MAQUANT: MODALITY-AWARE QUANTIZATION FOR OMNIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Omnimodal large language models (OLLMs) deliver strong capabilities but incur substantial memory and computation costs. Low-bit quantization is a practical remedy, yet existing methods mainly target text-only large language models (LLMs) or large vision-language models (LVLMs) and may inadequately address heterogeneous activation distributions, outlier-dominated degradation, and cross-modal error coupling in OLLMs. We propose **MAQuant**, a modality-aware post-training quantization framework that combines **Modality-Specific Learnable Transformation** (MSL), **Exponent-weighted Activation Penalty** (EAP), and **Cross-modal Error Exchange** (CME) to learn modality-specific quantization-friendly spaces, protect salient activations, and share channel-wise sensitivity across modalities. EAP and CME are calibration-only objectives and add no inference-time loss computation. On Qwen2.5-Omni-7B and 3B under W4A8 and W4A4, MAQuant achieves the highest six-task average among the evaluated baselines across both model sizes and bit-width settings. On Qwen2.5-Omni-7B, W4A4 retains 97.8% of the full-precision six-task average, with estimated deployment storage of 4.530 GiB including auxiliary transformations.

1 INTRODUCTION

Omnimodal large language models (OLLMs) have emerged as a promising direction toward general-purpose artificial intelligence (AI) by processing vision, audio, video, and text within a unified framework (Hurst et al., 2024; Wu et al., 2024; Zhan et al., 2024; Xu et al., 2025). With modality-specific encoders and a powerful language decoder, OLLMs support diverse tasks such as multimodal reasoning (Hudson & Manning, 2019; Singh et al., 2019; Mathew et al., 2021; Lu et al., 2022; Wang et al., 2024b), instruction following (Ouyang et al., 2022; Zhu et al., 2023), grounded generation (Peng et al., 2023; Chen et al., 2023), and interactive assistance (Team et al., 2023; Xu et al., 2025). However, jointly deploying encoders, projection modules, and the language decoder leads to substantial memory and inference costs (Zhou et al., 2024; Jin et al., 2025), making efficient deployment crucial.

Quantization is an effective way to reduce model size and computation (Wan et al., 2023; Gong et al., 2025). Existing methods work well for large language models (LLMs) and large vision-language models (LVLMs), typically by minimizing weight error, activation discrepancy, or layer-wise output mismatch. These objectives are suitable when representations are relatively homogeneous, where local numerical fidelity can approximate model behavior.

However, OLLMs present a more challenging quantization scenario. Heterogeneous modality tokens originate from different input structures and encoders but are mapped into a shared language representation space. Therefore, quantization errors depend not only on local approximation quality, but also on modality-specific activation distributions, outliers, and cross-modal error interactions.

As shown in Figure 1, existing PTQ methods typically adopt globally shared calibration objectives or equivalent transformations, which are effective for LLM compression (Frantar et al., 2022; Xiao et al., 2023; Lin et al., 2024b; Liu et al., 2024; Sun et al., 2024). However, LVLMs project heterogeneous modality inputs into a shared decoder space (Alayrac et al., 2022; Li et al., 2023), where token distributions and quantization sensitivities can vary significantly across modalities. Recent LVLM quantization studies show that uniform calibration over visual and textual tokens may overfit insensitive modalities and hurt low-bit performance (Li et al., 2025; Wang et al., 2024a; Yu

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.

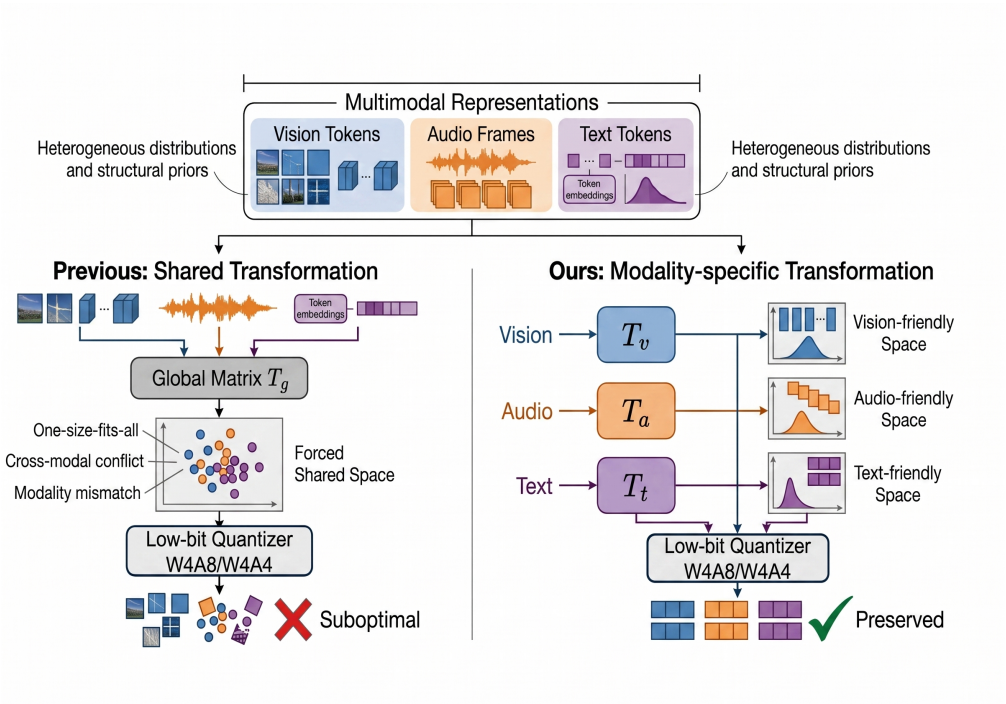


Figure 1: Motivation for modality-aware calibration: encoder branches use separate transformations, while vision, audio, and text tokens share one transformation at each decoder site. Modality-conditioned losses and cross-modal error exchange guide this common decoder quantization space; separate text, vision, and audio decoder weights are not required.

et al., 2025). This motivates a modality-aware PTQ strategy that learns quantization-friendly spaces for different modalities while maintaining consistency in the shared decoder representation space.

As shown in Figure 2, modality-aware quantization for OLLMs remains challenging due to three key issues. First, modality-dependent quantization geometry arises because different modalities exhibit distinct activation statistics and structural properties, motivating branch-specific encoder transformations and modality-aware decoder calibration. Second, outlier-dominated activation distortion occurs because low-bit quantization errors are often dominated by a few salient tokens or channels, while uniform reconstruction objectives may allocate calibration capacity inefficiently. Third, cross-modal error coupling emerges because all modalities are ultimately processed by shared decoder layers, causing quantization-sensitive directions to be coupled across modalities.

