(R) = \int_{[0,1]^{|R|}} \frac{\partial^{|R|} f_i^\pi}{\prod_{j \in R} \partial t_j} \left(\mathbf{t}_R, \mathbf{0}_{\mathcal{J} \setminus R} \right) d\mathbf{t}_R. \tag{O.4}$$

Proof For one coordinate, $f(1) - f(0) = \displaystyle\int_0^1 f'(t)\, dt$. Applying this identity iteratively to all coordinates in R yields

$$\Delta_R f(\mathbf{0}) = \int_{[0,1]^{|R|}} \partial_R f(\mathbf{t}_R, \mathbf{0}_{\bar{R}}) d\mathbf{t}_R. \tag{O.6}$$

The iterated finite difference satisfies

$$\Delta_R f(\mathbf{0}) = \sum_{T \subseteq R} (-1)^{|R|-|T|} f(\mathbf{1}_T), \tag{O.7}$$

which equals $H_i^\pi(R)$ by Eq. (O.3). □

Thus the discrete Harsanyi interaction is precisely an integrated mixed precision sensitivity over the interpolation hypercube.

Corollary O.2 If, over the interpolation region,

$$\left| \frac{\partial^{|R|} f_i^\pi}{\prod_{j \in R} \partial t_j} \right| \geq m_R > 0 \tag{O.9}$$

and the mixed derivative has a constant sign, then

$$|H_i^\pi(R)| \geq m_R. \tag{O.10}$$

Hence persistent mixed sensitivity necessarily induces non-zero configuration interaction.

O.2 Expressiveness limits of scalar projection attribution

Let $v : 2^{\mathcal{J}} \to \mathbb{R}$ be a fixed-context local precision game, with Möbius representation $v(A) = \sum_{R \subseteq A} H_v(R)$. Define the additive function family $\mathfrak{A}_1 = \left\{ g : g(A) = c + \sum_{j \in A} s_j \right\}$.

Theorem O.3. Exact additive representation criterion There exists $g \in \mathfrak{A}_1$ satisfying $v(A) = g(A), \quad \forall A \subseteq \mathcal{J}$, if and only if

$$H_v(R) = 0, \qquad \forall R \subseteq \mathcal{J}, \quad |R| \geq 2. \tag{O.14}$$

Proof The Möbius representation of a set function is unique. Every additive set function has zero Möbius coefficients of order at least two. Therefore equality over the complete Boolean lattice is possible if and only if the higher-order Möbius coefficients of v vanish. □

Theorem O.4. Irreducible L_∞ approximation error For every additive projection-scoring function $g \in \mathfrak{A}_1$,

$$\|v - g\|_\infty \geq \max_{\substack{R \subseteq \mathcal{J} \\ |R| \geq 2}} 2^{-|R|} |H_v(R)|. \tag{O.15}$$

Proof Let $e(A) = v(A) - g(A)$. Since $H_g(R) = 0$ for $|R| \geq 2$, $H_v(R) = \sum_{T \subseteq R} (-1)^{|R|-|T|} e(T)$. Hence $|H_v(R)| \leq \sum_{T \subseteq R} |e(T)| \leq 2^{|R|} \|e\|_\infty$. Rearranging proves Eq. (O.15). □

For the three projections $\mathcal{J} = \{u, g, d\}$, this gives

$$\inf_{g \in \mathfrak{A}_1} \|v - g\|_\infty \geq \max \left\{ \frac{|H_{ug}|}{4}, \frac{|H_{ud}|}{4}, \frac{|H_{gd}|}{4}, \frac{|H_{ugd}|}{8} \right\}. \tag{O.18}$$

O.3 Connection to the interaction ratio used in the experiments

Define $E_i = \sum_{\emptyset \neq R \subseteq \mathcal{J}} |H_i(R)|$ and $\rho_i = \dfrac{\sum_{|R| \geq 2} |H_i(R)|}{E_i}$. For $|\mathcal{J}| = 3$, there are four higher-order subsets. Hence $\max_{|R| \geq 2} |H_i(R)| \geq \dfrac{\rho_i E_i}{4}$. Since $2^{|R|} \leq 8$,

Corollary O.5

$$\inf_{g \in \mathfrak{A}_1} \|v_i - g\|_\infty \geq \frac{\rho_i E_i}{32}. \tag{O.22}$$

Therefore, the interaction ratio reported in the main experiments has a direct approximation-theoretic interpretation: non-negligible higher-order Harsanyi energy implies a non-zero worst-case error floor for every purely additive projection-level representation.

O.4 Orthogonal spectral characterization

The Harsanyi representation is natural for activated interactions, whereas an orthogonal representation provides an exact squared-error decomposition.

Encode a precision state by $x \in \{-1,+1\}^{|\mathcal{J}|}$. For $R \subseteq \mathcal{J}$, define the Walsh character $\chi_R(x) = \prod_{j\in R} x_j$. Every local precision response admits the unique Fourier–Walsh expansion

$$v(x) = \sum_{R\subseteq\mathcal{J}} \widehat{v}(R)\chi_R(x), \tag{O.24}$$

with Walsh coefficients

$$\widehat{v}(R) = 2^{-|\mathcal{J}|} \sum_{x\in\{-1,+1\}^{|\mathcal{J}|}} v(x)\chi_R(x). \tag{O.25}$$

Orthogonality gives $\mathbb{E}[\chi_R(X)\chi_T(X)] = \mathbb{I}[R=T]$. Consequently, $\mathrm{Var}[v(X)] = \sum_{R\neq\emptyset} \widehat{v}(R)^2$. Define the orthogonal higher-order interaction ratio

$$\eta(v) = \frac{\sum_{|R|\geq 2} \widehat{v}(R)^2}{\sum_{|R|\geq 1} \widehat{v}(R)^2}. \tag{O.28}$$

Theorem O.6. Optimal additive L_2 approximation Let the additive Walsh subspace be

$$\mathfrak{P}_1 = \mathrm{span}\{1, \chi_{\{u\}}, \chi_{\{g\}}, \chi_{\{d\}}\}. \tag{O.29}$$

The corresponding orthogonal projection and residual are

$$(\Pi_1 v)(x) = \widehat{v}(\emptyset) + \sum_{|R|=1} \widehat{v}(R)\chi_R(x),$$
$$(I-\Pi_1)v(x) = \sum_{|R|\geq 2} \widehat{v}(R)\chi_R(x).$$

Since the two components are orthogonal in $L_2(\{-1,+1\}^{|\mathcal{J}|})$, the Pythagorean decomposition is

$$v = \Pi_1 v + (I-\Pi_1)v, \qquad \langle \Pi_1 v, (I-\Pi_1)v\rangle = 0,$$
$$\mathbb{E}[(v-g)^2] = \mathbb{E}[(v-\Pi_1 v)^2] + \mathbb{E}[(\Pi_1 v - g)^2], \qquad g \in \mathfrak{P}_1.$$

Minimizing the second term yields

$$\inf_{g\in\mathfrak{P}_1} \mathbb{E}[(v(X)-g(X))^2] = \sum_{|R|\geq 2} \widehat{v}(R)^2. \tag{O.30}$$

Therefore,

$$\frac{\inf_{g\in\mathfrak{P}_1} \mathbb{E}[(v-g)^2]}{\mathrm{Var}[v]} = \eta(v). \tag{O.31}$$

Proof Because the Walsh characters form an orthonormal basis, the orthogonal projection onto $\mathfrak{P}_1$ retains exactly the coefficients of degree zero and one. Parseval's identity then gives Eq. (O.30). □

Thus higher-order spectral mass is exactly the fraction of local configuration variance that no additive projection model can explain.

