MDMIXQ: MODALITY-DECOUPLED MIXED-PRECISION QUANTIZATION FOR MULTIMODAL MOE COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training quantization is widely adopted for reducing the memory footprint of large-scale Mixture-of-Experts (MoE) models, and expert-level mixed precision further improves the accuracy–compression trade-off by allocating heterogeneous bit-widths across experts. In multimodal MoE settings, however, we find that existing mixed-precision pipelines overlook inter- and intra-modality routing non-uniformities that skew expert bit allocation and degrade accuracy under compression. To address these failure modes, we propose **MDMixQ**, a mixed-precision quantization framework comprising three targeted components: Modality-Decoupled Calibration (MDC), which isolates a clean text-routing signal from multimodal inputs; Salience-Guided Filtering and Rectification, which retains only confident text-side routing events and rectifies sign-inconsistent logits; and Dynamic Mixed-Precision Allocation, which ranks experts by the refined salience scores under a global bit budget and applies modality-aware activation quantization during MoE inference. At a 2.52-bit average expert budget, the best MDMixQ configurations retain 96.29%, 96.05%, and 95.33% of FP16 performance on Qwen3-VL-30B-A3B, Kimi-VL-16B-A3B, and Qwen3-Omni-30B-A3B, respectively, outperforming prior expert-priority methods across two model families. The framework also transfers to the MXFP format with 99.72% performance retained on Qwen3-VL-30B-A3B, confirming compatibility with emerging hardware-oriented quantization standards.

1 INTRODUCTION

Multimodal foundation models have achieved remarkable progress, yet their deployment cost scales rapidly with model capacity. Mixture-of-Experts (MoE) architectures mitigate this via sparse activation (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022; Du et al., 2022; Zoph et al., 2022), but still require all expert parameters to reside in memory, creating a significant bottleneck for memory footprint and bandwidth (Frantar & Alistarh, 2023; Kim et al., 2023). This challenge is particularly acute for multimodal MoE models, where complex expert specialization and cross-modal routing amplify system overhead.

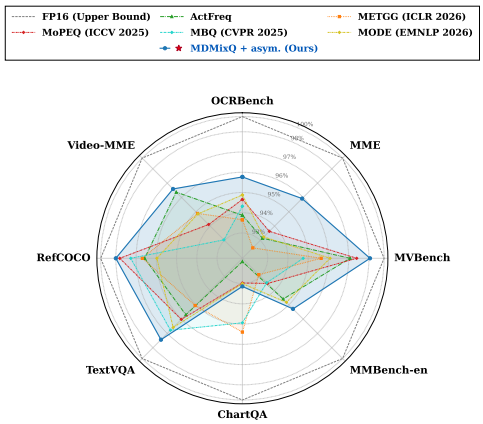


Figure 1: Per-benchmark retention on Qwen3-VL-30B-A3B at a 2.52-bit average expert budget, normalized to the FP16 reference.

Post-training quantization (PTQ) is a natural remedy (Yao et al., 2022; Frantar et al., 2022; Xiao et al., 2023; Lin et al., 2024; Shao et al., 2024). Recent modality-aware methods have shown that visual and text tokens exhibit different quantization sensitivities: MBQ reweights reconstruction according to modality-specific sensitivity, while SPEED-Q applies staged sensitivity adaptation and distillation-enhanced low-

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.

bit optimization to vision–language models (Li et al., 2025; Guo et al., 2026). This principle has also been extended to multimodal MoE models. VEQ incorporates modality and token–expert affinity into reconstruction-based calibration, and MODE decomposes expert statistics by modality and combines them with quantization sensitivity for expert-level mixed-precision allocation (Qin et al., 2026; Chen et al., 2026). In parallel, language-only MoE quantizers exploit heterogeneous expert importance through mixed-precision allocation, but their importance signals remain modality-agnostic (Huang et al., 2025; Duanmu et al., 2025).

Despite this progress, separating modalities alone does not guarantee a reliable expert-importance signal. We identify two residual failure modes in routing-based calibration. First, even within one modality, many tokens induce flat routing distributions whose weak and potentially unstable expert preferences accumulate as background noise. Second, raw router logits are signed: directly summing them can suppress frequently selected experts, whereas bounded normalizations may collapse the discriminative range of strongly preferred experts. These effects compound the familiar cross-modal token-volume imbalance, so expert ranking can remain biased even after the calibration stream is partitioned by modality. The resulting challenge is therefore not only to identify which modality produced a routing trace, but also to determine whether each routing event provides reliable and numerically consistent evidence of expert importance.

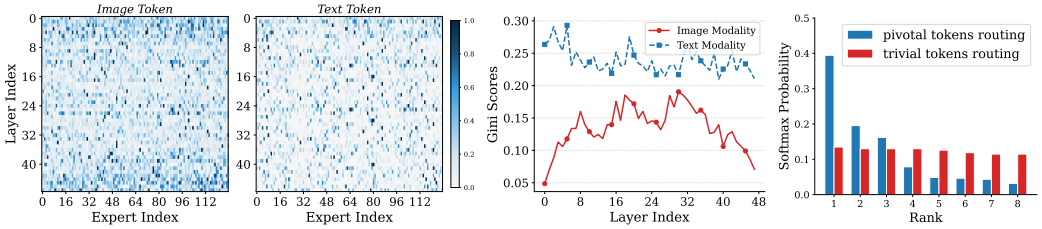
To address these challenges, we propose **MDMixQ**, a modality-decoupled mixed-precision quantization framework that constructs expert importance directly from routing structure. At a 2.52-bit average expert-weight budget, the best MDMixQ configurations retain 96.29%, 96.05%, and 95.33% of FP16 performance on Qwen3-VL-30B-A3B, Kimi-VL-16B-A3B, and Qwen3-Omni-30B-A3B, respectively, and transfer to MXFP quantization with 99.72% retention.

The main contributions of this work are summarized as follows:

- We identify two overlooked sources of bias in routing-based expert-importance estimation for multimodal MoEs. Beyond cross-modal token-volume imbalance, ambiguous routing events within each modality inject background noise, while direct accumulation of signed router logits can cancel otherwise frequent and important expert selections. This analysis shows why merely separating modality streams is insufficient for robust mixed-precision allocation.
- We propose a lightweight, routing-only salience estimator whose modality-decoupled calibration isolates a clean text-routing signal from multimodal inputs, adaptively removes low-confidence token–expert events, and rectifies signed logits while preserving the discriminative range of strong positive responses. The resulting text-anchored salience supports expert-level bit allocation using only calibration-time router outputs, without gradients, per-expert quantization-error profiling, or integer programming.
- We extend modality asymmetry from expert-weight allocation to activation precision inside MoE layers and validate the resulting framework on three multimodal MoE backbones spanning vision–language and audio–visual–language tasks, as well as under MXFP quantization, demonstrating effectiveness across model families, modalities, and numerical formats.

2 Related Work

Mixed-precision MoE quantization. Expert-wise mixed precision exploits the non-uniform importance and quantization sensitivity of sparse experts. **ActFreq** (Muzio et al., 2024) assigns higher precision to frequently activated experts, but conflates routing load with quantization sensitivity. **MoPEQ** (Chitty-Venkata et al., 2025) uses a Hessian-trace approximation to estimate expert sensitivity and clusters experts into precision groups, whereas **METGG** (Chowdhury et al., 2026) ranks experts using changes in router norms together with intra-expert weight variation. **MC-MoE** (Huang et al., 2025) combines pre-loading mixed-precision quantization, whose expert bit allocation balances activation reconstruction error, routing score, and activation frequency through linear programming, with online dynamic expert pruning. **MxMoE** (Duanmu et al., 2025) co-designs accuracy and efficiency by searching mixed-precision configurations from linear-block sensitivity, expert activation dynamics, and hardware constraints, and generates mixed-precision GroupGEMM kernels. These approaches primarily study language-only MoE models and therefore do not explicitly account for modality-dependent routing statistics.