Motivated by these observations, we propose MAQuant, a modality-aware post-training quantization framework for OLLMs. MAQuant jointly considers modality-specific transformation, activation salience, and cross-modal error sensitivity during calibration.

MAQuant consists of three components: (i) **M**odality-**S**pecific **L**earnable Transformation (**MSL**), which learns branch-specific encoder transformations and optimizes a shared decoder quantization space using modality-conditioned reconstruction and cross-modal guidance. (ii) **E**xponent-weighted **A**ctivation **P**enalty (EAP), which emphasizes activation-salient tokens and channels to reduce outlier-induced distortion. (iii) **C**ross-**M**odal Error **E**xchange (CME), which transfers normalized channel-wise error profiles from source modalities to a different target modality, coupling their reconstruction objectives when learning the shared decoder transformations. By mitigating modality conflicts under uniform calibration and preserving key activation structures, MAQuant enables robust W4A4 full-modality quantization for the targeted linear layers in modality encoders, connectors, and the shared language decoder while maintaining strong omnimodal understanding and reasoning performance.

Our contributions are summarized as follows:

- We propose **MAQuant**, a full-modality low-bit post-training quantization framework for encoder-decoder OLLMs, enabling unified quantization of the targeted linear layers in modality encoders, connectors, and the shared decoder.

- Within MAQuant, we develop **Modality-Specific Learnable Transformation** (MSL), **Exponent-weighted Activation Penalty** (EAP), and **Cross-modal Error Exchange** (CME) to address modality-dependent quantization geometry, outlier-dominated activation distortion, and cross-modal error coupling, respectively. MSL provides the shared learnable coordinate system, while CME explicitly couples modality-conditioned reconstruction objectives within that system.
- Extensive experiments on Qwen2.5-Omni-7B and Qwen2.5-Omni-3B show that MAQuant achieves average scores of 68.00 and 63.30 under W4A4, indicating better robustness than previous methods under aggressive activation quantization.

2 Related Work

Quantization of Large Language Models. Post-training quantization has become an important technique for efficient LLM deployment. Existing methods improve low-bit performance from two directions: reconstruction-based optimization and representation transformation. Reconstruction-based methods, such as GPTQ (Frantar et al., 2022), AWQ (Lin et al., 2024b), and OmniQuant (Shao et al., 2023), reduce quantization error through second-order reconstruction, activation-aware channel protection, or learnable calibration. Transformation-based methods, including SmoothQuant (Xiao et al., 2023), SpinQuant (Liu et al., 2024), FlatQuant (Sun et al., 2024), and DuQuant (Lin et al., 2024a), reshape activations and weights into more quantization-friendly distributions via scaling, rotation, or distribution flattening. These methods perform well under 4-bit or mixed-precision settings but primarily target decoder-only LLMs. OLLMs combine modality-specific encoders, projection modules, and a shared decoder, introducing heterogeneous activation distributions and error patterns that complicate direct application of LLM quantization methods.

Quantization for Large Vision Language Models. Recent studies have extended quantization from LLMs to LVLMs. Compared with LLM quantization, multimodal quantization needs to preserve cross-modal interactions under low-bit compression (Wang et al., 2024a; Li et al., 2025; Yu et al., 2025). Q-VLM (Wang et al., 2024a) shows that visual and textual tokens exhibit heterogeneous distributions and require special treatment during calibration. MBQ (Li et al., 2025) introduces modality-balanced calibration to alleviate modality imbalance, while MQuant (Yu et al., 2025) explores static quantization for multimodal large language models. These works demonstrate that modality-aware calibration is crucial for reducing quantization degradation in multimodal models. Most LVLM quantization methods target decoder or projection layers while retaining higher-precision modality encoders. Encoders and decoders are also typically compressed with separate heuristics, leaving unified modality-aware quantization across the full OLLM architecture underexplored.

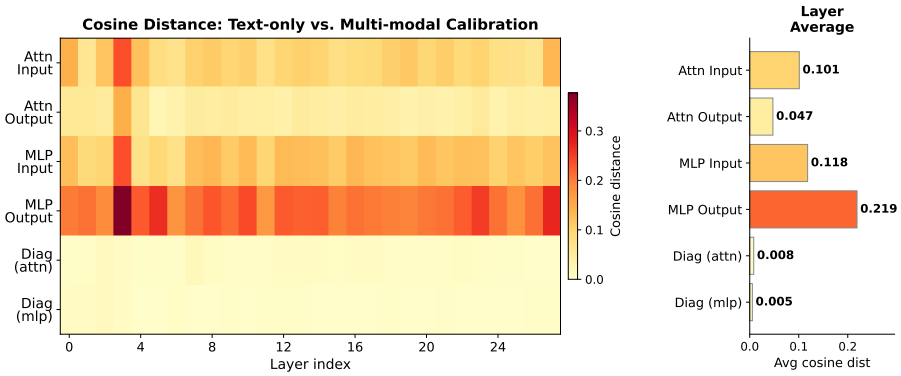
Omnimodal Large Language Models. Recent OLLMs generalize vision-language models to unified systems that process image, video, audio, and text (Team et al., 2023; Xu et al., 2025). They typically adopt an encoder-connector-decoder architecture, where modality-specific encoders extract heterogeneous features, connectors align them with the language space, and a shared decoder performs reasoning and generation (Alayrac et al., 2022; Wang et al., 2024b; Xu et al., 2025). Modality encoders differ in feature distributions, saliency, and numerical sensitivity, and their quantization errors can propagate through connectors into decoding. These coupled errors challenge full-architecture OLLM quantization beyond the efficient inference settings widely studied for LLMs and LVLMs (Wan et al., 2023; Zhou et al., 2024; Gong et al., 2025).

3 Observations

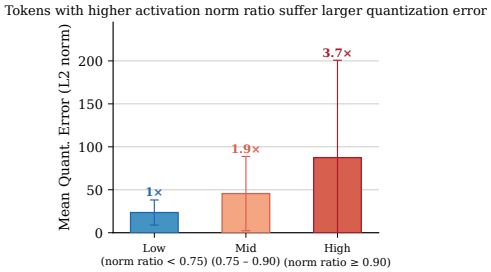
We identify three challenges for low-bit OLLM quantization arising from heterogeneous activations and cross-modal interactions.