Proposition O.7. Möbius–Walsh transform Let the Boolean variables in the Harsanyi representation be $z_j \in \{0,1\}$, let $x_j = 2z_j - 1$, and write $A(x) = \{j : x_j = +1\}$. Then

$$\begin{aligned}\mathbb{I}[T \subseteq A(x)] &= \prod_{j\in T} z_j = 2^{-|T|} \prod_{j\in T}(1+x_j) \\ &= 2^{-|T|} \sum_{R\subseteq T} \chi_R(x), \\ v(x) &= \sum_{T\subseteq \mathcal{J}} H_v(T)\mathbb{I}[T \subseteq A(x)] \\ &= \sum_{R\subseteq \mathcal{J}} \left(\sum_{T\supseteq R} 2^{-|T|} H_v(T) \right) \chi_R(x).\end{aligned}$$

Matching Walsh coefficients gives

$$\widehat{v}(R) = \sum_{T\supseteq R} 2^{-|T|} H_v(T), \tag{O.32}$$

and conversely,

$$\begin{aligned}2^{-|T|} H_v(T) &= \sum_{R\supseteq T} (-1)^{|R|-|T|} \widehat{v}(R), \\ H_v(T) &= 2^{|T|} \sum_{R\supseteq T} (-1)^{|R|-|T|} \widehat{v}(R).\end{aligned} \tag{O.33}$$

Hence the Harsanyi and Walsh views are two invertible coordinate systems of the same local precision landscape.

P BI-LEVEL MÖBIUS THEORY OF CONFIGURATION-VALUED OWEN ATTRIBUTION

This section formalizes the two sources of interaction modeled by the proposed method: local interactions among the up, gate, and down projections, and contextual interactions induced by the precision states of other routing experts.

P.1 PRODUCT BOOLEAN GAME

Fix expert i and let $\mathcal{E}_{-i} = \mathcal{E} \setminus \{i\}$. For $R \subseteq \mathcal{J}, \qquad C \subseteq \mathcal{E}_{-i}$, define $v_i(R, C) = V_{\mathcal{C}}\left(S_C^{i\leftarrow A_R}\right)$, where A_R assigns high precision to projections in R and baseline precision to projections outside R; experts in C are at their reference state and all remaining routing experts are at their baseline state.

Therefore, $v_i : 2^{\mathcal{J}} \times 2^{\mathcal{E}_{-i}} \to \mathbb{R}$. Define the bi-level Möbius interaction tensor

$$\Gamma_i(R,T) = \sum_{R'\subseteq R} \sum_{T'\subseteq T} (-1)^{|R|-|R'|+|T|-|T'|} v_i(R',T'). \tag{P.3}$$

The first coordinate captures within-expert projection interaction; the second captures interaction between the local precision change of expert i and the outer precision context.

P.2 MATRIX AND TENSOR REPRESENTATION

Let $p = |\mathcal{J}|, \qquad n = |\mathcal{E}|-1$. Arrange all values $v_i(R,C)$ into $\mathbf{V}_i \in \mathbb{R}^{2^p \times 2^n}$. Define Boolean-lattice zeta matrices $(\mathbf{Z}_p)_{A,R} = \mathbb{I}[R \subseteq A]$, $(\mathbf{Z}_n)_{C,T} = \mathbb{I}[T \subseteq C]$,

and Möbius matrices $\mathbf{M}_p = \mathbf{Z}_p^{-1}, \qquad \mathbf{M}_n = \mathbf{Z}_n^{-1}$.

Theorem P.1. Bi-level Möbius transform The complete interaction tensor satisfies

$$\mathbf{\Gamma}_i = \mathbf{M}_p \mathbf{V}_i \mathbf{M}_n^\top, \tag{P.8}$$

with inverse transform

$$\mathbf{V}_i = \mathbf{Z}_p \mathbf{\Gamma}_i \mathbf{Z}_n^\top. \tag{P.9}$$

Vectorization gives

$$\mathrm{vec}(\mathbf{\Gamma}_i) = (\mathbf{M}_n \otimes \mathbf{M}_p)\,\mathrm{vec}(\mathbf{V}_i). \tag{P.10}$$

Proof The Möbius function of the Cartesian product of two finite posets factorizes into the product of the Möbius functions of the two component posets. Applying Möbius inversion along the local axis and outer-context axis gives Eqs. (P.8)–(P.10). □

This shows that the two-level method can be viewed as a change of basis from raw configuration/context evaluations to a tensor of pure hierarchical interaction coordinates.

P.3 Exact Config-Owen contraction formula

Define the local marginal under context C as $\Delta_i(A \mid C) = v_i(A, C) - v_i(\emptyset, C)$. From Eq. (P.9),

$$\Delta_i(A \mid C) = \sum_{\emptyset \neq R \subseteq A} \sum_{T \subseteq C} \Gamma_i(R, T). \tag{P.12}$$

The population configuration-valued utility is

$$F_i(A) = \mathbb{E}_\pi[\Delta_i(A \mid C_\pi(i))]. \tag{P.13}$$

For every fixed outer set $T \subseteq \mathcal{E}_{-i}$,

$$\Pr_\pi[T \subseteq C_\pi(i)] = \frac{1}{|T| + 1}, \tag{P.14}$$

because $T \subseteq C_\pi(i)$ exactly when i appears before every member of T in the relative ordering of $T \cup \{i\}$.

Theorem P.2. Exact hierarchical interaction decomposition

$$F_i(A) = \sum_{\emptyset \neq R \subseteq A} \sum_{T \subseteq \mathcal{E}_{-i}} \frac{\Gamma_i(R, T)}{|T| + 1}. \tag{P.15}$$

Proof Substitute Eq. (P.12) into Eq. (P.13):

$$\begin{aligned} F_i(A) &= \mathbb{E}_\pi \left[\sum_{\emptyset \neq R \subseteq A} \sum_{T \subseteq C_\pi(i)} \Gamma_i(R, T) \right] \\ &= \sum_{\emptyset \neq R \subseteq A} \sum_{T \subseteq \mathcal{E}_{-i}} \Gamma_i(R, T) \Pr[T \subseteq C_\pi(i)]. \end{aligned}$$

Applying Eq. (P.14) proves Eq. (P.15). □

Equation (P.15) is the central theoretical representation of the proposed attribution object: Config-Owen integrates local projection interactions jointly with outer expert-context interactions.

Define $r_A(R) = \mathbb{I}[\emptyset \neq R \subseteq A]$, and $\omega(T) = \dfrac{1}{|T|+1}$. Let $\mathbf{r}_A = (r_A(R))_R, \qquad \boldsymbol{\omega} = (\omega(T))_T$. Then Eq. (P.15) can be written compactly as

$$F_i(A) = \mathbf{r}_A^\top \boldsymbol{\Gamma}_i \boldsymbol{\omega}. \tag{P.19}$$

The equivalent Frobenius contraction is

$$F_i(A) = \left\langle \boldsymbol{\Gamma}_i,\, \mathbf{r}_A \boldsymbol{\omega}^\top \right\rangle_F. \tag{P.20}$$

This is a compact characterization of Config-Owen as a contraction of the complete hierarchical interaction tensor with a configuration-context kernel.

P.4 Projection-Owen as an additive reconstruction

For a fixed local game, the Shapley allocation to projection j distributes each local interaction $R \ni j$ uniformly among its $|R|$ members. After outer-context averaging, define

$$\phi_{ij} = \sum_{\substack{R \subseteq \mathcal{J} \\ j \in R}} \sum_T \frac{\Gamma_i(R,T)}{|R|(|T|+1)}. \tag{P.21}$$

An additive configuration reconstruction induced by projection-level Owen scores is $F_i^{\mathrm{proj}}(A) = \sum_{j \in A} \phi_{ij}$. Therefore,

$$F_i^{\mathrm{proj}}(A) = \sum_{\emptyset \neq R \subseteq \mathcal{J}} \sum_T \frac{|R \cap A|}{|R|(|T|+1)} \Gamma_i(R,T). \tag{P.23}$$

Comparing Eqs. (P.15) and (P.23) gives:

Theorem P.3. Exact projection-reconstruction mismatch

$$F_i(A) - F_i^{\mathrm{proj}}(A) = \sum_{\emptyset \neq R \subseteq \mathcal{J}} \sum_T \left[\mathbb{I}[R \subseteq A] - \frac{|R \cap A|}{|R|} \right] \frac{\Gamma_i(R,T)}{|T|+1}. \tag{P.24}$$

Only interacting sets R that are partially selected by A can generate a discrepancy. Indeed, if $R \subseteq A$, both coefficients equal one; if $R \cap A = \emptyset$, both coefficients equal zero; and if $0 < |R \cap A| < |R|$, projection-wise scoring allocates only a fraction of an interaction whose configuration-level activation is all-or-nothing.

