MORPHOQUANT: MORPHOLOGY-DRIVEN QUANTIZATION FOR OMNI-MODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Conventional Post-Training Quantization (PTQ) methods struggle with 4-bit Omni-modal Large Language Models (OLLMs) due to the extreme distribution heterogeneity and disparate outlier patterns across modalities. To address this, we propose MorphoQuant, a morphology-directed PTQ framework engineered to preserve cross-modal morphology and mitigate outlier loss. Specifically, we introduce Morphology-Driven Sparse Compensation (MDSC), which identifies high-dispersion activation channels and retains large activation residuals into a channel-wise sparse computational branch. Complementing this, we propose Morphology-Preserving Optimization (MPO), which optimizes the activation clipping boundary using a reconstruction objective with a cosine-similarity term; the boundary also determines which residuals enter the compensation branch. We evaluate the method on Qwen2.5-Omni-3B and 7B using vision-language, video-language, and audio-language benchmarks, and on InternVL2.5-8B using vision-language benchmarks. Under the evaluated W4A4 setting, MorphoQuant maintains non-collapsed task performance where the reproduced W4A4 multimodal baselines frequently fail. Alternative activation formats further improve results in some settings. These findings suggest that fine-grained residual selection coupled with clipping-boundary optimization is a promising approach to low-bit quantization of omni-modal models.

1 INTRODUCTION

Omni-modal Large Language Models (Omni-modal LLMs, OLLMs) (Wang et al., 2024c; Xu et al., 2025b; Team et al., 2024) have recently achieved unprecedented breakthroughs across a myriad of complex reasoning tasks, ranging from visual question answering (Jin et al., 2025; Ghosh et al., 2024; Caffagni et al., 2024; Xu et al., 2023; Bhuyan et al., 2025) to audio/video comprehension. By aligning diverse sensory inputs, OLLMs exhibit cognitive paradigms that closely mirror human perception, primarily driven by their large-scale multimodal datasets (Liu et al., 2025; Li et al., 2024; Yin et al., 2023) and massive parameter scales. However, their exorbitant memory demands and computational cost pose severe impediments to real-world deployment, particularly on resource-constrained or edge devices. Furthermore, the auto-regressive nature of OLLM response generation necessitates extensive memory bandwidth for repetitive forward passes, severely bottlenecking inference latency. Consequently, compressing model size and mitigating memory consumption without compromising reasoning capability has emerged as a paramount challenge.

To alleviate these computational bottlenecks, various model compression techniques have been explored, including pruning (Ye et al., 2025), low-rank decomposition (Chen et al., 2024), effective architecture design (Bai et al., 2024), and quantization (Xiao et al., 2023; Lin et al., 2024). Among these, Post-Training Quantization (PTQ) (Xie et al., 2024; Wang et al., 2024a; Yu et al., 2025; Wang et al., 2024b;a) stands out as a highly practical paradigm that maps floating-point representations into low-bit integers using a minuscule calibration dataset, which significantly accelerates inference while circumventing the prohibitive costs of full retraining. However, pushing PTQ to the ultra-low W4A4 extreme for true *omni-modal* models which simultaneously ingest audio, video, images, and text like Qwen2.5-Omni (Xu et al., 2025a), poses a formidable challenge.

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.

The inherent dynamics of these diverse sensory inputs introduce highly heterogeneous cross-modal information, giving rise to disparate outlier patterns and severe distribution discrepancies. Existing quantization frameworks (Huang et al., 2024b) fail to adequately address this cross-modal distribution chasm due to applying uniform quantization functions across these heterogeneous modalities, leading to catastrophic outlier truncation with substantial discretization errors for dense inliers. Existing methods address outliers through distribution transformations or higher-precision preservation of selected values. These strategies motivate finer-grained control over activation residual compensation. (Xiao et al., 2023; Dettmers et al., 2022; 2023b) Conversely, mixed-precision workarounds (Zhao et al., 2024b) protect outliers by assigning higher precision to selected features. Their accuracy and computational cost depend on the selection strategy and its implementation, motivating finer-grained control over the values retained in the correction branch. Therefore, discovering a mechanism to simultaneously tackle disparate outlier patterns and align heterogeneous cross-modal distributions remains a critical, unresolved challenge.

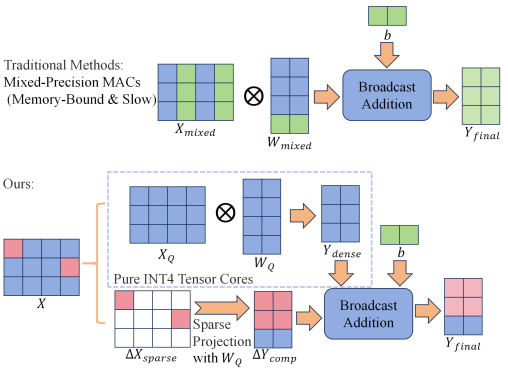

Figure 1: Comparison with traditional methods. Traditional mixed-precision methods necessitate memory-interleaved operations (blue-INT4, green-FP16), suffering from fragmented memory access and severe overhead. Our MDSC strictly decouples the computation, dividing the activation into the vast majority in pure INT4 Tensor Cores and the sparse outliers (red) undergoing a lightweight projection.