Modality-dependent Transformation Preference. On Qwen2.5-Omni-7B under W4A4, transformations learned with text-only and multimodal calibration diverge more strongly for rotations than for diagonal scaling (Figure 2a). This suggests that the learned coordinates depend on calibration composition, motivating joint multimodal calibration of shared decoder transformations.

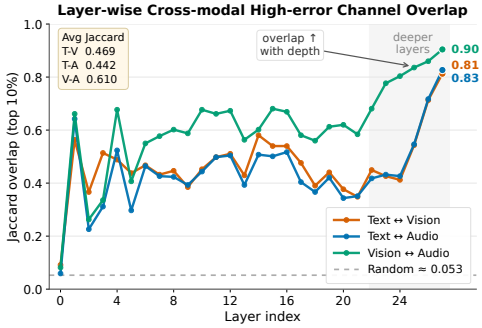
Challenge 1: Quantization should accommodate modality-dependent transformation preferences across the encoder–decoder architecture.



(a) Text-only versus multimodal calibration.



(b) Token sensitivity.



(c) Cross-modal error-channel overlap.

Figure 2: **Key observations motivating MAQuant.** **(a)** Differences between text-only and multimodal calibration are more pronounced in rotational transformations than in diagonal scaling. **(b)** Activation-prominent tokens suffer larger quantization distortion. **(c)** High-error channels substantially overlap across modalities and become increasingly shared in deeper layers.

Outlier-dominated Activation Quantization Error. Quantization error concentrates on a small number of activation-prominent tokens and channels (Figure 2b). Uniform reconstruction objectives can therefore over-emphasize low-impact activations and under-protect vulnerable outliers.

Challenge 2: Calibration should prioritize activation-prominent tokens and channels that dominate quantization error.

Cross-modal Error Coupling. High-error channels overlap across modalities well beyond the random baseline, increasingly so in deeper layers (Figure 2c). The incomplete overlap, especially in shallow layers, indicates complementary modality-specific sensitivities. Independent calibration can miss shared fragile directions, while uniform calibration can overlook these differences. Appendix C.3 provides per-channel visualizations.

Challenge 3: Modality-aware quantization should exploit shared fragile channels across modalities while preserving modality-specific quantization sensitivity.

4 METHOD

To solve the aforementioned challenges, we present **MAQuant**, a modality-aware post-training quantization framework for omnimodal large language models. MAQuant improves low-bit quantization by learning modality-specific pre-quantization transformations, which reshape activations and weights into quantization-friendly coordinate systems.

As illustrated in Figure 3, MAQuant consists of three components. First, **Modality-Specific Learnable Transformation** (MSL) learns branch-specific encoder transformations and a shared decoder coordinate system optimized with modality-conditioned losses. Second, **Exponent-weighted Activation Penalty** (EAP) imposes a salience-aware activation constraint to emphasize quantization-sensitive tokens and channels. Third, **Cross-modal Error Exchange** (CME) transfers channel-wise

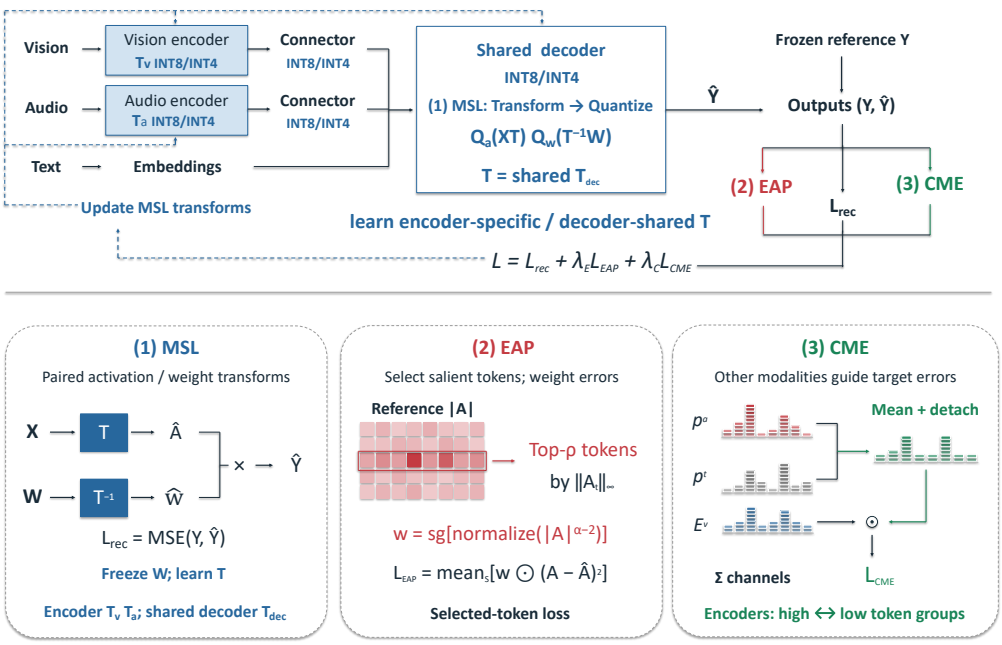


Figure 3: Overview of MAQuant. Encoder branches learn independent transformations. At each decoder site, modality masks separate the reconstruction errors of vision, audio, and text tokens; CME transfers detached channel-error profiles between these groups to guide the shared MSL transformation. Modality conditioning occurs in calibration, and all decoder tokens use the same transformed weight at inference.

quantization error profiles across modalities, encouraging different modalities to account for shared error-sensitive directions.

Given a pretrained omnimodal model, we apply MAQuant through post-training calibration. Specifically, we use a small multimodal calibration set to optimize MSL transformations jointly with the auxiliary objectives. EAP imposes salience-aware reconstruction penalties on module output activations to emphasize quantization-sensitive tokens and channels. Meanwhile, CME exchanges channel-wise quantization error profiles among modalities, enabling each modality to consider shared error-sensitive directions. After calibration, the learned transformations are folded into the quantized computation, and the model is converted into its final low-bit form.

4.1 Modality-Specific Learnable Transformation

MSL learns invertible transformations to reshape activations and weights into quantization-friendly representations. Its modality dependence follows the encoder–decoder architecture: vision and audio encoders have independent transformations, whereas each decoder site learns a common coordinate system for interacting visual, audio, and text representations. In the decoder, modality information conditions the calibration signals used to optimize this shared space.