(a) Act Frequency (b) Gini Coefficient (c) Softmax Distribution

Figure 2: Cross-modal activation patterns on Qwen3-VL-30B-A3B. (a) Expert activation frequency heatmaps for image and text tokens across all MoE layers. Image tokens exhibit a dispersed activation pattern, whereas text tokens concentrate their activations on a smaller subset of experts. (b) Gini coefficient comparison between image and text tokens. Text tokens consistently yield higher Gini coefficients than image tokens across layers, confirming that text inputs rely on a narrower set of experts while image inputs engage a broader expert distribution. (c) Softmax probability distributions of routing scores, sorted by rank. Pivotal tokens exhibit a concentrated distribution with a dominant top-ranked expert, while trivial tokens distribute probability near-uniformly across all experts.

Modality-aware multimodal quantization. Recent work has shown that directly transferring language-model quantization recipes to multimodal models can yield modality-imbalanced calibration. **MBQ** (Li et al., 2025) measures the different gradient sensitivities of visual and language tokens and reweights the reconstruction objective during post-training quantization, preventing the more numerous but less sensitive visual tokens from dominating calibration. **SPEED-Q** (Guo et al., 2026) targets end-to-end low-bit deployment of compact VLMs by quantizing the vision and language components through staged sensitivity adaptation and distillation-enhanced optimization. Extending modality-aware quantization to sparse multimodal architectures, **VEQ** (Qin et al., 2026) weights expert reconstruction error using modality-aware activation frequency and augments Hessian-based calibration with token–expert affinity and modality information. **MODE** (Chen et al., 2026) further decomposes expert selection frequency by modality, filters redundant visual tokens, complements the resulting frequencies with modality-specific quantization sensitivity, and solves an integer linear program for expert-level bit allocation. Collectively, these methods demonstrate that modality identity and modality-specific sensitivity should be explicitly preserved during multimodal quantization and calibration.

How MDMixQ differs. MDMixQ differs primarily in how it constructs expert importance. Unlike VEQ's reconstruction- and Hessian-based calibration and MODE's combination of modality-decomposed frequency, quantization sensitivity, and integer programming, MDMixQ uses a lightweight routing-only signal: it filters low-confidence routing events, rectifies signed logits, and ranks experts by the resulting text-anchored salience. This avoids gradients and per-expert sensitivity profiling, while extending the same modality-asymmetry principle to activation precision inside MoE layers.

3 Motivation

3.1 Inter-Modality Asymmetry in Expert Activation

Multimodal workloads involve distinct token modalities (e.g., text and non-text patches) that exhibit significant differences in cardinality and redundancy. Empirically, non-text tokens are more numerous but highly redundant, often capturing overlapping low-level structures. In contrast, text tokens are sparser and carry task-relevant semantic content critical for reasoning.

As illustrated in Figure 2, these modalities induce divergent expert activation patterns. Image tokens exhibit dispersed activation across the expert pool, while text tokens concentrate on a narrower subset. To quantify this, we employ the Gini coefficient:

$$G = \frac{\sum_{i=1}^{N}(2i - N - 1)f_i}{N\sum_{i=1}^{N} f_i}, \tag{1}$$

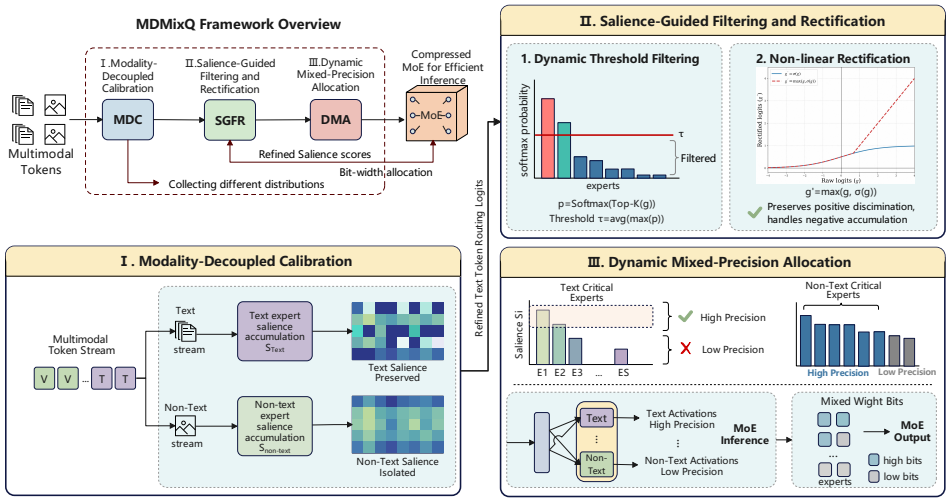


Figure 3: Overview of MDMixQ. The framework isolates a clean text-routing signal from multimodal calibration inputs, refines text-side routing logits through salience-guided filtering and rectification, and assigns expert-level weight precision from the resulting importance scores. We additionally evaluate modality-aware activation quantization as an optional extension.

where N is the number of experts and $\{f_i\}$ are sorted cumulative activation frequencies. A value closer to 1 implies higher concentration. Our measurements confirm that text tokens consistently yield higher Gini coefficients across layers, indicating a more specialized expert usage compared to the diffuse non-text activation.

This inter-modality asymmetry poses a challenge for fused-stream calibration. Importance statistics derived from combined streams become biased toward the modality with higher token volume or aggregate routed mass, rather than the one most informative for reasoning. This results in a systematic bias where quantization budgets favor experts processing redundant perceptual patterns, effectively diluting the protection afforded to text-specialized experts essential for multimodal reasoning.

3.2 Intra-Modality Heterogeneity in Token-Level Routing

Even within a single modality, expert activation is highly token-conditional. As illustrated in Figure 2c, a small subset of pivotal tokens typically triggers routing distributions with high confidence and strong top-expert dominance, which are often the tokens most aligned with salient entities, relations, or discourse cues. In contrast, the bulk of trivial tokens exhibit comparatively flat routing: padding tokens, template-like text, and other low-information units spread mass across experts with small margins, reflecting weak preference rather than decisive specialization. Individually, such trivial activations are numerous; in aggregate, their near-uniform contributions behave like additive noise that swamps the sparse, high-contrast signal from the pivotal tokens when experts are scored by naive summation or frequency counting. Consequently, coarse calibration that does not separate pivotal from trivial tokens within each modality risks misranking experts, masking which units truly pin down reasoning structure behind a thick background of uninformative routing events.

4 Methodology

The non-uniformities identified in Section 3 demonstrate that multimodal MoE compression fails under fused calibration or naive routing aggregation. To address this, we propose MDMixQ, a mixed-precision quantization framework comprising three components (Figure 3): (1) Modality-Decoupled Calibration (MDC), which isolates a clean text-routing signal from multimodal inputs; (2) Salience-Guided Filtering and Rectification (SGFR), which suppresses routing noise and corrects

sign-inconsistent logits; and (3) Dynamic Mixed-Precision Allocation (DMA), which maps refined salience scores to heterogeneous expert-weight bit-widths.

4.1 Modality-Decoupled Calibration

Fused-stream calibration conflates modalities with distinct routing behaviors. High-volume, diffusely routed non-text tokens often dominate aggregate importance scores, obscuring the sharper, text-centric specialization (Figure 2a). Modality-Decoupled Calibration (MDC) counters this asymmetry by isolating a clean text-routing signal from multimodal inputs while maintaining separate routing and activation accumulators for text and non-text tokens.

For a MoE layer with N experts, input tokens are split into $\mathcal{T} = \{\mathcal{T}_{\text{text}}, \mathcal{T}_{\text{non-text}}\}$. During calibration, each expert E_i updates two accumulators: text salience $\mathbf{S}_{\text{text},i}$ and non-text salience $\mathbf{S}_{\text{non-text},i}$. This separation prevents dispersed non-text routing from contaminating the clean text-side signal used by the main configuration for weight-budget allocation. Reserving a small Top-K set of non-text-critical experts is evaluated as an optional ablation. This decoupled strategy consistently outperforms fused-stream baselines.