Taking absolute values yields

$$|F_i(A) - F_i^{\mathrm{proj}}(A)| \leq \sum_{\substack{R: \\ 0<|R\cap A|<|R|}} \sum_T \frac{|\Gamma_i(R,T)|}{|T|+1}. \tag{P.25}$$

This theorem gives a direct theoretical interpretation to the Projection-Owen ablation: the discrepancy arises precisely when a mixed-precision configuration splits an interacting group of projections.

P.5 FIXED-CONTEXT LOCAL CALIBRATION

Let $C_0 \subseteq \mathcal{E}_{-i}$ denote the fixed outer context used by a local-calibration baseline. Define $F_i^{C_0}(A) = \Delta_i(A \mid C_0)$. By Eq. (P.12),

$$F_i^{C_0}(A) = \sum_{\emptyset \neq R \subseteq A} \sum_{T \subseteq C_0} \Gamma_i(R, T). \tag{P.27}$$

Theorem P.4. Fixed-context deviation

$$F_i(A) - F_i^{C_0}(A) = \sum_{\emptyset \neq R \subseteq A} \sum_{T} \left[\frac{1}{|T|+1} - \mathbb{I}[T \subseteq C_0] \right] \Gamma_i(R, T). \tag{P.28}$$

Thus a single local-calibration context is exact only when the omitted contextual interaction terms collectively vanish.

The special baseline context $C_0 = \emptyset$ gives

$$F_i^{\emptyset}(A) = \sum_{\emptyset \neq R \subseteq A} \Gamma_i(R, \emptyset), \tag{P.29}$$

Removing the baseline outer-context term gives

$$F_i(A) - F_i^{\emptyset}(A) = \sum_{\emptyset \neq R \subseteq A} \sum_{\emptyset \neq T} \frac{\Gamma_i(R, T)}{|T|+1}. \tag{P.30}$$

P.6 TWO-DIMENSIONAL INTERACTION SPECTRUM

Define the hierarchical interaction spectrum

$$\mathcal{S}_i(r, s) = \sum_{\substack{|R|=r \\ |T|=s}} \frac{|\Gamma_i(R, T)|}{s+1}. \tag{P.31}$$

The first coordinate r measures within-expert projection interaction order; the second coordinate s measures outer-context interaction order.

For the three-projection case,

$$\mathbf{S}_i = \begin{pmatrix} \mathcal{S}_i(1,0) & \mathcal{S}_i(1,1) & \mathcal{S}_i(1,2) & \cdots \\ \mathcal{S}_i(2,0) & \mathcal{S}_i(2,1) & \mathcal{S}_i(2,2) & \cdots \\ \mathcal{S}_i(3,0) & \mathcal{S}_i(3,1) & \mathcal{S}_i(3,2) & \cdots \end{pmatrix}. \tag{P.32}$$

This spectrum separates the two forms of structure targeted by the method:

$$r \geq 2 \quad \Longleftrightarrow \quad \text{non-additive local projection interaction,} \tag{P.33}$$

$$s \geq 1 \quad \Longleftrightarrow \quad \text{dependence on outer expert context.} \tag{P.34}$$

Theorem P.5. Hierarchical truncation bound Let $F_i^{(r_0, s_0)}(A)$ retain only interaction terms satisfying $|R| \leq r_0, \qquad |T| \leq s_0$. Then

$$|F_i(A) - F_i^{(r_0, s_0)}(A)| \leq \sum_{\substack{R \subseteq A,\ R \neq \emptyset \\ |R| > r_0 \ \vee \ |T| > s_0}} \frac{|\Gamma_i(R, T)|}{|T|+1}. \tag{P.35}$$

Equivalently, with weighted interaction norm $\|\Gamma\|_{1,\omega} = \sum_{R,T} \frac{|\Gamma(R,T)|}{|T|+1}$,

$$|F_i - F_i^{(r_0,s_0)}| \le \|(\mathbf{I} - \mathbf{P}_{r_0,s_0})\Gamma_i\|_{1,\omega}. \tag{P.37}$$

This formally links controlled attribution ablations to discarded hierarchical interaction mass.

P.7 Shapley endpoint consistency

Define the binary expert-reference game $\nu(C) = V_{\mathcal{C}}(S_C^{\mathrm{ref}})$. At the complete local reference configuration A_{ref}, $v_i(\mathcal{J}, C) = \nu(C \cup \{i\})$, whereas $v_i(\emptyset, C) = \nu(C)$.

Theorem P.6. Expert-level Shapley consistency

$$F_i(A_{\mathrm{ref}}) = \phi_i^{\mathrm{Shapley}}(\nu). \tag{P.39}$$

Shapley efficiency consequently gives

$$\sum_{i\in\mathcal{E}} F_i(A_{\mathrm{ref}}) = V_{\mathcal{C}}(S_{\mathrm{ref}}) - V_{\mathcal{C}}(S_0). \tag{P.40}$$

Moreover, $F_i(A_0) = 0$. Thus configuration-valued attribution extends, rather than replaces, the standard expert-level cooperative-game attribution: the latter is recovered exactly at the reference endpoint.

Q Joint variance theory of router-prior order sampling

Appendix J analyzes importance sampling separately for a single utility entry. Here we analyze the statistically relevant object induced by RPOS: the entire expert-configuration utility table computed from shared sampled orders.

Q.1 Vector-valued marginal process

Index the complete sampled utility table by

$$\mathcal{Q} = \left\{(i, A) : i \in \mathcal{E},\ A \in \mathcal{B}^{|\mathcal{J}|}\right\}, \qquad D = |\mathcal{Q}|. \tag{Q.1}$$

For a permutation π, define the vector of suffix-pruned configuration marginals

$$\mathbf{Y}(\pi) = \left(\mathbf{1}_{\Omega_i(\pi)}\Delta_i^{\pi}(A)\right)_{(i,A)\in\mathcal{Q}} \in \mathbb{R}^D. \tag{Q.2}$$

Let the target permutation distribution be $p(\pi) = \frac{1}{|\mathcal{E}|!}$. Then $\boldsymbol{\mu} = \mathbb{E}_{\pi\sim p}[\mathbf{Y}(\pi)]$. Under proposal q,

$$\widehat{\boldsymbol{\mu}} = \frac{1}{M}\sum_{m=1}^{M} \frac{p(\pi_m)}{q(\pi_m)}\mathbf{Y}(\pi_m), \qquad \pi_m \overset{\mathrm{iid}}{\sim} q. \tag{Q.4}$$

Proposition Q.1. Joint unbiasedness for the suffix-pruned table Provided $q(\pi) > 0$ wherever $p(\pi) > 0$,

$$\mathbb{E}_q[\widehat{\boldsymbol{\mu}}] = \boldsymbol{\mu}. \tag{Q.5}$$

This is the vector analogue of the entry-wise importance-correction identity in Appendix J.

Q.2 FULL COVARIANCE GEOMETRY

Define $\mathbf{Z}_q(\pi) = \dfrac{p(\pi)}{q(\pi)}\mathbf{Y}(\pi)$. Then

$$\mathrm{Cov}(\widehat{\boldsymbol{\mu}}) = \frac{1}{M}\boldsymbol{\Sigma}_q, \tag{Q.7}$$

with full covariance

$$\boldsymbol{\Sigma}_q = \sum_{\pi} \frac{p(\pi)^2}{q(\pi)}\mathbf{Y}(\pi)\mathbf{Y}(\pi)^\top - \boldsymbol{\mu}\boldsymbol{\mu}^\top. \tag{Q.8}$$

The off-diagonal entries of $\boldsymbol{\Sigma}_q$ quantify dependence between configuration utilities computed from the same sampled permutation.