We address this question with MorphoQuant, a PTQ framework that combines residual compensation with clipping-boundary optimization. Building on the use of higher-precision correction for outliers, we introduce Morphology-Driven Sparse Compensation (MDSC). During calibration, a channel-wise dispersion score identifies candidate channels with pronounced activation ranges relative to their aggregate magnitudes. Within these channels, an element-wise threshold selects large-magnitude activations whose quantization residuals are retained in a compensation branch. This two-stage selection provides finer granularity than protecting all values in the selected channels. We further introduce Morphology-Preserving Optimization (MPO) to couple the quantization range with residual selection. MPO optimizes the activation clipping boundary using an ℓ_p reconstruction loss and a cosine-similarity loss. The element-wise residual threshold is defined relative to the clipping boundary, so changing the boundary also changes which residuals are compensated. The candidate channel mask remains fixed after calibration, while the element-wise selection depends on the activation values. This coupling allows the clipping boundary to be calibrated for the combined quantized representation and residual correction.

We evaluate MorphoQuant on Qwen2.5-Omni-3B and 7B using vision-language, video-language, and audio-language benchmarks, and on InternVL2.5-8B using vision-language benchmarks. Under the evaluated W4A4 settings, MorphoQuant achieves higher accuracy than the reproduced multimodal baselines on most reported tasks. We additionally evaluate alternative activation formats and progressively assess the combined method components. Our main contributions are summarized as follows:

- We present a systematic investigation into the ultra-low bit (W4A4) quantization bottlenecks of all-in-one OLLMs, successfully addressing a critical gap in current omni-modal compression research.
- We propose MDSC, which combines channel-wise dispersion-based selection with element-wise thresholding to retain selected activation quantization residuals.
- We introduce MPO, which couples residual selection with clipping-boundary optimization through a reconstruction objective incorporating magnitude and directional errors.
- We evaluate the resulting framework across visual, video, and audio understanding tasks on Qwen2.5-Omni, with additional vision-language evaluation on InternVL2.5, demonstrating its effectiveness under the evaluated low-bit quantization settings.

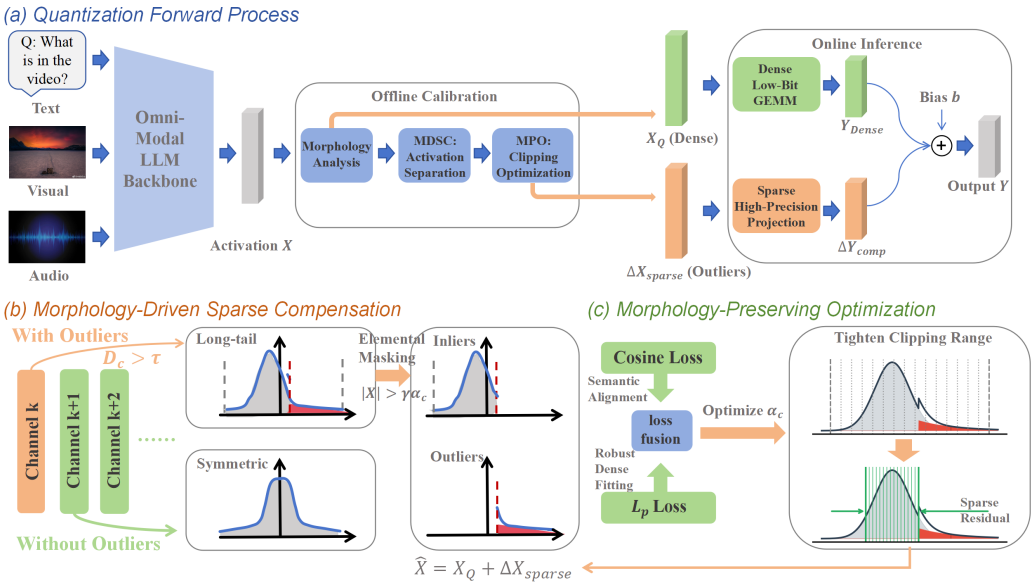

Figure 2: Overview of our omni-modal quantization framework. **(a) Quantization Forward Process** : Offline calibration extracts quantization parameters, enabling online inference to be expressed as a dense quantized matrix multiplication and an additional residual correction. **(b) Morphology-Driven Sparse Compensation**: A dual-grained approach (channel-wise $\mathcal{D}_c$ and elemental thresholding $\gamma\alpha_c$) precisely extracts extreme outliers, leaving symmetric inliers. **(c) Morphology-Preserving Optimization**: A composite loss ($\mathcal{L}_p$ & $\mathcal{L}_{cos}$) guides the dynamic tightening of the clipping boundary α_c, co-optimizing the quantization grid and sparse residuals to preserve cross-modal semantics.