4.2 Salience-Guided Filtering and Rectification

Naive logit accumulation mixes decisive specialization signals with background noise, and suffers from sign inconsistency when negative logits penalize frequently selected experts. To refine the text stream for expert ranking, we apply a two-stage procedure:

Dynamic threshold filtering. We isolate token–expert pairs that demonstrate confident expert preference. Let $\mathcal{I}_t^l$ be the indices of the top-K_r router logits for token t at layer l ($K_r = 8$ in our experiments). We apply Softmax only over these selected logits and scatter the resulting probabilities back to the full expert-index space:

$$p_{t,i}^l = \begin{cases} \dfrac{\exp(g_{t,i}^l)}{\sum_{j \in \mathcal{I}_t^l} \exp(g_{t,j}^l)}, & i \in \mathcal{I}_t^l, \\ 0, & i \notin \mathcal{I}_t^l. \end{cases} \tag{2}$$

For a modality-specific calibration-token set $\mathcal{T}_m$, where $m \in \{\text{text}, \text{non-text}\}$, the layerwise confidence threshold is

$$\tau_m^l = \frac{1}{|\mathcal{T}_m|} \sum_{t \in \mathcal{T}_m} \max_i p_{t,i}^l. \tag{3}$$

Only pairs that both belong to the router top-K_r set and satisfy $p_{t,i}^l \geq \tau_m^l$ contribute to importance accumulation. The indicator is applied outside the rectification function, so every filtered pair contributes exactly zero. This filters out padding and low-information routing events that would otherwise inflate frequency-based scores.

Non-linear rectification. To establish a monotonic, positive-sensitive importance metric and prevent negative accumulation, we rectify raw logits for salient tokens:

$$\mathbf{g}' = \max(\mathbf{g},\ \sigma(\mathbf{g})), \tag{4}$$

where $\sigma(\cdot)$ is the Sigmoid function. As shown in Figure 3 II, this maps negative logits to $(0, 0.5)$ while applying an identity map to large positive values, preserving linear discriminability for high activations without universal saturation.

4.3 Dynamic Mixed-Precision Allocation

Post-calibration, each expert E_i receives separate text and non-text structural salience scores. In the main MDMixQ configuration, we globally rank all layer–expert pairs by text salience and set $K_{\text{nt}} = 0$. Under the 2.52-bit average budget with precision set $\{2, 3\}$ used in our experiments, the 3-bit tier contains the highest-ranked pairs and all remaining experts use 2-bit precision. Top-K_{nt} non-text reservation is an optional extension evaluated only in the ablation study: reserved non-text pairs enter the fixed-size 3-bit tier first, after which the remaining slots are filled by text-salience

Algorithm 1 MDMixQ Framework

Require: Model $\mathcal{M}$; token sets $\mathcal{T}_{\text{text}}, \mathcal{T}_{\text{nt}}$; router K_r; budget $\bar{B}$; optional K_{nt} and A_{asym}
Ensure: Quantized model $\mathcal{M}_Q$
for $l = 1, \ldots, L$ **do**
  for $m \in \{\text{text}, \text{nt}\}$ **do**
    Collect logits $g^l_{m,t,i}$ and Top-K_r index sets $\mathcal{I}^l_{m,t}$
    Scatter Top-K_r Softmax probabilities to the full expert space using Equation (2)
    Compute τ^l_m using Equation (3)
    $S^l_{m,i} \leftarrow \sum_{t \in \mathcal{T}_m} \mathbf{1}[i \in \mathcal{I}^l_{m,t} \wedge p^l_{m,t,i} \geq \tau^l_m] \max(g^l_{m,t,i}, \sigma(g^l_{m,t,i}))$
  end for
end for
Globally rank layer–expert pairs by $S^l_{\text{text},i}$
$H \leftarrow \lfloor (\bar{B} - 2) \sum_l N_l \rfloor$ for equal-sized experts and $\{2, 3\}$-bit precision
Initialize $\mathcal{E}_3$ with the top-K_{nt} pairs by $S^l_{\text{nt},i}$; use $K_{\text{nt}} = 0$ for the main configuration
Fill $\mathcal{E}_3$ to size H by text-salience rank; set its experts to 3-bit and all others to 2-bit
if A_{asym} **then**
  In MoE layers only, use 4-bit non-text and 8-bit text activations
else
  Keep MoE activations at 16-bit
end if
return Quantized model $\mathcal{M}_Q$

ranking. Thus, enabling the reserve does not change the average expert-weight budget. By eliminating cross-modal bias and filtering intra-modal noise, text-side salience serves as a reliable proxy for the default bit assignment.

Beyond weight compression, we evaluate an optional modality-decoupled activation-quantization extension, reported as *MDMixQ + asym. act.* Because text activations contain dense symbolic content essential for reasoning, whereas non-text activations exhibit higher spatial redundancy, this variant uses 4-bit non-text and 8-bit text activations exclusively inside MoE layers. The base *MDMixQ* configuration keeps MoE activations at 16-bit. The full procedure and its optional switches are detailed in Algorithm 1.

5 Experiments

5.1 Experimental Setup

Models and benchmarks. We evaluate MDMixQ on three multimodal MoE backbones: **Qwen3-VL-30B-A3B** (Bai et al., 2025), **Kimi-VL-16B-A3B** (Team et al., 2025), and **Qwen3-Omni-30B-A3B** (Xu et al., 2025). For the two vision–language backbones, we evaluate image understanding on OCRBench (Liu et al., 2024b), MME (Fu et al., 2026), MMB-en (Liu et al., 2024a), ChartQA (Masry et al., 2022), and TextVQA (Singh et al., 2019); we additionally use RefCOCO (Yu et al., 2016) for Qwen3-VL and InfoVQA (Mathew et al., 2022) for Kimi-VL. Qwen3-VL is further evaluated for video understanding on MVBench (Li et al., 2024) and Video-MME (Fu et al., 2025). The corresponding Kimi-VL video scores are unavailable because the vLLM backend used in our evaluation does not support video inputs for this model. For Qwen3-Omni-30B-A3B, we evaluate image understanding on ChartQA and MMB-en, audio understanding on LibriSpeech ASR (Panayotov et al., 2015), SpeechQA (Nachmani et al., 2024; Berant et al., 2013), and OmniBench (Li et al., 2026), and video understanding on MVBench and Video-MME. Detailed descriptions of all benchmarks are deferred to Appendix A.8.

Baselines and protocol. We compare three groups of methods. (i) Reference models: FP16 and uniform W4A16. (ii) Expert-priority baselines: **ActFreq**, **MoPEQ** (Chitty-Venkata et al., 2025) (Hessian-based importance), **METGG** (Chowdhury et al., 2026) (router-norm-change-based importance), **MODE**, and **MBQ**. (iii) Our framework variants: **MDMixQ** and **MDMixQ + asym. act.** (with cross-modal asymmetric activation quantization). Within each backbone and calibration re-sample, all methods use the same 50 examples, weight quantizer, target average expert bits, and evaluation pipeline; they differ in expert-importance estimation and bit allocation.

Table 1: Main results on three multimodal MoE models. Panel (a): Qwen3-VL-30B-A3B; panel (b): Kimi-VL-16B-A3B; panel (c): Qwen3-Omni-30B-A3B. "Avg. Bit" is the average expert-weight precision. "Retain (%)" summarizes normalized performance relative to the corresponding FP16 reference model (Section A.8). **NT** and **T** in Act. Bits denote **non-text tokens** and **text tokens**, respectively, and our asymmetric activation policy (**asym.**) is **only applied to MoE layers**. Bold and underline indicate the best and second-best results within the 2.52-bit comparison group {ActFreq, MoPEQ, METGG, MODE, MBQ, MDMixQ, MDMixQ + asym.}; FP16 and uniform W4A16 are reported as higher-budget references and are not eligible for bold/underline. All quantized runs use 4-bit attention projections unless otherwise noted. MDMixQ results are averaged over three calibration-set re-samples.