For a calibrated linear layer, let $\mathbf{X}^m$ denote the activations associated with modality m, $\mathbf{W}$ its weight, and $\mathbf{T}$ its transformation. We omit layer and branch indices for clarity. The equivalent mapping $\mathbf{X}^m\mathbf{W} = (\mathbf{X}^m\mathbf{T})(\mathbf{T}^{-1}\mathbf{W})$ gives the quantized output

$$\widehat{\mathbf{Y}}^m = Q_a(\mathbf{X}^m\mathbf{T})Q_w(\mathbf{T}^{-1}\mathbf{W}), \tag{1}$$

where Q_a and Q_w denote activation and weight quantization. Encoder branches use their own $\mathbf{T}$. Decoder modality masks identify token origins. After self-attention, these token states may already contain information from other modalities, so it does not imply an isolated unimodal representation.

The reconstruction objective can be written by token-origin group as

$$\mathcal{L}_{\text{rec}} = \sum_{m \in \mathcal{M}} \frac{n_m}{N} \underbrace{\frac{1}{n_m d'} \left\| \mathbf{Y}^m - \widehat{\mathbf{Y}}^m \right\|_F^2}_{\mathcal{L}_{\text{rec}}^m}, \qquad N = \sum_{m \in \mathcal{M}} n_m, \tag{2}$$

where $\mathcal{M}$ contains the token modalities, n_m is their token count, and d' is the output dimension.

MSL comes from optimizing $\mathbf{T}$ together with CME: a channel found sensitive for one modality receives additional reconstruction weight for other modalities, and these target-specific losses jointly update the same transformation. MSL supplies the learnable shared coordinates; CME supplies the source-to-target error guidance. This allows the common quantization space to account for complementary modality sensitivities without storing separate decoder weights for each modality.

Following FlatQuant, we use lightweight Kronecker-structured transformations and stabilized inverse computation. This differs from assigning modality-dependent quantizer settings or smoothing scales: CME transfers output-channel error information between token modalities to optimize the common transformation and associated quantization parameters.

4.2 Exponent-weighted Activation Penalty

EAP focuses calibration on activation-prominent tokens and channels. Let $\mathbf{A}, \widehat{\mathbf{A}} \in \mathbb{R}^{N \times d}$ be the reference and quantized output activations of a calibrated module, with batch and token dimensions flattened. We select tokens by their maximum absolute activation:

$$\mathcal{S} = \text{TopK}_t(\|\mathbf{A}_{t,:}\|_\infty, k), \qquad k = \max(1, \lfloor \rho N \rfloor). \tag{3}$$

For each selected activation, define

$$u_{t,c} = \begin{cases} |\mathbf{A}_{t,c}|^{\alpha-2}, & \alpha > 2, \\ 1, & \alpha \le 2, \end{cases} \qquad w_{t,c} = \text{sg}\left(\frac{u_{t,c}}{\max\left(\frac{1}{kd} \sum_{i \in \mathcal{S}} \sum_{j=1}^{d} u_{i,j}, \epsilon \right)} \right), \tag{4}$$

where sg stops gradients and $\epsilon = 10^{-8}$. The auxiliary loss is

$$\mathcal{L}_{\text{EAP}} = \frac{1}{kd} \sum_{t \in \mathcal{S}} \sum_{c=1}^{d} w_{t,c} (\mathbf{A}_{t,c} - \widehat{\mathbf{A}}_{t,c})^2. \tag{5}$$

For $\alpha > 2$, large selected activations receive greater weight; for $\alpha \le 2$, EAP reduces to MSE on the selected tokens. We use $\rho = 0.05$. In the decoder, EAP is evaluated on MLP outputs with matched reference inputs; in the encoders, it is evaluated on calibrated block outputs. EAP introduces no learnable parameters and is used only during calibration.

4.3 Cross-modal Error Exchange

CME transfers quantization sensitivity across modalities: channels with large errors in one modality receive greater reconstruction weight in the others. For example, audio-sensitive channels guide the reconstruction of visual and text tokens. This exchange uses shared output-channel coordinates and requires no token alignment.

Let $\mathbf{Y}^m, \widehat{\mathbf{Y}}^m \in \mathbb{R}^{n_m \times d'}$ denote the full-precision and quantized outputs of tokens from modality m. We compute and normalize their channel-wise reconstruction errors:

$$e_c^m = \frac{1}{n_m} \left\| \mathbf{Y}_{:,c}^m - \widehat{\mathbf{Y}}_{:,c}^m \right\|_2^2, \qquad p_c^m = \frac{e_c^m}{\sum_{j=1}^{d'} e_j^m + \epsilon}, \tag{6}$$

where ϵ ensures numerical stability. For each target modality m, the mean error profile of the other available modalities provides cross-modal guidance:

$$\bar{p}_c^m = \text{sg}\left(\frac{1}{|\mathcal{M}| - 1} \sum_{\substack{n \in \mathcal{M} \\ n \neq m}} p_c^n \right), \tag{7}$$

where $\mathcal{M}$ is the set of available modalities and sg stops gradients. CME then weights the target modality's reconstruction error by this profile:

$$\mathcal{L}_{\text{CME}} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{c=1}^{d'} \bar{p}_c^m \left\| \mathbf{Y}_{:,c}^m - \widehat{\mathbf{Y}}_{:,c}^m \right\|_2^2. \tag{8}$$

Thus, each modality is reconstructed under guidance from the others, and the resulting gradients jointly optimize the shared transformation. CME is used only during calibration and is zero when fewer than two modalities are available. Sampling details are given in Appendix A.5.

The overall calibration objective is

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_E \mathcal{L}_{\text{EAP}} + \lambda_C \mathcal{L}_{\text{CME}}, \tag{9}$$

where λ_E and λ_C control the auxiliary penalties.

5 Experiments

5.1 Experimental Setup

Models and Benchmarks. We evaluate MAQuant on Qwen2.5-Omni-7B and Qwen2.5-Omni-3B. We use six benchmarks covering visual reasoning, audio understanding, omnimodal comprehension, and text-only commonsense reasoning: GQA (Hudson & Manning, 2019), ScienceQA (Lu et al., 2022), OmniBench (Li et al., 2024), MMAU (Sakshi et al., 2024), PIQA (Bisk et al., 2020), and HellaSwag (Zellers et al., 2019). Together, these benchmarks test whether low-bit quantization preserves both modality-specific perception and general language reasoning.