Theorem Q.2. Trace-variance-optimal proposal Let $\mathcal{S}_+ = \{\pi : \|\mathbf{Y}(\pi)\|_2 > 0\}$. Among proposal distributions supported on $\mathcal{S}_+$ and positive throughout that support,

$$q^\star(\pi) = \frac{p(\pi)\|\mathbf{Y}(\pi)\|_2}{Z}, \quad \pi \in \mathcal{S}_+, \qquad Z = \mathbb{E}_p[\|\mathbf{Y}(\pi)\|_2], \tag{Q.9}$$

with $q^\star(\pi) = 0$ outside $\mathcal{S}_+$, minimizes $\mathrm{tr}\,\mathrm{Cov}(\widehat{\boldsymbol{\mu}})$.

Proof The proposal-dependent part of the trace covariance is

$$\mathcal{M}_2(q) = \sum_{\pi\in\mathcal{S}_+} \frac{p(\pi)^2\|\mathbf{Y}(\pi)\|_2^2}{q(\pi)}. \tag{Q.10}$$

By Cauchy–Schwarz,

$$\begin{aligned} Z^2 &= \left(\sum_{\pi\in\mathcal{S}_+} p(\pi)\|\mathbf{Y}(\pi)\|_2\right)^2 \\ &= \left(\sum_{\pi\in\mathcal{S}_+} \frac{p(\pi)\|\mathbf{Y}(\pi)\|_2}{\sqrt{q(\pi)}}\sqrt{q(\pi)}\right)^2 \\ &\le \left[\sum_{\pi\in\mathcal{S}_+} \frac{p(\pi)^2\|\mathbf{Y}(\pi)\|_2^2}{q(\pi)}\right]\left[\sum_{\pi\in\mathcal{S}_+} q(\pi)\right] \\ &= \mathcal{M}_2(q). \end{aligned} \tag{Q.11}$$

Equality holds precisely when $q(\pi) \propto p(\pi)\|\mathbf{Y}(\pi)\|_2$ on $\mathcal{S}_+$, which yields Eq. (Q.9). □

When global full support is required, the smoothed proposal

$$q_\eta^\star(\pi) = (1-\eta)q^\star(\pi) + \eta p(\pi), \qquad 0 < \eta < 1, \tag{Q.12}$$

is positive wherever p is positive and approaches the trace-optimal proposal as $\eta \downarrow 0$.

Unlike the single-entry proposal in Appendix J, Eq. (Q.9) optimizes computation of the complete utility table jointly.

Q.3 RÉNYI-DIVERGENCE CHARACTERIZATION

Because $p(\pi)\|\mathbf{Y}(\pi)\|_2 = Zq^\star(\pi)$, Eq. (Q.10) becomes $\mathcal{M}_2(q) = Z^2 \sum_\pi \frac{q^\star(\pi)^2}{q(\pi)}$. Recall the order-two Rényi divergence $D_2(P\|Q) = \log \sum_x \frac{P(x)^2}{Q(x)}$.

Theorem Q.3. Proposal-mismatch identity

$$\mathcal{M}_2(q) = Z^2 \exp\left(D_2(q^\star\|q)\right). \tag{Q.14}$$

The equivalent chi-square form is

$$\mathcal{M}_2(q) = Z^2\left[1 + \chi^2(q^\star\|q)\right]. \tag{Q.15}$$

Therefore,

$$\mathcal{M}_2(q) - \mathcal{M}_2(q^\star) = Z^2\chi^2(q^\star\|q). \tag{Q.16}$$

This equation gives a precise interpretation to the router-prior order proposal: its statistical efficiency is determined by how closely its probability mass matches the inaccessible variance-optimal permutation law.

Corollary Q.4 If $D_2(q^\star\|q_{\mathrm{RPOS}}) \le \varepsilon_q$, then

$$\mathcal{M}_2(q_{\mathrm{RPOS}}) \le e^{\varepsilon_q}\mathcal{M}_2(q^\star). \tag{Q.18}$$

This is a conditional rather than unconditional variance-reduction guarantee.

Q.4 INTRINSIC DIMENSION OF THE UTILITY TABLE

The nominal table dimension is D, but its effective statistical dimension may be substantially smaller because different configuration marginals share the same expert contexts.

The resulting effective dimension is

$$d_{\mathrm{eff}}(q) = \frac{\mathrm{tr}(\mathbf{\Sigma}_q)}{\|\mathbf{\Sigma}_q\|_{\mathrm{op}}}. \tag{Q.19}$$

Then $1 \le d_{\mathrm{eff}}(q) \le D$. A small d_{eff} indicates that most sampling uncertainty is concentrated in a low-dimensional subspace of the configuration-utility table, providing a statistical explanation for why one sampled outer permutation can simultaneously inform many candidate configurations.

R BUDGET GEOMETRY AND STABILITY OF THE MCKP ALLOCATION

The preceding appendices analyze the configuration utilities themselves. This section connects utility computation to the final mixed-precision allocation.

Let the budget-feasible allocation set be

$$\mathfrak{F}_B = \left\{\mathbf{A} = (A_i)_{i\in\mathcal{E}} : \sum_i m_i(A_i) \le B_{\mathrm{mem}}\right\}. \tag{R.1}$$

Define the population surrogate $\Phi(\mathbf{A}) = \sum_i F_i(A_i)$, and its empirical counterpart $\widehat{\Phi}(\mathbf{A}) = \sum_i \widehat{F}_i(A_i)$. Let $\mathbf{A}^\star = \arg\max_{\mathbf{A}\in\mathfrak{F}_B} \Phi(\mathbf{A})$ and

$$\widehat{\mathbf{A}} = \arg\max_{\mathbf{A}\in\mathfrak{F}_B} \widehat{\Phi}(\mathbf{A}). \tag{R.5}$$

R.1 Utility-table error

From the CASP suffix-pruning analysis and Appendix K, define $\varepsilon_{\text{prune}} = \max_{i,A} |F_i(A) - F_i^{\text{prune}}(A)|$. With probability at least $1-\delta$,

$$\varepsilon_{\text{stat}}(M,\delta) = G_{\max}\sqrt{\frac{2}{M}\log\frac{2|\mathcal{E}||\mathcal{B}|^{|\mathcal{J}|}}{\delta}}. \tag{R.7}$$

Therefore,

$$\|\widehat{F} - F\|_\infty \le \varepsilon_M := \varepsilon_{\text{prune}} + \varepsilon_{\text{stat}}(M,\delta). \tag{R.8}$$

Theorem R.1. MCKP stability under utility perturbation On the event in Eq. (R.8),

$$\Phi(\mathbf{A}^\star) - \Phi(\widehat{\mathbf{A}}) \le 2|\mathcal{E}|\varepsilon_M. \tag{R.9}$$

Proof For every feasible allocation,

$$|\widehat{\Phi}(\mathbf{A}) - \Phi(\mathbf{A})| \le |\mathcal{E}|\varepsilon_M. \tag{R.10}$$

The optimizer chain satisfies

$$\begin{aligned}\Phi(\mathbf{A}^\star) &\le \widehat{\Phi}(\mathbf{A}^\star) + |\mathcal{E}|\varepsilon_M \\ &\le \widehat{\Phi}(\widehat{\mathbf{A}}) + |\mathcal{E}|\varepsilon_M \\ &\le \Phi(\widehat{\mathbf{A}}) + 2|\mathcal{E}|\varepsilon_M.\end{aligned}$$

□

R.2 Exact allocation recovery

Define the MCKP allocation margin

$$\Gamma_B = \Phi(\mathbf{A}^\star) - \max_{\substack{\mathbf{A}\in\mathfrak{F}_B \\ \mathbf{A}\neq\mathbf{A}^\star}} \Phi(\mathbf{A}). \tag{R.11}$$