Qwen3-VL-30B-A3B

			Image Understanding						Video Understanding		
Method	Avg. Bit	Act. Bits	OCRBench	MME	MMB-en	ChartQA	TextVQA	RefCOCO	MVBench	Video-MME	Retain (%)
FP16	16.00	16	898	2426	87.37	85.12	84.65	96.14	69.95	67.22	100.00
W4A16	4.00	16	888	2426	85.31	83.32	82.98	95.91	69.30	64.85	98.47
ActFreq	2.52	16	843	2260	82.65	76.20	81.01	92.85	67.90	<u>64.78</u>	94.61
MoPEQ	2.52	16	850	2271	81.70	79.08	81.28	94.03	68.13	63.26	95.00
METGG	2.52	16	841	2215	81.62	**81.20**	80.46	93.27	66.90	63.78	94.54
MODE	2.52	16	852	2262	82.85	79.16	81.76	92.27	67.21	63.76	94.93
MBQ	2.52	16	847	2158	81.65	<u>80.81</u>	81.92	93.51	66.28	62.54	94.19
MDMixQ	2.52	16	**864**	<u>2298</u>	<u>82.90</u>	79.52	**82.48**	**94.54**	**68.75**	64.22	<u>96.11</u>
MDMixQ + asym.	2.52	NT4/T8	<u>860</u>	**2327**	**83.25**	79.28	**82.48**	<u>94.22</u>	<u>68.65</u>	**64.93**	**96.29**

Kimi-VL-16B-A3B

			Image Understanding						Video Understanding		
Method	Avg. Bit	Act. Bits	OCRBench	MME	MMB-en	ChartQA	TextVQA	InfoVQA	MVBench	Video-MME	Retain (%)
FP16	16.00	16	880	2177	79.47	86.68	92.57	83.71	–	–	100.00
W4A16	4.00	16	862	2101	76.74	84.16	90.82	81.25	–	–	97.22
ActFreq	2.52	16	822	2066	74.14	82.88	86.07	81.45	–	–	94.58
MoPEQ	2.52	16	827	2054	75.17	82.24	<u>87.51</u>	80.94	–	–	94.84
METGG	2.52	16	837	2058	75.52	83.96	87.27	79.82	–	–	95.19
MODE	2.52	16	823	**2081**	73.88	81.04	87.48	81.56	–	–	94.58
MBQ	2.52	16	827	<u>2071</u>	74.28	82.25	86.74	<u>82.15</u>	–	–	94.88
MDMixQ	2.52	16	**842**	2031	**76.72**	<u>84.12</u>	**88.10**	**82.51**	–	–	**96.05**
MDMixQ + asym.	2.52	NT4/T8	<u>840</u>	2065	<u>76.48</u>	**84.15**	87.25	81.24	–	–	<u>95.82</u>

Qwen3-Omni-30B-A3B

			Image Understanding		Audio Understanding			Video Understanding		
Method	Avg. Bit	Act. Bits	ChartQA	MMB-en	LibriSpeechASR↓	SpeechQA	OmniBench	MVBench	Video-MME	Retain (%)
FP16	16.00	16	83.00	85.00	1.64	54.53	56.46	66.40	63.96	100.00
W4A16	4.00	16	82.24	82.65	1.69	51.92	55.40	65.70	63.30	98.22
ActFreq	2.52	16	76.88	<u>81.19</u>	2.00	45.23	54.07	63.73	60.89	93.95
MoPEQ	2.52	16	77.44	80.33	1.98	40.65	53.98	64.47	60.96	92.86
METGG	2.52	16	78.24	79.38	2.04	42.08	**54.60**	63.10	60.00	92.85
MODE	2.52	16	79.25	80.12	2.12	43.53	54.12	63.61	60.21	93.55
MBQ	2.52	16	78.21	77.53	2.11	44.29	53.14	63.15	60.87	92.94
MDMixQ	2.52	16	<u>80.16</u>	**81.62**	**1.72**	<u>46.31</u>	54.16	<u>64.65</u>	<u>61.07</u>	<u>95.17</u>
MDMixQ + asym.	2.52	NT4/T8	**80.80**	80.24	<u>1.79</u>	**46.46**	<u>54.25</u>	**65.47**	**61.30**	**95.33**

Quantization configuration and calibration. Unless otherwise specified, attention QKV/O projections are quantized to 4-bit weights, while the audio encoder and ViT are kept in full precision, since the text pathway dominates multimodal post-training quantization accuracy (Zhang et al., 2026). To enable a controlled comparison, ActFreq, METGG, and our MDMixQ all use a $\{2, 3\}$-bit configuration that meets the same 2.52-bit average; MoPEQ retains its native $\{2, 3, 4\}$-bit allocation as specified in its original paper, which provides a strictly more expressive bit-allocation space. We verified that, for the other three methods, the $\{2, 3\}$-bit configuration outperforms their respective $\{2, 3, 4\}$-bit variant at the same average budget (Table 3b); we therefore report each method under its better-performing setting. Modality-decoupled calibration uses only the **COCO** dataset Lin et al. (2014) for the two VL models and **COCO-Audio** for the Omni model. These calibration sources

Table 2: Calibration and allocation ablations at 2.52 average expert bits. All inputs are multimodal; "img$x$k+txt$y$k" counts image and text tokens used for routing statistics ($k = 10^3$). Panel (b) varies the number of high-precision non-text experts.

(a) Calibration token composition.

Calibration split	OCRBench	MME	MMB-en	ChartQA	Video-MME	Retain (%)
img10k + txt0	856	2284	82.90	78.48	65.11	95.50
img5k + txt5k	866	2253	82.99	76.64	64.22	95.05
img2k + txt8k	868	2318	82.99	77.76	64.07	95.72
img0 + txt10k	861	2320	83.33	78.88	64.81	96.14

(b) Top-K non-text expert preservation.

Variant	OCRBench	MMB-en	ChartQA	TextVQA	Video-MME	Retain (%)
W/O NT	864	82.90	79.52	82.48	64.22	96.11
W/ NT Top-32	849	82.65	77.80	81.53	64.00	95.34
W/ NT Top-16	855	83.42	77.88	81.89	64.15	95.63
W/ NT Top-8	866	83.16	78.72	82.10	64.93	96.16

are unrelated to the evaluation benchmarks, and only 50 randomly sampled data points are used for each model. Quantizer details and analytical weight-storage estimates appear in Appendix A.1 and Appendix A.2.

5.2 Main Results

We summarize the comparison in Table 1 through two findings.

Higher retention at the same bit budget. At the matched 2.52-bit expert budget, MDMixQ achieves the best aggregate retention among the compared mixed-precision methods on Qwen3-VL-30B-A3B, Kimi-VL-16B-A3B, and Qwen3-Omni-30B-A3B. The consistent advantage across three different multimodal MoE backbones and across image, video, and audio benchmarks shows that the proposed modality-decoupled importance estimation generalizes beyond a particular model family or modality. Because all methods use the same calibration budget, average bit width, and evaluation pipeline within each backbone, the improvement can be attributed to the proposed expert-importance estimation and mixed-precision allocation strategy.

Cross-modal asymmetric activation reduces activation bits without additional performance degradation. The asymmetric policy uses 4-bit visual or speech activations and 8-bit text activations inside MoE layers, reducing activation precision without increasing the expert weight budget. Across all three backbones, MDMixQ with asymmetric activation remains comparable to the standard MDMixQ configuration and preserves its advantage over the competing methods. This supports using higher precision for the reasoning-critical text pathway than for non-text tokens, consistent with the asymmetry identified in Section 3.

5.3 Ablation Studies

We ablate MDMixQ along three axes: (1) modality-decoupled calibration and expert allocation, (2) cross-modal asymmetric activation quantization, and (3) budget sensitivity. Representative benchmarks are shown below; full per-benchmark numbers are deferred to the Appendix in Table 6. Generalization under MXFP quantization is reported separately in Appendix A.6.

Modality-decoupled calibration and expert allocation. This axis covers two tightly coupled components: the calibration token composition and the resulting Top-K non-text expert preservation.

Panel (a) of Table 2 varies which routing statistics are accumulated from the same multimodal COCO inputs at a fixed 2.52-bit expert budget. In "img0 + txt10k", images remain in the inputs, but only text-token routing contributes to expert scores. This setting achieves the highest Retain (96.14%), versus 95.05% for img5k + txt5k and 95.72% for img2k + txt8k. Thus, text-side routing provides a stronger allocation signal in this setting; the experiment does not compare against calibration on text-only inputs.