Baselines and Settings. We compare MAQuant with SmoothQuant (Xiao et al., 2023), FlatQuant (Sun et al., 2024), and MBQ (Li et al., 2025) across six benchmarks. We additionally compare full-scope adaptations of MQuant (Yu et al., 2025) and MASQuant (Hu et al., 2026) on Qwen2.5-Omni-3B using OmniBench, ScienceQA, and MMAU. These adaptations quantize the target linear layers in the vision encoder, audio encoder, connectors, and shared Thinker decoder, matching MAQuant's input-side quantization scope. Unless otherwise specified, we report W4A8 and W4A4 results, where weights are quantized to 4 bits and activations to 8 or 4 bits, respectively.

Implementation Details. We use VideoMME (Fu et al., 2025) as the calibration source, with video frames, available audio, and text prompts providing the decoder token groups. The calibration budget is 64 samples for each modality. For MAQuant, transformations and quantization parameters are optimized with reconstruction, EAP, and CME objectives. Additional hyperparameters and calibration details are provided in Appendix A.

5.2 Main Results

Table 1 summarizes the main results on Qwen2.5-Omni-3B/7B. MAQuant achieves the highest six-task average among the evaluated baselines for both model sizes and bit-width settings. Compared with generic quantization baselines, it better preserves multimodal understanding while maintaining competitive text-only reasoning, highlighting the value of modality-aware calibration for OLLMs.

Under W4A8, MAQuant stays close to the full-precision model on both backbones and performs well across visual, audio, and language tasks. The gains are evident on OmniBench and MMAU, covering omnimodal comprehension and audio understanding.

Under W4A4, MAQuant retains higher average accuracy than the baselines, although its own accuracy also decreases relative to W4A8. This behavior matches the design of MAQuant: MSL reduces modality-dependent mismatch, EAP protects sensitive activations, and CME improves consistency in the shared decoder space.

Comparison with MQuant and MASQuant. MAQuant achieves the highest three-task average at both precisions in Table 2, with gains over MQuant concentrated on OmniBench and MMAU. MQuant remains slightly stronger on ScienceQA under W4A8. The larger W4A4 advantage supports MAQuant's robustness when quantization covers both modality encoders and the shared decoder.

Table 1: Accuracy (%) on Qwen2.5-Omni models. Avg. is the mean of the six displayed tasks. NaN denotes numerical collapse during evaluation and contributes zero to this mean.

Model	Bitwidth	Method	GQA	ScienceQA	OmniBench	MMAU	PIQA	HellaSwag	Avg. ↑
Qwen2.5-Omni-7B	FP16	-	60.61	86.47	48.23	68.90	78.07	74.82	69.52
	W4A8	SmoothQuant	56.62	82.31	42.83	65.80	74.65	73.14	65.89
		FlatQuant	60.66	84.33	44.34	66.40	77.37	73.27	67.73
		MBQ	55.46	78.19	44.42	67.00	74.59	70.70	65.06
		MAQuant	**61.29**	**85.32**	**47.26**	**67.80**	**78.13**	**73.49**	**68.88**
	W4A4	SmoothQuant	NaN	NaN	NaN	NaN	25.86	49.24	12.52
		FlatQuant	59.08	80.47	45.31	65.52	**77.91**	**74.52**	67.14
		MBQ	NaN	NaN	9.91	NaN	50.49	27.19	14.60
		MAQuant	**59.85**	**84.43**	**45.84**	**67.80**	77.79	72.26	**68.00**
Qwen2.5-Omni-3B	FP16	-	61.95	80.76	45.39	66.80	75.41	68.36	66.45
	W4A8	SmoothQuant	56.77	74.39	34.69	65.20	71.46	66.35	61.48
		FlatQuant	60.78	77.79	43.57	65.40	69.52	67.10	64.03
		MBQ	**61.17**	78.76	44.45	63.60	**73.56**	68.25	64.97
		MAQuant	60.95	**79.67**	**44.51**	**66.80**	71.82	**68.30**	**65.34**
	W4A4	SmoothQuant	NaN	NaN	NaN	NaN	51.36	27.61	13.16
		FlatQuant	41.18	29.65	14.07	38.10	71.38	64.16	43.09
		MBQ	NaN	NaN	NaN	NaN	51.69	26.77	13.08
		MAQuant	**59.76**	**75.81**	**40.18**	**64.50**	**74.59**	**64.98**	**63.30**

Table 2: Comparison on Qwen2.5-Omni-3B. Avg. (3 tasks) averages only the three benchmarks shown.

Bits	Method	OmniBench	ScienceQA	MMAU	Avg. (3 tasks)
FP16	–	45.39	80.76	66.80	64.32
W4A8	MASQuant	39.44	74.60	56.70	56.91
	MQuant	36.89	**80.12**	50.20	55.74
	MAQuant	**44.51**	79.67	**66.80**	**63.66**
W4A4	MASQuant	14.07	36.90	37.13	29.37
	MQuant	32.69	75.56	48.10	52.12
	MAQuant	**40.18**	**75.81**	**64.50**	**60.16**

5.3 Ablation Studies

Effect of Each Component. We define the reconstruction-only control as optimizing weight/activation clipping with fixed transformations. MSL additionally trains the transformations, while EAP and CME augment the reconstruction loss. All variants use 64 VideoMME calibration samples, batch size 1, five epochs, and a base learning rate of 5×10^{-3}. In Figure 4(a), each component independently improves the two-task average. EAP and CME also improve upon MSL, and their combination performs best, supporting their complementary roles. Gains vary by task: CME leaves MMAU accuracy unchanged.

The EAP ablation uses $\alpha = 5$ and combines salient-token selection with magnitude-based weighting. We use $\rho = 0.05$ and $\lambda_E = 0.01$. CME uses $\lambda_C = 0.2$ and vision/audio/text token caps of 64/64/64.

5.4 Visualization and Analysis

Modality-specific Component Response. The transformation analysis in Section 3 motivates jointly calibrating the shared decoder coordinates. The largest text-only versus multimodal differences occur in MLP output rotations (Figure 2a); Appendix C.1 relates this pattern to the design.

Effect of EAP on Sensitive Activations. Figure 4(b) compares per-token maximum activation distributions with and without EAP. Appendix C.2 additionally measures residual outliers after the learned transformations, motivating salience-aware calibration alongside MSL.

Cross-modal Error Coupling. To quantify shared channel sensitivity, we select the top-error channels for each modality and compute their pairwise Jaccard overlap (Figure 2c). Appendix C.3 complements the overlap analysis in Section 3 with channel-level visualizations.