Theorem R.2 If

$$\Gamma_B > 2|\mathcal{E}|\varepsilon_M, \tag{R.12}$$

then exact recovery follows:

$$\widehat{\mathbf{A}} = \mathbf{A}^\star. \tag{R.13}$$

Indeed, for every feasible $\mathbf{A} \neq \mathbf{A}^\star$, the empirical objective gap obeys the complete perturbation chain

$$\begin{aligned}\widehat{\Phi}(\mathbf{A}^\star) - \widehat{\Phi}(\mathbf{A}) &= \Phi(\mathbf{A}^\star) - \Phi(\mathbf{A}) + [\widehat{\Phi} - \Phi](\mathbf{A}^\star) - [\widehat{\Phi} - \Phi](\mathbf{A}) \\ &\geq \Gamma_B - \left|[\widehat{\Phi} - \Phi](\mathbf{A}^\star)\right| - \left|[\widehat{\Phi} - \Phi](\mathbf{A})\right| \\ &\geq \Gamma_B - 2|\mathcal{E}|\varepsilon_M > 0,\end{aligned}$$

so no competing feasible allocation can maximize $\widehat{\Phi}$.

Corollary R.3. Finite-sample recovery condition Suppose $\Gamma_B > 2|\mathcal{E}|\varepsilon_{\text{prune}}$. Then exact recovery with probability at least $1 - \delta$ is guaranteed whenever

$$\begin{aligned}2|\mathcal{E}|\varepsilon_M < \Gamma_B \iff 2|\mathcal{E}|G_{\max}\sqrt{\frac{2}{M}\log\frac{2|\mathcal{E}||\mathcal{B}|^{|\mathcal{J}|}}{\delta}} < \Gamma_B - 2|\mathcal{E}|\varepsilon_{\text{prune}} \\ \Longleftarrow M > \frac{8|\mathcal{E}|^2 G_{\max}^2}{(\Gamma_B - 2|\mathcal{E}|\varepsilon_{\text{prune}})^2}\log\frac{2|\mathcal{E}||\mathcal{B}|^{|\mathcal{J}|}}{\delta}.\end{aligned} \tag{R.15}$$

This result connects the number of sampled expert permutations directly to recovery of the budget-optimal precision assignment.

R.3 Lagrangian geometry of the bit-budget frontier

Consider the Lagrangian relaxation

$$\mathcal{L}(\mathbf{A}, \lambda) = \sum_i F_i(A_i) - \lambda\left[\sum_i m_i(A_i) - B\right], \qquad \lambda \geq 0. \tag{R.16}$$

It can be rewritten as

$$\mathcal{L}(\mathbf{A}, \lambda) = \lambda B + \sum_i \left[F_i(A_i) - \lambda m_i(A_i)\right]. \tag{R.17}$$

Therefore, for fixed λ, the relaxed problem separates across experts:

$$A_i^\lambda \in \arg\max_A \left\{F_i(A) - \lambda m_i(A)\right\}. \tag{R.18}$$

Equivalently, define the upper envelope of the local affine utilities and the associated dual function by

$$\begin{aligned}g_i(\lambda) &:= \max_{A \in \mathcal{B}^{|\mathcal{J}|}} \{F_i(A) - \lambda m_i(A)\}, \\ g(\lambda) &:= \max_{\mathbf{A}} \mathcal{L}(\mathbf{A}, \lambda) = \lambda B + \sum_{i \in \mathcal{E}} g_i(\lambda), \\ A_i^\lambda &\in \arg\max_A \{F_i(A) - \lambda m_i(A)\}.\end{aligned}$$

The multiplier λ acts as a shadow price for memory.

For two configurations A, B with different memory costs, define the switching threshold

$$\begin{aligned}F_i(A) - \lambda m_i(A) = F_i(B) - \lambda m_i(B) \\ \iff \lambda = \lambda_{i;A,B} := \frac{F_i(A) - F_i(B)}{m_i(A) - m_i(B)}.\end{aligned} \tag{R.19}$$

Proposition R.4. Piecewise-stable precision path For every expert i, the Lagrangian-optimal configuration A_i^λ is piecewise constant in λ, with changes occurring only at the finite set $\Lambda_i = \{\lambda_{i;A,B} : A \neq B\}$. Thus changes in the mixed-precision allocation as the bit budget varies occur through discrete configuration switches rather than through a continuously varying scalar importance score.

R.4 Interaction-induced switching-threshold shifts

Let $F_i(A) = F_i^{(1)}(A) + I_i(A)$, where $F_i^{(1)}$ denotes an additive projection-level reconstruction and $I_i(A) = F_i(A) - F_i^{(1)}(A)$ is its interaction correction.

The configuration switching threshold under the full utility is $\lambda_{A,B} = \dfrac{F_i(A) - F_i(B)}{m_i(A) - m_i(B)}$, whereas the additive approximation produces

$$\lambda_{A,B}^{(1)} = \frac{F_i^{(1)}(A) - F_i^{(1)}(B)}{m_i(A) - m_i(B)}. \tag{R.24}$$

Therefore,

$$\lambda_{A,B} - \lambda_{A,B}^{(1)} = \frac{I_i(A) - I_i(B)}{m_i(A) - m_i(B)}. \tag{R.25}$$

Equation (R.25) shows that higher-order interaction directly shifts the budget threshold at which the optimizer changes precision configuration.

Theorem R.5. Interaction-induced decision reversal Assume $m_i(A) > m_i(B)$. Suppose an additive scorer prefers B at shadow price λ, $F_i^{(1)}(A) - F_i^{(1)}(B) < \lambda[m_i(A) - m_i(B)]$. The full configuration utility instead selects A whenever

$$I_i(A) - I_i(B) > \lambda[m_i(A) - m_i(B)] - \left[F_i^{(1)}(A) - F_i^{(1)}(B)\right]. \tag{R.27}$$

Thus sufficiently strong differential interaction can reverse the precision decision induced by projection-wise scoring.

R.5 Pareto-dominated configurations

For expert i, define its utility–memory point set

$$\mathcal{P}_i = \left\{(m_i(A), F_i(A)) : A \in \mathcal{B}^{|\mathcal{J}|}\right\}. \tag{R.28}$$

A configuration A is dominated if some B satisfies $m_i(B) \leq m_i(A)$ and $F_i(B) \geq F_i(A)$, with at least one strict inequality.

Proposition R.6. Dominated-state elimination A dominated local configuration cannot be uniquely required by any globally optimal MCKP allocation, and there always exists an optimal allocation that excludes it.

Proof Replacing A by its dominating configuration B weakly decreases total memory and weakly increases total utility. Therefore any feasible allocation containing A admits a feasible replacement with no worse objective value. □

R.6 REDUCED-COST STABILITY

Suppose $\mathbf{A}^\star$ is a supported optimum of the Lagrangian relaxation at multiplier $\lambda^\star$. Define the reduced-cost gap

$$\gamma_{i,A} = [F_i(A_i^\star) - \lambda^\star m_i(A_i^\star)] - [F_i(A) - \lambda^\star m_i(A)]\,. \tag{R.31}$$

Let $\gamma_{\min} = \min_i \min_{A \neq A_i^\star} \gamma_{i,A}$.

Theorem R.7. Local reduced-cost robustness If $\|\widehat{F} - F\|_\infty \leq \varepsilon$ and

$$\gamma_{\min} > 2\varepsilon, \tag{R.34}$$

then the Lagrangian-optimal local state of every expert is preserved:

$$\widehat{A}_i^{\lambda^\star} = A_i^{\lambda^\star} \qquad \forall i. \tag{R.35}$$

This local stability criterion can be substantially tighter than the global $2|\mathcal{E}|\varepsilon$ condition.

S FIDELITY OF THE REFERENCE-CONTEXT SURROGATE

The MCKP optimizes the separable surrogate $\sum_i F_i(A_i)$, whereas actual model performance depends jointly on the global configuration. This section characterizes their discrepancy.