Panel (b) then ablates how many non-text-specialised experts to reserve at the higher bit-width beyond the text-salience-driven allocation. "W/O NT" disables non-text reservation entirely; "W/ NT Top-K" additionally reserves the K experts with the largest non-text accumulated salience. The trend follows from the inter-modality asymmetry in Section 3: reserving too many non-text experts (Top-32) consumes the bit budget on coarsely-routed experts and drops Retain from 96.11% (W/O NT) to 95.34%, while reserving a sharp top slice (Top-8) recovers genuinely non-text-specialist experts and lifts Retain to 96.16%. This matches the design intuition: text routing drives most of the allocation, but a thin layer of high-confidence non-text experts remains worth protecting.

Table 3: Activation-precision and expert-weight-budget ablations. Panel (a) compares cross-modal activation policies at the 2.52-bit expert budget; **NT** and **T** denote non-text and text tokens, and "MoE" restricts the policy to MoE layers. Panel (b) varies the expert-weight budget; "Precision Set" lists the bit-widths available for per-expert assignment.

(a) Cross-modal asymmetric activation quantization.

Activation setting	OCRBench	MME	MMB-en	ChartQA	Video-MME	Retain (%)
NT16+T16	864	2298	82.90	79.52	64.22	96.11
NT4+T8	851	2300	82.65	78.40	64.63	95.69
NT4+T8-MoE	860	2327	83.25	79.28	64.93	96.29
NT16+T4-MoE	852	2310	79.64	79.16	64.37	95.21
NT16+T4	853	2236	79.64	74.52	61.26	92.96
NT4+T4-MoE	844	2329	81.70	77.36	64.15	95.16

(b) Expert-weight budget sensitivity.

Precision Set	Avg. Bit	OCRBench	MME	MMB-en	Video-MME	Retain (%)
{2, 3}	2.27	842	2280	82.13	62.00	93.14
{2, 3}	2.50	855	2315	83.08	64.48	95.83
{2, 3}	2.52	864	2298	82.90	64.22	96.11
{2, 3, 4}	2.52	860	2331	82.79	64.30	96.06
{2, 4}	2.52	851	2350	82.9	64.03	94.99
{2, 4}	3.00	864	2378	83.76	64.15	96.62

Cross-modal asymmetric activation quantization. Panel (a) of Table 3 compares activation quantization policies at the 2.52-bit expert budget. Three patterns emerge. (i) Reducing text activations to 4-bit is consistently the most harmful configuration, with the largest drops on reasoning-heavy benchmarks (MMB-en, ChartQA, Video-MME) and the lowest aggregate Retain (92.96% in the global variant). (ii) The NT4+T8 only-MoE policy matches or improves over the all-16-bit baseline on MME, MMB-en, and Video-MME, attaining the best Retain (96.29%). (iii) Restricting this policy to MoE layers outperforms applying it globally, indicating that the precision-sensitivity asymmetry is largely localized to expert computation rather than the dense backbone. These observations support using 8-bit text activations while quantizing non-text activations more aggressively.

Budget sensitivity. Panel (b) sweeps the average expert bit-width. As expected, performance increases monotonically with the overall budget. Crucially, the panel justifies our choice of the restricted $\{2, 3\}$-bit split over configurations that include 4-bit assignments at the same 2.52-bit average. While allocating 4 bits to top experts seems beneficial, maintaining this tight global budget forces a disproportionate number of remaining experts down to 2 bits. The severe degradation from these excessive 2-bit experts easily outweighs the local gains, dropping the overall Retain from 96.11% down to 94.99%. Finally, all results are averaged over three calibration re-samples (Retain standard deviation < 0.1), confirming these structural gaps are significant and not due to noise.

6 Conclusion

We have presented MDMixQ, a modality-decoupled mixed-precision quantization framework that addresses a specific failure mode of existing MoE compression: fused-modality calibration allows high-volume non-text routing noise to dominate expert importance estimates, systematically under-protecting the text-specialized experts that drive cross-modal reasoning. MDMixQ resolves this by using MDC to isolate a clean text-routing signal from multimodal inputs, refining the signal with dynamic threshold filtering and logit rectification, allocating per-expert bit-widths under a global budget, and applying a cross-modal asymmetric activation policy that protects text activations while compressing non-text activations inside MoE layers. At a 2.52-bit average expert budget, the best MDMixQ configurations retain 96.29%, 96.05%, and 95.33% of FP16 performance on Qwen3-VL-30B-A3B, Kimi-VL-16B-A3B, and Qwen3-Omni-30B-A3B, respectively, outperforming prior expert-priority methods across two model families. The allocation principle also transfers to MXFP quantization with 99.72% retention. More broadly, our results suggest that the routing asymmetry between modalities is a general property of multimodal MoE architectures and that exploiting it at the calibration stage yields gains that are orthogonal to weight-only or activation-only compression. We expect the same decoupling principle to combine naturally with non-text-token pruning, quantization-aware fine-tuning, and hardware-aware activation scheduling.

AI Use Statement. In this work, we used generative AI tools for improving the readability of the text, LaTeX typesetting, and organizing the released code; we have not used generative AI tools for designing or running experiments, analyzing results, or formulating the claims, and the remaining tasks listed in the ICLR AI Policy for Authors are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1533–1544, 2013.

Yuanteng Chen, Nanxin Zeng, Peisong Wang, Zhilei Liu, Yuantian Shao, Shiqiang Lang, Tao Liu, Chuangyi Li, Qinghao Hu, Gang Li, Jing Liu, and Jian Cheng. MODE: Modality-decomposed expert-level mixed-precision quantization for MoE multimodal LLMs. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026. URL https://arxiv.org/abs/2606.17118. Main Conference.

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*, pp. 4023–4032, 2025.

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. *CoRR*, abs/2604.06515, 2026. doi: 10.48550/ARXIV.2604.06515. URL https://doi.org/10.48550/arXiv.2604.06515.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026. doi: 10.48550/arXiv.2606.19348. URL https://arxiv.org/abs/2606.19348.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International conference on machine learning*, pp. 5547–5569. PMLR, 2022.

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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*, pp. 14793–14806. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/duanmu25a.html.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23:120:1–120:39, 2022. URL https://jmlr.org/papers/v23/21-0998.html.

Elias Frantar and Dan Alistarh. Qmoe: Practical sub-1-bit compression of trillion-parameter models. *arXiv preprint arXiv:2310.16795*, 2023.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *ArXiv*, abs/2210.17323, 2022. URL https://api.semanticscholar.org/CorpusID:253237200.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Tianyu Guo, Shanwei Zhao, Shiai Zhu, and Chenguang Ma. SPEED-Q: Staged processing with enhanced distillation towards efficient low-bit on-device VLM quantization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21486–21494, 2026. doi: 10.1609/aaai.v40i26.39296. URL https://ojs.aaai.org/index.php/AAAI/article/view/39296.

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

Young Jin Kim, Raffy Fahim, and Hany Hassan Awadalla. Mixture of quantized experts (moqe): Complementary effect of low-bit quantization and robustness. *arXiv preprint arXiv:2310.02410*, 2023.

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 *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. URL https://openreview.net/forum?id=qrwe7XHTmYb.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.

Shiyao Li, Yingchun Hu, Xuefei Ning, Xihui Liu, Ke Hong, Xiaotao Jia, Xiuhong Li, Yaqi Yan, Pei Ran, Guohao Dai, Shengen Yan, Huazhong Yang, and Yu Wang. MBQ: Modality-balanced quantization for large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4167–4177, 2025. doi: 10.1109/CVPR52734.2025.00394. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_MBQ_Modality-Balanced_Quantization_for_Large_Vision-Language_Models_CVPR_2025_paper.html.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Hangyu Guo, Yiming Liang, Jiaheng Liu, Noah Wang, Jian Yang, Siwei Wu, et al. Omnibench: Towards the future of universal omni-language models. *Advances in Neural Information Processing Systems*, 38, 2026.

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. *Proceedings of machine learning and systems*, 6: 87–100, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024a.

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, 2024b.

Ahmed Masry, Xuan Long Do, 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, 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. URL https://openaccess.thecvf.com/content/WACV2022/html/Mathew_InfographicVQA_WACV_2022_paper.html.