(a) Ablation under W4A4

Variant	MSL	EAP	CME	OmniBench	MMAU	Avg. (2 tasks)
Base	–	–	–	41.36	64.90	53.13
+ EAP	–	✓	–	45.49	65.50	55.49
+ CME	–	–	✓	43.36	64.90	54.13
+ MSL	✓	–	–	43.45	66.20	54.83
+ MSL + EAP	✓	✓	–	44.34	67.20	55.77
+ MSL + CME	✓	–	✓	44.36	67.40	55.88
MAQuant	✓	✓	✓	45.84	67.80	56.82

(b) Effect of EAP

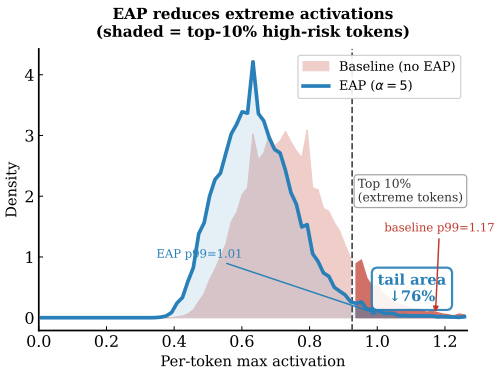


Figure 4: Ablation and EAP analysis on Qwen2.5-Omni-7B. (a) Component ablation under W4A4 with a common five-epoch calibration budget; Avg. (2 tasks) averages OmniBench and MMAU. EAP/CME averages use unrounded evaluation scores. (b) Effect of EAP on activation distribution.

Table 3: Prefill comparison on Qwen2.5-Omni-7B using an NVIDIA A100 and sequence length 2048. Quantized methods use W4A4. Speedup and memory reduction factors are relative to FP16 at the same batch size.

Batch	Method	Prefill (ms)	Tokens/s	Speedup	Peak mem. (GB)	Mem. reduction
1	FP16	191.77	10,679	1.00×	13.70	1.00×
1	QuaRot	89.2	22,960	2.15×	4.95	2.77×
1	SpinQuant	77.0	26,597	2.49×	5.42	2.53×
1	MAQuant	**75.5**	**27,126**	**2.54**×	5.21	2.63×
8	FP16	2140.77	7,653	1.00×	17.28	1.00×
8	QuaRot	748.5	21,889	2.86×	11.22	1.54×
8	SpinQuant	711.2	23,037	3.01×	9.00	1.92×
8	MAQuant	**704.2**	**23,266**	**3.04**×	9.82	1.76×

5.5 Deployment Efficiency

Storage. For Qwen2.5-Omni-7B, the estimated packed W4 model occupies 4.491 GiB under both W4A8 and W4A4. Including these transformations, the W4A4 deployment estimate is 4.530 GiB; the transformation overhead is 0.863% of the W4 model storage, computed from unrounded byte counts.

Prefill efficiency. We compare MAQuant with QuaRot and SpinQuant, two transformation-based quantization methods, on an NVIDIA A100 using the same inference backend, batch sizes, and sequence length. MAQuant has the lowest measured prefill latency at both batch sizes, with competitive peak memory (Table 3).

Benefits of full-architecture quantization. Quantizing the modality encoders, connectors, and shared decoder extends low-bit computation across the input processing path. With compatible fused kernels, adjacent quantized operators can reuse low-bit activations and reduce redundant dequantization, requantization, and intermediate memory transfers. This offers additional deployment savings beyond faster matrix multiplication.

6 Conclusion

In this paper, we study full-modality low-bit post-training quantization for omnimodal large language models. We identify three key challenges: modality-dependent quantization geometry, activation-sensitive distortion, and cross-modal error coupling. To address them, we propose MAQuant, which integrates modality-specific transformations, salience-aware activation preservation, and cross-modal error exchange. Experiments on Qwen2.5-Omni show that MAQuant improves low-bit quantization performance, especially under W4A4, while reducing storage and memory costs. Future work will explore data-efficient calibration, mixed-precision quantization, and hardware-aware deployment.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2025. URL https://arxiv.org/abs/2405.21075.

Ruihao Gong, Yifu Ding, Zining Wang, Chengtao Lv, Xingyu Zheng, Jinyang Du, Yang Yong, Shiqiao Gu, Haotong Qin, Jinyang Guo, et al. A survey of low-bit large language models: Basics, systems, and algorithms. *Neural networks*, pp. 107856, 2025.

Lulu Hu, Wenhu Xiao, Xin Chen, Xinhua Xu, Bowen Xu, Kun Li, and Yongliang Tao. Masquant: Modality-aware smoothing quantization for multimodal large language models, 2026. URL https://arxiv.org/abs/2603.04800.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Yizhang Jin, Jian Li, Tianjun Gu, Yexin Liu, Bo Zhao, Jinxiang Lai, Zhenye Gan, Yabiao Wang, Chengjie Wang, Xin Tan, et al. Efficient multimodal large language models: A survey. *Visual Intelligence*, 3(1):27, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Shiyao Li, Yingchun Hu, Xuefei Ning, Xihui Liu, Ke Hong, Xiaotao Jia, Xiuhong Li, Yaqi Yan, Pei Ran, Guohao Dai, et al. Mbq: Modality-balanced quantization for large vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4167–4177, 2025.

Yizhi Li, Yinghao Ma, Ge Zhang, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, et al. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*, 2024.

Haokun Lin, Haobo Xu, Yichen Wu, Jingzhi Cui, Yingtao Zhang, Linzhan Mou, Linqi Song, Zhenan Sun, and Ying Wei. Duquant: Distributing outliers via dual transformation makes stronger quantized llms. *Advances in Neural Information Processing Systems*, 37:87766–87800, 2024a.

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, 2024b.

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. *arXiv preprint arXiv:2405.16406*, 2024.

Pan Lu, Swaroop Mishra, Tony 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, 2022. URL https://arxiv.org/abs/2209.09513.

Minesh Mathew, Dimosthenis Karatzas, and CV 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, 2021.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024.

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. *arXiv preprint arXiv:2308.13137*, 2023.

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.

Yuxuan Sun, Ruikang Liu, Haoli Bai, Han Bao, Kang Zhao, Yuening Li, Jiaxin Hu, Xianzhi Yu, Lu Hou, Chun Yuan, et al. Flatquant: Flatness matters for llm quantization. *arXiv preprint arXiv:2410.09426*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, et al. Efficient large language models: A survey. *Transactions on Machine Learning Research*, 2023.