S.1 ALLOCATION-SPECIFIC EXPERT GAME

For a global configuration $\mathbf{A} = (A_1, \ldots, A_N)$, let $S_C^{\mathbf{A}}$ denote the state in which experts in C use their assigned configurations A_i, whereas experts outside C use the baseline configuration.

Define $\nu_{\mathbf{A}}(C) = V_{\mathcal{C}}(S_C^{\mathbf{A}})$. The allocation-specific Shapley marginal is

$$\psi_i^{\mathbf{A}} = \mathbb{E}_\pi\left[\nu_{\mathbf{A}}(C_\pi(i) \cup \{i\}) - \nu_{\mathbf{A}}(C_\pi(i))\right]. \tag{S.2}$$

By Shapley efficiency,

$$\sum_i \psi_i^{\mathbf{A}} = V_{\mathcal{C}}(S^{\mathbf{A}}) - V_{\mathcal{C}}(S_0). \tag{S.3}$$

S.2 CONTEXT-SUBSTITUTION SENSITIVITY

For arbitrary outer configuration $\mathbf{z}_{-i}$, define

$$\Delta_i(A_i \mid \mathbf{z}_{-i}) = V(A_i, \mathbf{z}_{-i}) - V(A_0, \mathbf{z}_{-i}). \tag{S.4}$$

For $k \neq i$, define

$$\beta_{ik}(A_i, A_k) = \sup_{\mathbf{z}} \left|\Delta_i(A_i \mid \mathbf{z}, z_k = A_{\text{ref}}) - \Delta_i(A_i \mid \mathbf{z}, z_k = A_k)\right|. \tag{S.5}$$

This quantity measures the sensitivity of expert i's configuration marginal to replacing the actual precision state of expert k with the reference state.

Theorem S.1. Reference-context surrogate fidelity For every feasible global allocation $\mathbf{A}$,

$$\left| \sum_i F_i(A_i) - \left[V_{\mathcal{C}}(S^{\mathbf{A}}) - V_{\mathcal{C}}(S_0) \right] \right| \leq \rho(\mathbf{A}), \tag{S.6}$$

where the context-sensitivity radius is

$$\rho(\mathbf{A}) = \frac{1}{2} \sum_i \sum_{k \neq i} \beta_{ik}(A_i, A_k). \tag{S.7}$$

Proof Compare the configuration marginal used by $F_i(A_i)$ with the allocation-specific marginal in Eq. (S.2) under the same permutation.

For every successor $k \in C_\pi(i)$, Config-Owen uses the reference state of k, whereas $\psi_i^{\mathbf{A}}$ uses its assigned state A_k. Replacing these states sequentially and applying Eq. (S.5),

$$|F_i(A_i) - \psi_i^{\mathbf{A}}| \leq \mathbb{E}_\pi \left[\sum_{k \in C_\pi(i)} \beta_{ik}(A_i, A_k) \right]. \tag{S.8}$$

Under a uniform random permutation, $\Pr[i \prec_\pi k] = \frac{1}{2}$. Hence

$$|F_i(A_i) - \psi_i^{\mathbf{A}}| \leq \frac{1}{2} \sum_{k \neq i} \beta_{ik}(A_i, A_k). \tag{S.10}$$

Summing over i and using Eq. (S.3) proves the result. □

Corollary S.2 If $\beta_{ik}(A_i, A_k) = 0$ for all expert pairs and admissible local states, then

$$\sum_i F_i(A_i) = V_{\mathcal{C}}(S^{\mathbf{A}}) - V_{\mathcal{C}}(S_0). \tag{S.11}$$

In that regime, the MCKP objective is an exact separable decomposition of the global configuration gain.

S.3 Global configuration regret

Define $J(\mathbf{A}) = V_{\mathcal{C}}(S^{\mathbf{A}}) - V_{\mathcal{C}}(S_0)$. Let $\mathbf{A}^\dagger = \arg\max_{\mathbf{A} \in \mathfrak{F}_B} J(\mathbf{A})$ be the truly optimal configuration under calibration utility.

Let $\rho_B = \sup_{\mathbf{A} \in \mathfrak{F}_B} \rho(\mathbf{A})$.

Theorem S.3. Global regret of the sampled Config-Owen allocation On the event $\|\widehat{F} - F\|_\infty \leq \varepsilon_M$,

$$J(\mathbf{A}^\dagger) - J(\widehat{\mathbf{A}}) \leq 2\rho_B + 2|\mathcal{E}|\varepsilon_M. \tag{S.15}$$

Proof For every feasible $\mathbf{A}$, $|J(\mathbf{A}) - \Phi(\mathbf{A})| \leq \rho_B$, and

$$|\Phi(\mathbf{A}) - \widehat{\Phi}(\mathbf{A})| \leq |\mathcal{E}|\varepsilon_M. \tag{S.17}$$

Therefore,

$$
\begin{aligned}
J(\mathbf{A}^\dagger) &\le \widehat{\Phi}(\mathbf{A}^\dagger) + \rho_B + |\mathcal{E}|\varepsilon_M \\
&\le \widehat{\Phi}(\widehat{\mathbf{A}}) + \rho_B + |\mathcal{E}|\varepsilon_M \\
&\le J(\widehat{\mathbf{A}}) + 2\rho_B + 2|\mathcal{E}|\varepsilon_M.
\end{aligned}
$$

□

Equation (S.15) separates final allocation error into reference-context approximation error and finite-sample utility error.

T EXTENSION TO MULTI-LEVEL PRECISION PRODUCT LATTICES

The main experiments instantiate two local precision levels, but the configuration-valued formulation extends naturally to arbitrary finite precision sets.

For projection j, let

$$
\mathcal{B}_j = \left\{ b_j^{(0)} \prec b_j^{(1)} \prec \cdots \prec b_j^{(K_j-1)} \right\}. \tag{T.1}
$$

The complete local precision space is

$$
\mathfrak{L}_i = \prod_{j \in \mathcal{J}} \mathcal{B}_j. \tag{T.2}
$$

For $\mathbf{a}, \mathbf{b} \in \mathfrak{L}_i$, define $\mathbf{a} \preceq \mathbf{b} \iff a_j \preceq b_j \quad \forall j$. Let $\mu_{\mathfrak{L}_i}$ denote the Möbius function on this product lattice.

The Möbius function factorizes:

$$
\mu_{\mathfrak{L}_i}(\mathbf{a}, \mathbf{b}) = \prod_{j \in \mathcal{J}} \mu_{\mathcal{B}_j}(a_j, b_j). \tag{T.4}
$$

Let $\mathbf{Z}_j$ be the zeta matrix of chain $\mathcal{B}_j$ and $\mathbf{M}_j = \mathbf{Z}_j^{-1}$ its Möbius matrix. Stack $v_i^\pi(\mathbf{a})$ and $\mathfrak{h}_i^\pi(\mathbf{a})$ over $\mathbf{a} \in \mathfrak{L}_i$ into $\mathbf{v}_i^\pi$ and $\mathbf{h}_i^\pi$. The product order induces the Kronecker factorization

$$
\begin{aligned}
\mathbf{Z}_{\mathfrak{L}_i} = \bigotimes_{j \in \mathcal{J}} \mathbf{Z}_j, \quad \mathbf{M}_{\mathfrak{L}_i} = \mathbf{Z}_{\mathfrak{L}_i}^{-1} = \bigotimes_{j \in \mathcal{J}} \mathbf{M}_j, \\
\mathbf{h}_i^\pi = \mathbf{M}_{\mathfrak{L}_i} \mathbf{v}_i^\pi, \qquad \mathbf{v}_i^\pi = \mathbf{Z}_{\mathfrak{L}_i} \mathbf{h}_i^\pi.
\end{aligned}
$$

Define generalized interaction coordinates

$$
\mathfrak{h}_i^\pi(\mathbf{a}) = \sum_{\mathbf{b} \preceq \mathbf{a}} \mu_{\mathfrak{L}_i}(\mathbf{b}, \mathbf{a}) v_i^\pi(\mathbf{b}). \tag{T.5}
$$

Theorem T.1. Product-lattice endpoint reconstruction For every candidate precision vector $\mathbf{a}$,

$$
v_i^\pi(\mathbf{a}) - v_i^\pi(\mathbf{a}_0) = \sum_{\mathbf{a}_0 \prec \mathbf{b} \preceq \mathbf{a}} \mathfrak{h}_i^\pi(\mathbf{b}), \tag{T.6}
$$

where $\mathbf{a}_0$ is the minimum local precision element.