Alexandre Muzio, Alex Sun, and Churan He. Seer-moe: Sparse expert efficiency through regularization for mixture-of-experts. *arXiv preprint arXiv:2404.05089*, 2024.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michele Tadmor Ramanovich. Spoken question answering and speech continuation using spectrogram-powered llm. In *International Conference on Learning Representations*, volume 2024, pp. 51883–51898, 2024.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Guangshuo Qin, Zhiteng Li, Zheng Chen, Weihang Zhang, Linghe Kong, and Yulun Zhang. VEQ: Modality-adaptive quantization for MoE vision-language models. *arXiv preprint arXiv:2602.01037*, 2026. doi: 10.48550/arXiv.2602.01037. URL https://arxiv.org/abs/2602.01037.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Gao Peng, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 45472–45496, 2024.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

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.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025. URL https://arxiv.org/abs/2504.07491.

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 *International conference on machine learning*, pp. 38087–38099. PMLR, 2023.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

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. *Advances in neural information processing systems*, 35:27168–27183, 2022.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pp. 69–85. Springer, 2016.

Manyi Zhang, Ji-Fu Li, Zhongao Sun, Haoli Bai, Hui-Ling Zhen, Zhenhua Dong, and Xianzhi Yu. Benchmarking post-training quantization of large language models under microscaling floating point formats. *arXiv preprint arXiv:2601.09555*, 2026.

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.

A Appendix

A.1 Implementation Details

Calibration and salience collection. For each backbone, all methods use the same 50 calibration examples in each of three independently re-sampled subsets. Qwen3-VL-30B-A3B and Kimi-VL-16B-A3B use COCO, whereas Qwen3-Omni-30B-A3B uses COCO-Audio. During each multimodal forward pass, tokens are partitioned using the model-specific modality boundaries. MDC uses text-token router outputs for default expert ranking; non-text salience is used only in the Top-K preservation ablation. At every MoE layer, we retain the router top-K_r candidates with $K_r = 8$, compute the layerwise confidence threshold in Equation (3), and accumulate the rectified logits in Equation (4).

Quantization configuration. Unless otherwise specified, the vision and audio encoders remain in FP16, and the attention QKV/O projections use 4-bit weights. Routed expert weights are assigned from $\{2, 3\}$ bits under a 2.52-bit average budget by globally ranking all layer–expert pairs according to text salience. The default MDMixQ configuration does not reserve additional non-text experts. The base variant keeps MoE activations at 16 bits; MDMixQ + asym. uses 4-bit non-text and 8-bit text activations only inside MoE layers. All compared INT weight quantizers use group-wise symmetric quantization with 128 weights per group and max-absolute-value clipping: for bit-width b, $s_G = \max_{w \in G} |w| / (2^{b-1} - 1)$ and $q = \mathrm{clip}(\mathrm{round}(w/s_G), -2^{b-1} + 1, 2^{b-1} - 1)$. The group scale is stored in FP16. Evaluation is run through the vLLM backend; its Kimi-VL implementation does not support video inputs, so video results are not reported for that backbone.

A.2 Analytical Weight Storage and Compute

Let P_E, P_A, and P_R be the numbers of routed-expert, decoder-attention projection, and remaining parameters. With FP16 scales per 128-weight group, the estimated storage in bytes is

$$M = 2P_R + \left(\frac{4}{8} + \frac{2}{128}\right) P_A + \left(\frac{\bar{b}}{8} + \frac{2}{128}\right) P_E, \tag{5}$$

where $\bar{b} = 4$ for W4A16 and approximately 2.52 for MDMixQ. FP16 storage is $2(P_E + P_A + P_R)$. This estimate includes packed weights and group scales, but excludes checkpoint metadata, alignment padding, and runtime memory. For Qwen3-Omni, only the Thinker decoder is quantized; Talker weights remain FP16. Kimi-VL's shared experts remain FP16, and its four latent-attention projections are included in P_A.

Table 4: Analytical weight storage (decimal GB). MDMixQ + asym. act. has the same persistent weight storage as MDMixQ.

Backbone	FP16	W4A16	MDMixQ	FP16 / MDMixQ
Qwen3-VL-30B-A3B	62.14	17.76	12.40	5.01×
Kimi-VL-16B-A3B	32.82	10.90	8.23	3.99×
Qwen3-Omni-30B-A3B	70.52	26.14	20.78	3.39×

Quantization changes arithmetic precision, but neither the network structure nor the number of routed experts. Consequently, the conventional matrix-operation FLOP count has a 0% theoretical reduction relative to FP16 at a fixed input and routing trace; this count does not measure low-precision operator speedup.

A.3 Text-Only Evaluation

We evaluate text-only capabilities on Qwen3-VL-30B-A3B and Kimi-VL-16B-A3B using the same quantization configurations as in the main experiments. Table 5 reports the unweighted average over six benchmarks. MDMixQ exceeds MODE on both backbones, improving the average by 5.34 points on Kimi-VL and 2.98 points on Qwen3-VL. Its averages remain below the corresponding FP16 baselines by 5.23 and 6.03 points, respectively. These results suggest that the text-guided expert allocation also better preserves text-only capability than the compared mixed-precision baseline, while leaving a measurable gap to FP16.

Table 5: Text-only benchmark results (%). Baseline denotes FP16; MODE and MDMixQ use the same 2.52-bit expert-weight budget and quantization settings as in the main experiments. Avg is the unweighted mean of six scores. Green shading marks MDMixQ; bold marks the best quantized average.

Backbone	Method	GPQA-Diamond	MATH-500	IFEval	HellaSwag	MMLU	WinoGrande	Avg
Kimi-VL-16B-A3B	Baseline	27.10	62.80	43.07	62.09	79.79	73.87	58.12
	MODE	17.34	38.60	40.30	53.24	66.87	68.95	47.55
	MDMixQ	20.76	56.24	41.76	57.97	70.47	70.12	**52.89**
Qwen3-VL-30B-A3B	Baseline	68.87	88.24	85.48	63.28	81.28	77.39	77.42
	MODE	56.29	69.21	82.42	57.29	74.86	70.41	68.41
	MDMixQ	60.48	75.28	81.58	60.61	76.18	74.22	**71.39**

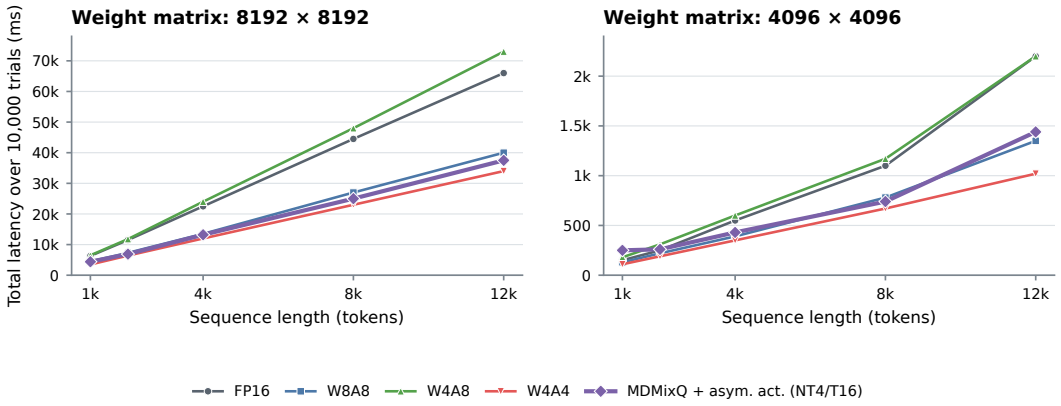


Figure 4: Measured single-layer NPU operator latency over 10,000 trials at two weight-matrix shapes. WbAa denotes b-bit integer weights and a-bit integer activations. The MDMixQ mixed-activation policy uses 4-bit non-text and 8-bit text activations inside MoE layers (NT4/T8). Curves are redrawn from the provided latency plot.