Changyuan Wang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. Q-vlm: Post-training quantization for large vision-language models. *Advances in Neural Information Processing Systems*, 37:114553–114573, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 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 *International conference on machine learning*, pp. 38087–38099. PMLR, 2023.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

JiangYong Yu, Sifan Zhou, Dawei Yang, Shuoyu Li, Shuo Wang, Xing Hu, Chen Xu, Zukang Xu, Changyong Shu, and Zhihang Yuan. Mquant: Unleashing the inference potential of multimodal large language models via static quantization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1783–1792, 2025.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9637–9662, 2024.

Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, et al. A survey on efficient inference for large language models. *arXiv preprint arXiv:2404.14294*, 2024.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

AI USE STATEMENT

In this work, we used generative AI tools (OpenAI Codex) for polishing the language of the manuscript and for assisting with modifications to the experimental code. We have not used generative AI tools for generating research ideas or writing the core technical content and claims of this paper. Additionally, we used generative AI tools for no other tasks. We have reviewed all AI-assisted work. Specifically, all AI-assisted code modifications were tested and verified by the authors, and the polished language was checked against the original technical content to ensure no factual or terminological distortion. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

A Additional Implementation Details

A.1 Calibration Data Construction

The calibration protocol uses VideoMME for the main configuration and component ablations, with 64 samples for each modality. Video frames supply vision inputs, available audio supplies audio inputs, and text prompts supply text inputs. Decoder modality masks identify token origins within the mixed sequence; visual tokens include the encoded video frames. Only modalities with valid tokens contribute to CME. OmniBench is reserved for evaluation rather than calibration. Calibration sample budgets, update budgets, and evaluation tasks are held fixed across the ablation variants as described in Section 5.3.

A.2 Quantization Settings

We evaluate W4A8 and W4A4: 4-bit weights with 8-bit or 4-bit activations in the targeted linear modules. Full-scope quantization covers the input-side vision encoder, audio encoder, connectors, and shared Thinker decoder. The scope excludes Talker, token2wav, and output heads; embeddings, normalization, softmax, and convolutional front ends retain their original precision. The original pretrained weights are frozen during calibration.

Following prior post-training quantization methods, we calibrate each target linear module using a frozen full-precision reference branch and a quantized branch. During calibration, gradients are back-propagated only through the quantized branch and the associated transformation or quantization parameters. The full-precision backbone remains fixed throughout the entire calibration process. After calibration, all quantized weights and modality-specific transformations are exported for inference.

A.3 Modality-Specific Learnable Transformation

MSL learns independent transformations in the vision and audio encoder branches and a shared transformation at each calibrated decoder site. Decoder modality masks partition reconstruction errors by token origin while retaining the mixed-modality forward context. Equation 2 decomposes the pooled MSE; Equation 8 adds explicit source-to-target error guidance. Gradients from these modality-conditioned objectives jointly update the common decoder transformation. Separate modality-specific decoder transformations or weight copies are not introduced.

Each linear layer uses the equivalent mapping $\mathbf{XW} = (\mathbf{XT})(\mathbf{T}^{-1}\mathbf{W})$. We pre-compute its transformed quantized weight offline and apply the corresponding activation transformation before low-bit linear computation.

A.4 Exponent-weighted Activation Penalty

EAP selects the top 5% of tokens by the infinity norm of their reference activations. It computes reconstruction loss only on these selected tokens. For $\alpha > 2$, each squared error is weighted by the corresponding reference magnitude raised to $\alpha - 2$, normalized by the mean weight over selected elements and detached from gradients. For $\alpha \leq 2$, all selected elements receive unit weight. Decoder EAP uses MLP output activations; encoder EAP uses calibrated block outputs. The EAP ablation uses $\alpha = 5$ and combines salient-token selection with magnitude-based weighting.

EAP is an auxiliary calibration loss with no additional trainable parameters or inference-time loss computation.

A.5 Cross-modal Error Exchange

Modality masks identify visual, audio, and text tokens. At most 64 tokens per modality are sampled from each sequence. Each modality's channel-error profile is computed from mean token errors and normalized across channels. The detached mean profile of the other modalities weights its reconstruction loss, transferring channel sensitivity across modalities. The loss sums token errors and averages over available modalities; it is skipped when fewer than two modalities are present. CME adds no inference-time loss computation.

Table 4: Accuracy drop in percentage points from FP16 on Qwen2.5-Omni-7B. Lower is better; negative values indicate improvement. Avg. drop averages all six tasks, using the main table's zero-score convention for numerical collapse. Values are computed from the displayed per-task scores.

Method	Bits	GQA	ScienceQA	OmniBench	MMAU	PIQA	HellaSwag	Avg. drop
SmoothQuant	W4A8	3.99	4.16	5.40	3.10	3.42	1.68	3.63
FlatQuant	W4A8	-0.05	2.14	3.89	2.50	0.70	1.55	1.79
MBQ	W4A8	5.15	8.28	3.81	1.90	3.48	4.12	4.46
MAQuant	W4A8	**-0.68**	**1.15**	**0.97**	**1.10**	**-0.06**	**1.33**	**0.64**
SmoothQuant	W4A4	60.61	86.47	48.23	68.90	52.21	25.58	57.00
FlatQuant	W4A4	1.53	6.00	2.92	3.38	**0.16**	**0.30**	2.38
MBQ	W4A4	60.61	86.47	38.32	68.90	27.58	47.63	54.92
MAQuant	W4A4	**0.76**	**2.04**	**2.39**	**1.10**	0.28	2.56	**1.52**

A.6 Optimization Details

MAQuant is calibrated on multimodal samples. The original full-precision model parameters are frozen, and only the transformation parameters and optional quantization-related parameters are optimized. Depending on the quantization backend, these quantization-related parameters may include scales, zero-points, or clipping thresholds.

We calibrate MAQuant layer by layer using $\mathcal{L}_{\text{rec}} + \lambda_E \mathcal{L}_{\text{EAP}} + \lambda_C \mathcal{L}_{\text{CME}}$. The component ablation uses five epochs, batch size 1, and a base learning rate of 5×10^{-3}. The loss runs train clipping parameters with fixed transformations: EAP-only uses $(\alpha, \lambda_E, \lambda_C) = (5, 0.01, 0)$, and CME-only uses $(\lambda_E, \lambda_C) = (0, 0.2)$. These runs use seed 0, AdamW for decoder and vision blocks, and SGD with momentum 0.9 for audio blocks, with cosine learning-rate decay.