Proof Möbius inversion gives

$$\begin{aligned} v_i^\pi(\mathbf{a}) &= \sum_{\mathbf{b}\preceq\mathbf{a}} \mathfrak{h}_i^\pi(\mathbf{b}), \\ v_i^\pi(\mathbf{a}_0) &= \mathfrak{h}_i^\pi(\mathbf{a}_0), \\ v_i^\pi(\mathbf{a}) - v_i^\pi(\mathbf{a}_0) &= \sum_{\mathbf{b}\preceq\mathbf{a}} \mathfrak{h}_i^\pi(\mathbf{b}) - \mathfrak{h}_i^\pi(\mathbf{a}_0) \\ &= \sum_{\mathbf{a}_0\prec\mathbf{b}\preceq\mathbf{a}} \mathfrak{h}_i^\pi(\mathbf{b}). \end{aligned} \tag{T.7}$$

This is Eq. (T.6). □

Corollary T.2 The binary setting used in the main experiments is the special case $K_j = 2 \qquad \forall j \in \mathcal{J}$, for which $\mathfrak{L}_i \cong 2^{\mathcal{J}}$. Therefore the Boolean-lattice endpoint equivalence in Appendix H is a special case of Eq. (T.6).

U CALIBRATION-TO-POPULATION TRANSFER AND END-TO-END GUARANTEES

The proposed utility is computed on a finite calibration set. We finally study the discrepancy between calibration utility and its population counterpart.

U.1 CALIBRATION MODEL

Let $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mathcal{D}$ denote calibration examples. Let $\ell(S; Z)$ be the example-normalized negative log-likelihood associated with precision state S.

The empirical and population losses, together with their inverse-perplexity utilities, are

$$\begin{aligned} \widehat{L}_{\mathcal{C}}(S) &= \frac{1}{n}\sum_{r=1}^{n} \ell(S; Z_r), \quad L_{\mathcal{D}}(S) = \mathbb{E}_{Z\sim\mathcal{D}}[\ell(S; Z)], \\ V_{\mathcal{C}}(S) &= \exp[-\widehat{L}_{\mathcal{C}}(S)], \qquad V_{\mathcal{D}}(S) = \exp[-L_{\mathcal{D}}(S)]. \end{aligned}$$

Let $\mathfrak{S}$ denote the finite family of admissible precision states relevant to the utility computation, with cardinality $N_{\mathfrak{S}} = |\mathfrak{S}|$.

Assumption U.1. Sub-Gaussian calibration loss For every $S \in \mathfrak{S}$, $\ell(S; Z) - L_{\mathcal{D}}(S)$ is σ_{cal}^2-sub-Gaussian:

$$\mathbb{E}\exp\left[\lambda\left(\ell(S; Z) - L_{\mathcal{D}}(S)\right)\right] \le \exp\left(\frac{\lambda^2\sigma_{\text{cal}}^2}{2}\right). \tag{U.6}$$

Theorem U.2. Uniform calibration concentration With probability at least $1 - \delta$,

$$\sup_{S\in\mathfrak{S}} \left|\widehat{L}_{\mathcal{C}}(S) - L_{\mathcal{D}}(S)\right| \le \sigma_{\text{cal}}\sqrt{\frac{2\log(2N_{\mathfrak{S}}/\delta)}{n}}. \tag{U.7}$$

Proof For any fixed S, sub-Gaussian concentration gives

$$\Pr\left(|\widehat{L}_{\mathcal{C}}(S) - L_{\mathcal{D}}(S)| > \eta\right) \le 2\exp\left(-\frac{n\eta^2}{2\sigma_{\text{cal}}^2}\right). \tag{U.8}$$

Applying a union bound over $\mathfrak{S}$ and solving for η proves Eq. (U.7). □

Set the uniform calibration radius to

$$\eta_{\mathrm{cal}} = \sigma_{\mathrm{cal}} \sqrt{\frac{2\log(2N_{\mathfrak{S}}/\delta)}{n}}. \tag{U.9}$$

For nonnegative losses, the mean-value theorem gives an intermediate ξ_S between $\widehat{L}_{\mathcal{C}}(S)$ and $L_{\mathcal{D}}(S)$ such that

$$\begin{aligned} |V_{\mathcal{C}}(S) - V_{\mathcal{D}}(S)| &= \left| e^{-\widehat{L}_{\mathcal{C}}(S)} - e^{-L_{\mathcal{D}}(S)} \right| \\ &= e^{-\xi_S} \left| \widehat{L}_{\mathcal{C}}(S) - L_{\mathcal{D}}(S) \right| \\ &\leq \left| \widehat{L}_{\mathcal{C}}(S) - L_{\mathcal{D}}(S) \right|, \end{aligned}$$

which makes the 1-Lipschitz transfer explicit.

Corollary U.3. Utility transfer With probability at least $1-\delta$,

$$\sup_{S \in \mathfrak{S}} |V_{\mathcal{C}}(S) - V_{\mathcal{D}}(S)| \leq \eta_{\mathrm{cal}}. \tag{U.10}$$

Since every configuration marginal is the difference between two state utilities,

$$\max_{i,A,\pi} |\Delta^{\pi}_{i,\mathcal{C}}(A) - \Delta^{\pi}_{i,\mathcal{D}}(A)| \leq 2\eta_{\mathrm{cal}}. \tag{U.11}$$

Averaging over permutations yields

$$\max_{i,A} |F_{i,\mathcal{C}}(A) - F_{i,\mathcal{D}}(A)| \leq 2\eta_{\mathrm{cal}}. \tag{U.12}$$

Proof Writing each configuration utility as an expectation of state differences gives

$$\begin{aligned} |F_{i,\mathcal{C}}(A) - F_{i,\mathcal{D}}(A)| &= \left| \mathbb{E}_{\pi}\left[\Delta^{\pi}_{i,\mathcal{C}}(A) - \Delta^{\pi}_{i,\mathcal{D}}(A)\right] \right| \\ &\leq \mathbb{E}_{\pi}\left[\left|\Delta^{\pi}_{i,\mathcal{C}}(A) - \Delta^{\pi}_{i,\mathcal{D}}(A)\right|\right] \leq 2\eta_{\mathrm{cal}}. \end{aligned}$$

□

U.2 Complete utility error decomposition

Define the Monte Carlo component by

$$\varepsilon_{\mathrm{MC}} = G_{\max} \sqrt{\frac{2}{M} \log \frac{2|\mathcal{E}||\mathcal{B}|^{|\mathcal{J}|}}{\delta}}. \tag{U.13}$$

Then the population utility-table error satisfies

$$\varepsilon_{\mathrm{util}} = 2\eta_{\mathrm{cal}} + \varepsilon_{\mathrm{prune}} + \varepsilon_{\mathrm{MC}}. \tag{U.14}$$

Therefore the total utility error obeys

$$\|\widehat{F} - F_{\mathcal{D}}\|_{\infty} \leq \varepsilon_{\mathrm{util}}. \tag{U.15}$$