A.4 Operator-Level Speedup from Asymmetric Activation Quantization

Kernel implementation. For an MoE operator with m text tokens, r non-text tokens, and an INT4 weight tile $W_4 \in \mathbb{Z}^{k\times n}$, we quantize the non-text branch to INT4 and the text branch to INT8. To execute both branches with the NPU's INT4 matrix-multiply primitive, each signed text code $q_8 \in [-128, 127]$ is decomposed into two signed INT4 codes:

$$h = \left\lfloor \frac{q_8 + 128}{16} \right\rfloor - 8, \qquad \ell = ((q_8 + 128) \bmod 16) - 8, \qquad q_8 = 16h + \ell + 8. \tag{6}$$

The non-text branch needs one INT4 matrix multiplication, while the text branch uses two, accumulating in INT32. Its integer result is reconstructed as $16(HW_4) + LW_4 + 8\mathbf{1}_m\mathbf{1}_k^\top W_4$; the last term is a reusable per-output-channel correction for the centered INT4 codes. We apply the corresponding quantization scales, cast the outputs to FP16, and restore the original token order. This decomposition lets the mixed NT4/T8 activation policy use the same INT4 NPU primitive for both modalities.

Operator latency. Figure 4 compares single-layer NPU operator latency across sequence lengths for two weight-matrix shapes. The mixed-precision curve corresponds to *MDMixQ + asym. act.* (NT4/T8) in the main paper; FP16, W8A8, W4A8, and W4A4 are fixed-precision references. At longer sequences, the asymmetric policy has lower latency than FP16 and W4A8 in both settings, although W4A4 remains faster. The comparison reflects operator-level execution rather than end-to-end model inference, and the reference configurations do not share the same expert-weight budget.

Table 6: Full per-benchmark results for the main-paper ablations at the same 2.52-bit expert budget: modality-decoupled calibration, Top-K non-text expert preservation, and cross-modal asymmetric activation. Bold marks the best value within each sub-table.

Modality-decoupled calibration — *All inputs are multimodal; "imgxk+txtyk" counts tokens used for routing statistics ($k = 10^3$).*

Variant	OCRBench	MME	MVBench	MMB-en	ChartQA	TextVQA	RefCOCO	Video-MME	Retain (%)
img10k + txt0	856	2284.75	67.80	82.90	78.48	81.70	93.36	**65.11**	95.50
img5k + txt5k	866	2253.31	67.75	82.99	76.64	81.62	93.50	64.22	95.05
img2k + txt8k	**868**	2318.41	67.80	82.99	77.76	82.08	94.18	64.07	95.72
img0 + txt10k	861	**2320.81**	**68.50**	**83.33**	**78.88**	**82.20**	**94.29**	64.81	**96.14**

Top-K non-text expert preservation — *"W/O NT" uses text salience only; "W/ NT Top-K" reserves K non-text experts at the higher bit-width.*

Variant	OCRBench	MME	MVBench	MMB-en	ChartQA	TextVQA	RefCOCO	Video-MME	Retain (%)
W/O NT	864	2298.61	**68.75**	82.90	**79.52**	**82.48**	**94.54**	64.22	96.11
W/ NT Top-8	**866**	2307.96	68.80	83.16	78.72	82.10	94.33	**64.93**	**96.16**
W/ NT Top-16	855	2306.02	68.03	**83.42**	77.88	81.89	94.56	64.15	95.63
W/ NT Top-32	849	**2322.94**	68.03	82.65	77.80	81.53	93.88	64.00	95.34

Cross-modal asymmetric activation — ***NT*** *and* ***T*** *in Act. Bits denote* ***non-text tokens*** *and* ***text tokens****, and "only-MoE" restricts the policy to MoE layers.*

Variant	OCRBench	MME	MVBench	MMB-en	ChartQA	TextVQA	RefCOCO	Video-MME	Retain (%)
NT16 + T16	**864**	2298.61	**68.75**	82.90	**79.52**	**82.48**	**94.54**	64.22	96.11
NT4 + T4 only-MoE	844	**2329.52**	67.92	81.70	77.36	81.79	93.93	64.15	95.16
NT4 + T8	851	2300.37	68.53	82.65	78.40	82.40	94.01	64.63	95.69
NT4 + T8 only-MoE	860	2327.23	68.65	**83.25**	79.28	**82.48**	94.22	**64.93**	**96.29**
NT16 + T4 only-MoE	852	2310.25	68.08	79.64	79.16	81.84	93.86	64.37	95.21
NT16 + T4	853	2236.38	66.17	79.64	74.52	80.91	92.81	61.26	92.96

Table 7: MXFP quantization results on Qwen3-VL-30B-A3B. "W4A4" denotes uniform MXFP4 weights and activations; "asym. act." applies non-text MXFP4 / text MXFP8 activation quantization inside MoE layers. Bold marks the best quantized result in each column.

Method	Avg. Bit	OCRBench	MME	MVBench	MMB-en	ChartQA	TextVQA	RefCOCO	Video-MME	Retain (%)
FP16	16	898	2426	69.95	87.37	85.12	84.65	96.14	67.22	100.00
MXFP W4A4	4	885	2406	68.40	86.42	83.28	83.68	95.49	65.38	98.46
MDMixQ	6	**898**	**2430**	69.70	86.08	**84.68**	**84.32**	**96.08**	**67.48**	**99.72**
MDMixQ + asym. act.	6	896	**2430**	**70.05**	**86.72**	84.44	84.00	95.99	66.63	99.59

A.5 Full Ablation Results

Table 6 reports the full per-benchmark numbers for the three ablations summarized in Section 5.3 of the main paper — modality-decoupled calibration, Top-K non-text expert preservation, and cross-modal asymmetric activation, all at the 2.52-bit expert budget. The design intuitions and per-row analyses are discussed in Section 5.3 of the main paper and are not repeated here. The budget sensitivity sweep is discussed in full in the main paper (Table 3b) and is not duplicated in this appendix.

A.6 Generalization under MXFP Quantization

We further evaluate MDMixQ under MXFP quantization, where the available precision levels are restricted to MXFP4 and MXFP8. On Qwen3-VL-30B-A3B, MDMixQ assigns MXFP8 to the experts ranked most important by modality-decoupled calibration and MXFP4 to the remaining experts, yielding an average expert-weight precision of 6 bits. As shown in Table 7, this allocation retains 99.72% of FP16 performance, compared with 98.46% for uniform MXFP W4A4. Applying the cross-modal asymmetric activation policy retains 99.59%, remaining within 0.5 Retain points of FP16. These results show that the proposed allocation principle is not tied to integer quantization and transfers effectively to a hardware-oriented floating-point format.

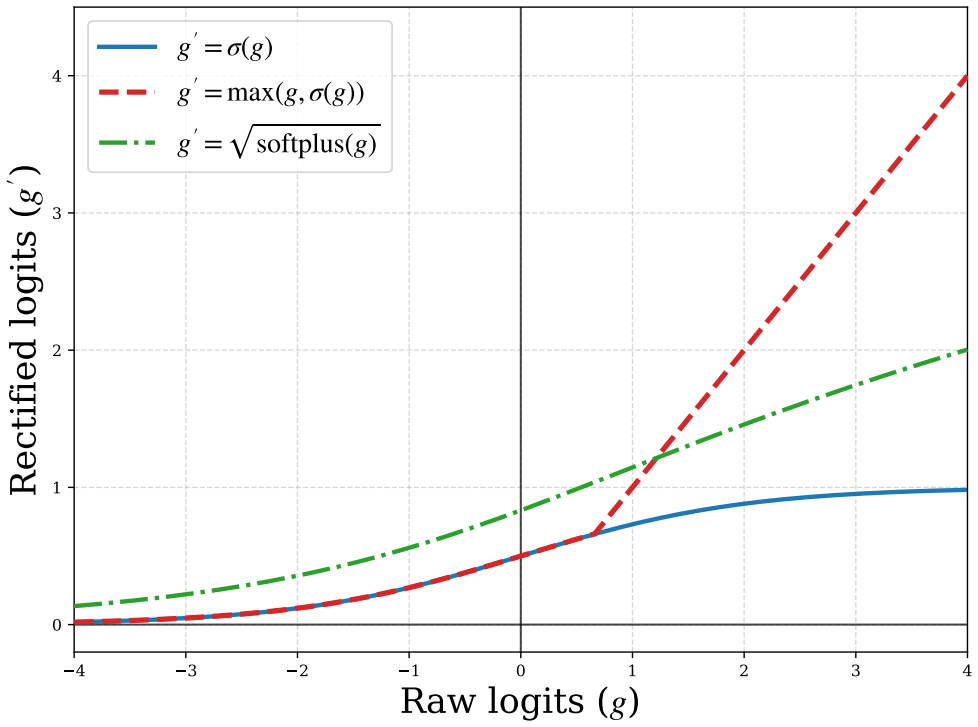


Figure 5: Candidate rectifiers as a function of the raw routing logit g: $\sigma(g)$ saturates on large positive logits; $\sqrt{\text{softplus}(g)}$ compresses them sub-linearly; $\max(g, \sigma(g))$ (ours) is bounded on the negative side and reduces to the identity on the positive side, preserving linear discriminability in the regime that matters for salience ranking.