A.7 Inference and Efficiency Measurement

For deployment experiments, we implement MAQuant with low-bit matrix multiplication kernels compatible with QuaRot-style INT4 GEMM execution. After calibration, each linear layer's transformed quantized weight is pre-computed offline. During inference, encoder branches apply their own learned activation transformations, while each shared decoder site applies its shared transformation across modalities.

Section 5.5 reports estimated packed-model storage, prefill latency, derived input-token throughput, and peak GPU memory. Prefill throughput is computed as batch size times input length divided by prefill latency; it does not measure autoregressive decoding or complete generation. The packed-weight estimate includes retained high-precision parameters, and auxiliary transformation tensors are counted separately before reporting the W4A4 total.

A.8 Compute Resources

The prefill comparison uses an NVIDIA A100 with a common inference backend, input length 2048, and batch sizes 1 and 8. Peak GPU memory is reported for the same prefill configurations. Packed-model storage is reported in GiB, and peak GPU memory during prefill is reported in GB.

B Additional Ablation Results

B.1 Robustness under Aggressive Activation Quantization

Table 4 complements Section 5.2 by reporting per-task accuracy drops from FP16 on Qwen2.5-Omni-7B. W4A8 and W4A4 keep weights at 4 bits and use separately calibrated activation precisions; negative drops indicate improvements over FP16.

B.2 EAP under W4A4

Four-bit activation quantization has fewer representable levels than eight-bit quantization, making reconstruction more sensitive to large activation ranges. EAP focuses the auxiliary loss on high-magnitude tokens and, for $\alpha > 2$, emphasizes large elements within those tokens. The W4A4 ablation

shows a 0.94-point improvement when EAP is added to MSL on the two-task average; it does not isolate EAP's relative benefit under W4A8.

B.3 Why CME is Important for Omnimodal Models

Section 4.3 describes how other modalities' error profiles guide target reconstruction. Adding CME to MSL improves the two-task average by 1.05 points in Figure 4(a), complementing the channel-sensitivity analysis in Appendix C.3.

C More Analysis and Visualizations

C.1 Modality-specific Response of Different Components

The comparison between text-only and multimodal calibration in Figure 2a shows stronger divergence in MLP-related rotations than in attention output transformations, with diagonal scaling changing less. This locates the largest calibration-dependent changes in feed-forward coordinates and motivates optimizing the shared transformations described in Section 4.1.

C.2 Analysis of Sensitive Activations

To examine whether outliers persist after the learned transformations, we measure decoder MLP projection inputs on a calibrated Qwen2.5-Omni-3B W4A4 checkpoint using 16 VideoMME samples (22,034 tokens). At the down-projection input of layer 35 (zero-indexed), sampled absolute activations have a median of 0.865 and a 99.9th percentile of 11.09. The top 1% and top 5% of tokens, ranked by maximum absolute activation, account for 43.1% and 54.8% of the total activation-quantization squared error, respectively. Percentiles are estimated from sampled activation elements; error shares use the collected tokens. These measurements show residual outliers in the transformed space and motivate targeted reconstruction penalties. They characterize a calibrated checkpoint rather than isolating the causal effect of MSL or EAP.

C.3 Per-channel Quantization Error across Modalities

To further understand the modality-dependent error patterns under low-bit activation quantization, we visualize the per-channel quantization error of text, vision, and audio tokens across representative layers. As shown in Figure 5, the error distribution is highly non-uniform across channels and varies substantially across modalities. Vision tokens generally exhibit denser high-error channels, while text and audio show sparser but still prominent error spikes. This observation suggests that different modalities do not share identical quantization-sensitive directions, motivating modality-conditioned error weighting within shared layers.

Meanwhile, several high-error regions appear repeatedly across modalities and layers, indicating that some fragile channels are shared after heterogeneous modality representations are mapped into the common decoder space. These channel-level views complement the aggregate overlap in Figure 2c.

C.4 Limitations and Future Directions

MAQuant depends on representative calibration data and currently uses fixed bit-widths. Evaluation is limited to the Qwen2.5-Omni family and input-side understanding tasks; video benchmarks and speech-generation modules are not evaluated. Deployment measurements cover prefill, while decoding and full-generation throughput remain to be measured. The small prefill advantage over SpinQuant requires repeated measurements to assess its stability, and eliminating intermediate precision conversions requires additional kernel fusion.

Future work will explore broader model families, data-efficient calibration, mixed-precision allocation, and fused low-bit execution across the encoder–decoder architecture.

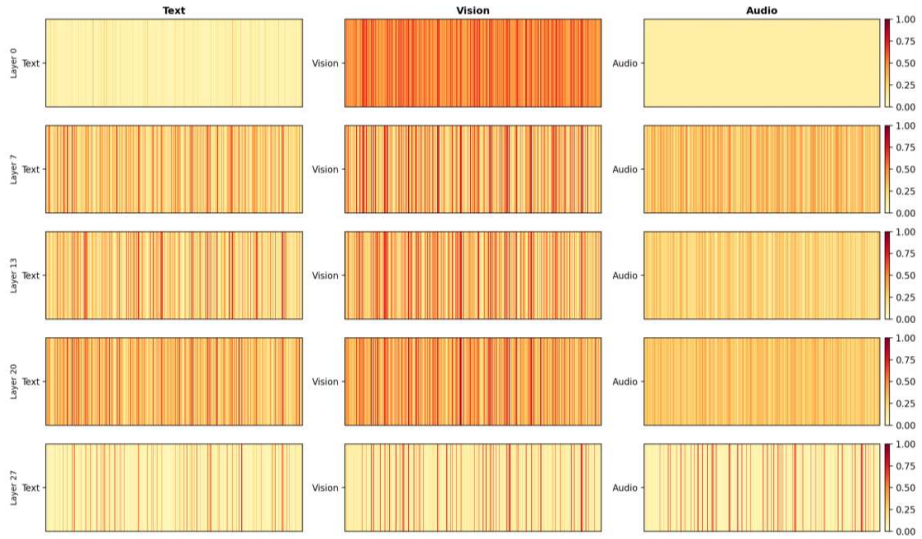


Figure 5: Per-channel quantization error of text, vision, and audio tokens under A4 activation quantization. Each row is normalized by its own maximum value and sorted by channel index. We visualize representative layers from shallow to deep layers. The results show that quantization errors are highly channel-dependent and modality-dependent: vision tokens contain denser high-error channels, while text and audio exhibit sparser but clear error spikes. The repeated high-error patterns across modalities further indicate cross-modal error coupling in the shared decoder space, motivating the proposed CME objective.