Proof For every table entry, insert the suffix-pruned calibration utility and the exact calibration utility:

$$\begin{aligned}\left|\widehat{F}_i(A) - F_{i,\mathcal{D}}(A)\right| &\le \left|\widehat{F}_i(A) - F^{\mathrm{prune}}_{i,\mathcal{C}}(A)\right| + \left|F^{\mathrm{prune}}_{i,\mathcal{C}}(A) - F_{i,\mathcal{C}}(A)\right| \\ &\quad + |F_{i,\mathcal{C}}(A) - F_{i,\mathcal{D}}(A)| \\ &\le \varepsilon_{\mathrm{MC}} + \varepsilon_{\mathrm{prune}} + 2\eta_{\mathrm{cal}} = \varepsilon_{\mathrm{util}}.\end{aligned}$$

Taking the maximum over (i, A) proves Eq. (U.15). □

U.3 Population-level configuration objective

For $J_{\mathcal{D}}(\mathbf{A}) = V_{\mathcal{D}}(S^{\mathbf{A}}) - V_{\mathcal{D}}(S_0)$, let $\mathbf{A}^{\dagger}_{\mathcal{D}} = \arg\max_{\mathbf{A}\in\mathfrak{F}_B} J_{\mathcal{D}}(\mathbf{A})$, and define

$$\rho_{B,\mathcal{D}} = \sup_{\mathbf{A}\in\mathfrak{F}_B} \left|\sum_i F_{i,\mathcal{D}}(A_i) - J_{\mathcal{D}}(\mathbf{A})\right|. \tag{U.19}$$

Theorem U.4. End-to-end population guarantee With probability at least $1-\delta$,

$$J_{\mathcal{D}}(\mathbf{A}^{\dagger}_{\mathcal{D}}) - J_{\mathcal{D}}(\widehat{\mathbf{A}}) \le 2\rho_{B,\mathcal{D}} + 2|\mathcal{E}|\varepsilon_{\mathrm{util}}. \tag{U.20}$$

Proof For every $\mathbf{A} \in \mathfrak{F}_B$, the reference-context and utility-table errors combine as

$$\begin{aligned}\left|J_{\mathcal{D}}(\mathbf{A}) - \widehat{\Phi}(\mathbf{A})\right| &\le \left|J_{\mathcal{D}}(\mathbf{A}) - \sum_i F_{i,\mathcal{D}}(A_i)\right| + \sum_i \left|F_{i,\mathcal{D}}(A_i) - \widehat{F}_i(A_i)\right| \\ &\le \rho_{B,\mathcal{D}} + |\mathcal{E}|\varepsilon_{\mathrm{util}}.\end{aligned} \tag{U.21}$$

Applying this bound to both $\mathbf{A}^{\dagger}_{\mathcal{D}}$ and $\widehat{\mathbf{A}}$, and using the optimality of $\widehat{\mathbf{A}}$ for $\widehat{\Phi}$, gives

$$\begin{aligned}J_{\mathcal{D}}(\mathbf{A}^{\dagger}_{\mathcal{D}}) &\le \widehat{\Phi}(\mathbf{A}^{\dagger}_{\mathcal{D}}) + \rho_{B,\mathcal{D}} + |\mathcal{E}|\varepsilon_{\mathrm{util}} \\ &\le \widehat{\Phi}(\widehat{\mathbf{A}}) + \rho_{B,\mathcal{D}} + |\mathcal{E}|\varepsilon_{\mathrm{util}} \\ &\le J_{\mathcal{D}}(\widehat{\mathbf{A}}) + 2\rho_{B,\mathcal{D}} + 2|\mathcal{E}|\varepsilon_{\mathrm{util}}.\end{aligned} \tag{U.22}$$

□

U.4 Exact population-optimal recovery

Define the population allocation gap

$$\Gamma_{\mathrm{pop}} = J_{\mathcal{D}}(\mathbf{A}^{\dagger}_{\mathcal{D}}) - \max_{\substack{\mathbf{A}\in\mathfrak{F}_B \\ \mathbf{A}\neq\mathbf{A}^{\dagger}_{\mathcal{D}}}} J_{\mathcal{D}}(\mathbf{A}). \tag{U.23}$$

Corollary U.5 If

$$\Gamma_{\mathrm{pop}} > 2\rho_{B,\mathcal{D}} + 2|\mathcal{E}|\varepsilon_{\mathrm{util}}, \tag{U.24}$$

then $\widehat{\mathbf{A}} = \mathbf{A}^{\dagger}_{\mathcal{D}}$.

Proof If $\widehat{\mathbf{A}} \neq \mathbf{A}^{\dagger}_{\mathcal{D}}$, the population gap and Theorem U.4 imply

$$\Gamma_{\mathrm{pop}} \le J_{\mathcal{D}}(\mathbf{A}^{\dagger}_{\mathcal{D}}) - J_{\mathcal{D}}(\widehat{\mathbf{A}}) \le 2\rho_{B,\mathcal{D}} + 2|\mathcal{E}|\varepsilon_{\mathrm{util}} < \Gamma_{\mathrm{pop}}, \tag{U.25}$$

a contradiction. □

V THEORETICAL INTERPRETATION OF THE CONTROLLED ABLATIONS

Local-Calib-MCKP evaluates each joint local configuration directly under a fixed outer-expert context. Config-Owen evaluates the same local state across sampled expert-ordering contexts and averages the resulting utilities.

Variant	Joint local state	Outer context	Omitted structure
Expert-only	No	Yes	Configuration resolution
Projection-Owen	Partial	Yes	Higher-order projection terms
Local-Calib-MCKP	Yes	Fixed	Contextual interactions
Config-Owen	Yes	Averaged	None of the two axes

Table 8: Structural interpretation of the controlled attribution variants.

The two structural omissions admit exact formula-level diagnostics. For a fixed outer context, the local non-additive residual is the orthogonal energy outside the additive Walsh subspace; for a fixed context C_0, the contextual residual is the omitted part of the hierarchical interaction tensor:

$$\begin{aligned}\mathcal{R}_{\mathrm{local}}(v) &:= \inf_{g\in\mathfrak{P}_1}\mathbb{E}[(v(X)-g(X))^2] = \sum_{|R|\geq 2}\widehat{v}(R)^2 = \eta(v)\,\mathrm{Var}[v(X)],\\ \mathcal{R}_{\mathrm{ctx},i}(A;C_0) &:= F_i(A) - F_i^{C_0}(A)\\ &= \sum_{\emptyset\neq R\subseteq A}\sum_{T\subseteq\mathcal{E}_{-i}}\left(\frac{1}{|T|+1} - \mathbb{I}[T\subseteq C_0]\right)\Gamma_i(R,T).\end{aligned} \tag{V.1}$$

The same local residual controls the displacement of the budget switching frontier. For any pair with $m_i(A)\neq m_i(B)$,

$$\begin{aligned}\Delta\lambda_{i;A,B} &:= \lambda_{i;A,B} - \lambda^{(1)}_{i;A,B} = \frac{I_i(A)-I_i(B)}{m_i(A)-m_i(B)},\\ |\Delta\lambda_{i;A,B}| &\leq \frac{|I_i(A)|+|I_i(B)|}{|m_i(A)-m_i(B)|}.\end{aligned} \tag{V.2}$$

Finally, substituting the complete utility-error decomposition into the population guarantee separates the three finite-data contributions:

$$\begin{aligned}J_{\mathcal{D}}(\mathbf{A}^{\dagger}_{\mathcal{D}}) - J_{\mathcal{D}}(\widehat{\mathbf{A}}) &\leq 2\rho_{B,\mathcal{D}} + 2|\mathcal{E}|\varepsilon_{\mathrm{util}}\\ &= 2\rho_{B,\mathcal{D}} + 4|\mathcal{E}|\eta_{\mathrm{cal}} + 2|\mathcal{E}|\varepsilon_{\mathrm{prune}}\\ &\quad + 2|\mathcal{E}|G_{\max}\sqrt{\frac{2}{M}\log\frac{2|\mathcal{E}||\mathcal{B}|^{|\mathcal{J}|}}{\delta}}.\end{aligned} \tag{V.3}$$

Thus the ablations isolate configuration resolution, local interaction mass, and outer-context averaging, while the final bound keeps approximation, calibration, suffix-pruning, and sampling effects explicit.