Table 8: Rectification function ablation at the same 2.52-bit expert budget. "Softmax" replaces the rectifier with a Softmax aggregation; σ, $\sqrt{\text{softplus}}$, and $\max(g, \sigma(g))$ correspond to the three rectifiers in Figure 5.

Rectifier	OCRBench	MME	MVBench	MMB-en	ChartQA	TextVQA	RefCOCO	Video-MME	Retain (%)
Softmax	846	**2337.50**	67.88	83.25	77.32	**82.51**	93.32	64.00	95.43
$\sqrt{\text{softplus}(g)}$	850	2309.37	68.70	82.90	79.48	82.05	94.39	64.26	95.88
$\max(g, \sigma(g))$ (ours)	**864**	2298.61	**68.75**	82.90	**79.52**	82.48	**94.54**	64.22	**96.11**

A.7 Rectification Function Ablation

The non-linear rectifier in Section 4.2 converts raw routing logits into a positive, monotonic importance signal before aggregation. Section 3 and Section 4.2 argued that the rectifier must (i) bound the magnitude of negative logits so that frequently penalised experts are not subtracted out of the running salience sum, and (ii) preserve the linear, discriminative regime on large positive logits so that the small set of experts that dominate confident text-side routing remains separable. Figure 5 plots the three candidate rectifiers under consideration:

$$\begin{aligned} &\text{(i) } g' = \sigma(g), \quad \text{(ii) } g' = \max(g, \sigma(g)), \\ &\text{(iii) } g' = \sqrt{\text{softplus}(g)}. \end{aligned} \tag{7}$$

$\sigma(g)$ alone saturates to 1 on large positive logits, collapsing the signal that distinguishes the most confident experts. $\sqrt{\text{softplus}(g)}$, the rectifier used to compute expert affinity scores in DeepSeek-V4 (DeepSeek-AI, 2026), is smoother and avoids the hard corner at $g = 0$, but it sub-linearly compresses large positive logits, again attenuating the high-confidence regime. Only $\max(g, \sigma(g))$ satisfies both desiderata simultaneously: it is bounded in $(0, 0.5)$ for negative logits and reduces to the identity on large positive logits.

Table 8 reports the per-benchmark numbers at the same 2.52-bit expert budget. The empirical ordering matches the analysis. Softmax aggregation drops aggregate Retain to 95.43%: by normalising logits within each token before accumulation, it amplifies the contribution of trivial tokens that routed near-uniformly across experts and dilutes the sharp signal from pivotal tokens identified in Section 3. The $\sqrt{\mathrm{softplus}(g)}$ rectifier closes most of this gap (95.88%) because it suppresses negative logits more gracefully than Softmax, but the sub-linear compression on large positive logits costs about 0.2 Retain points relative to $\max(g, \sigma(g))$. Our default rectifier achieves the highest Retain (96.11%) and the best score on five of eight benchmarks, directly validating the design argument in Section 4.2: among the candidates that bound the negative regime, the one that additionally preserves identity on the positive regime gives the most reliable salience ranking.

A.8 Benchmarks and Evaluation Details

We evaluate MDMixQ on three backbones from two model families: Qwen3-VL-30B-A3B, Kimi-VL-16B-A3B (Team et al., 2025), and Qwen3-Omni-30B-A3B. Together, they cover vision–language and audio–visual–language scenarios. We describe the evaluation benchmarks by task type below.

OCR and document understanding. **OCRBench** evaluates the optical character recognition capability of the model across various text recognition, text-based visual question answering, and key information extraction tasks. It measures the ability of the model to perceive and reason over textual content embedded in natural images.

InfoVQA, formally introduced as InfographicVQA (Mathew et al., 2022), evaluates visual question answering over information-rich infographics. Its questions require joint reasoning over document layout, textual content, graphical elements, and data visualizations. We include InfoVQA in the Kimi-VL-16B-A3B evaluation.

Chart understanding. **ChartQA** contains chart images paired with question–answer annotations that require numerical reasoning, logical inference, and visual structure parsing. Performance on ChartQA reflects the model's ability to extract quantitative information from structured visual representations.

Visual grounding. **RefCOCO** is a referring expression comprehension benchmark where models must localize objects in images based on natural language descriptions. This task probes the alignment between visual perception and fine-grained language understanding.

Image–text reasoning. This category comprises three benchmarks. **MME** provides a comprehensive evaluation across 14 subtasks spanning perception and cognition, including object recognition, scene understanding, commonsense reasoning, and numerical calculation. **MMB-en** assesses multimodal understanding through multiple-choice questions covering a broad range of real-world visual reasoning scenarios, with an English evaluation split. **TextVQA** requires models to answer questions by reading and reasoning over text present in images, combining OCR with high-level semantic understanding.

Video understanding. Two benchmarks capture temporal reasoning. **MVBench** evaluates video understanding through temporally grounded tasks such as action recognition, temporal localization, and event ordering. **Video-MME** is a comprehensive video evaluation benchmark covering both short- and long-form videos with diverse question types, probing a model's ability to track events, characters, and causal relationships across extended temporal windows.

Speech and audio benchmarks. For the Qwen3-Omni-30B-A3B backbone, we additionally evaluate on speech and omnimodal benchmarks. **LibriSpeech ASR** measures the word error rate on English speech recognition to evaluate the ability to transcribe spoken language under varied acoustic conditions. **SpeechQA** probes spoken question answering, requiring the model to understand spoken queries and produce accurate textual answers. **OmniBench** is an omnimodal benchmark that simultaneously assesses vision, language, and speech understanding in a unified setting.

Retain metric. To aggregate performance over heterogeneous benchmarks with different scales and directions, we report a normalized *Retain* score against the corresponding FP16 backbone. For a higher-is-better benchmark i,

$$R_i = \frac{s_i}{s_i^{\text{ref}}}, \tag{8}$$

and for a lower-is-better benchmark (LibriSpeech ASR word error rate),

$$R_i = \frac{100 - s_i}{100 - s_i^{\text{ref}}}. \tag{9}$$

The overall *Retain* (%) metric is defined as the arithmetic mean of the normalized scores:

$$\text{Retain} = 100 \cdot \frac{1}{N} \sum_{i=1}^{N} R_i, \tag{10}$$

where s_i^{ref} is the FP16 score on the same backbone. Retain (%) thus reflects the aggregate normalized performance, representing the average fraction of FP16 accuracy preserved across the evaluation suite.

A.9 LIMITATIONS

We discuss two main limitations. First, we report analytical weight storage (Section A.2) and single-operator NPU latency (Section A.4) but no end-to-end speedup: expert-level mixed precision assigns heterogeneous bit-widths (e.g., a $\{2, 3\}$-bit split across layer–expert pairs), while current inference backends (including the vLLM pipeline used in our evaluation) provide optimized kernels only for uniform-precision expert weights, so mixed-precision expert execution falls back to slower generic paths or requires repacking at load time, obscuring the latency benefit of the reduced average bit-width. The reported gains of MDMixQ should therefore be interpreted as memory-footprint and accuracy–compression improvements; realized speedup would require co-designed mixed-precision GroupGEMM kernels, which we leave to future work. Second, MDMixQ quantizes only the routed-expert weights, keeping attention projections at uniform 4 bits and the vision/audio encoders in FP16, so the compression ratio of the full model is bounded by these components.