2 Related Works

2.1 Quantization of Language Models

PTQ (Yao et al., 2022; Wu et al., 2020; Huang et al., 2024a) of single-modal LLMs has been widely studied to reduce memory footprint and inference cost while preserving accuracy. SmoothQuant (Xiao et al., 2023) proposes a training-free framework that smooths channel-wise activation outliers into weights via per-channel scaling, enabling hardware-friendly W8A8 quantization on very large LLMs with minimal performance degradation. QLoRA (Dettmers et al., 2023a) focuses on memory-efficient fine-tuning by storing pretrained weights in a 4-bit NormalFloat (NF4) format, combining double quantization of scales with low-rank adapters trained in higher precision, which allows models with tens of billions of parameters to be finetuned on a single GPU. AWQ (Lin et al., 2024) targets on-device deployment and uses activation statistics to identify a small set of salient channels; it applies equivalent per-channel scaling to protect roughly 0.1–1% of weights, achieving effective W4 weight-only quantization. Outlier Suppression+ (OS+) (Wei et al., 2023) systematically analyzes activation outliers and finds that they are concentrated in a few channels and strongly asymmetric across channels.

More recent methods introduce learnable transformations to further improve low-bit PTQ. OmniQuant (Shao et al., 2023) combines learnable weight clipping (LWC) with learnable equivalent transformations (LET) on activations, optimized in a block-wise manner to calibrate quantization parameters across bit-widths including W4A4. LRQuant (Zhao et al., 2024a) learns smoothing parameters using an objective based on the negative logarithm of cosine similarity between full-precision and quantized block outputs. Our MPO builds on calibration-based optimization, combining an ℓ_p reconstruction term with a cosine-similarity term to optimize the activation clipping boundary. This boundary also controls the element-wise residual selection threshold. SpinQuant (Liu et al., 2024) learns orthogonal rotations on residual streams, attention projections, and KV caches, parameterized on the Stiefel manifold, to redistribute outliers into more isotropic directions and improve

W4A4KV4 quantization of text-only LLMs. These approaches largely focus on unimodal transformers and provide the foundation for low-bit compression of large language models.

2.2 Quantization of MLLMs or VLMs

Quantization of multimodal large language models (MLLMs) or vision–language models (VLMs) (Wang et al., 2025; Shinde et al., 2025) has recently attracted increasing interest as these models are deployed in broader application scenarios. QSLAW (Xie et al., 2024) studies on quantization-aware adaptation of MLLMs and introduces quantization-aware scale learning with a multimodal warmup schedule, where group-wise weight scales are learned jointly with visual instruction tuning. For large vision–language models, Q-VLM (Wang et al., 2024a) proposes a PTQ framework that explicitly considers cross-layer dependency of discretization errors; it employs Cross-layer Dependency Mining (CDM) to group layers into blocks based on activation entropy and Visual Encoder Optimization to reshape visual activations by jointly optimizing entropy and reconstruction error. MBQ (Modality-Balanced Quantization) (Li et al., 2025) observes that language and vision tokens exhibit different sensitivities to quantization and introduces gradient-based sensitivity indicators to reweight reconstruction losses for different modalities within channel-wise equalization, along with a mean-absolute-error objective that better correlates with loss change. MQuant (Yu et al., 2025) further tailors PTQ to MLLMs by identifying challenges in prefill latency, modality-wise distribution gaps, and rotation-induced outliers; it proposes Modality-Specific Static Quantization (MSQ) with distinct static scales for visual and textual tokens, Attention-Invariant Flexible Switching (AIFS) for token reordering under preserved attention, and Rotation Magnitude Suppression (RMS) to handle channels affected by Hadamard-based rotations. FreeAct (Liu et al., 2026) accounts for token-type-dependent activation distributions through distinct activation transformations while maintaining a shared transformation for weights. These studies motivate quantization strategies that account for differences between modalities. Our work investigates activation residual compensation on Qwen2.5-Omni, with evaluations covering vision-language, video-language, and audio-language tasks, and additional vision-language evaluation on InternVL2.5.

2.3 Mixed-Precision Quantization and Sparse Outlier Preservation

Mixed-precision quantization protects selected values using higher precision. LLM.int8() (Dettmers et al., 2022) separates outlier feature dimensions into a higher-precision matrix multiplication. QUIK (Ashkboos et al., 2024) quantizes most weights and activations to 4 bits while retaining selected outlier components in higher precision, with corresponding GPU kernels. ATOM (Zhao et al., 2024b) combines mixed-precision quantization with reordering and operator fusion, including INT8 computation for selected outlier channels. For weight compression, SpQR (Dettmers et al., 2023b) isolates outlier weights in higher precision, while SqueezeLLM (Kim et al., 2023) combines a quantized dense representation with sparse storage of outlier and sensitive weights.

3 Approach

In this section, we present our post-training quantization framework tailored for OLLMs. We begin by reviewing standard PTQ and then explicitly identify the cross-modal distribution chasm. To address this, we introduce the MDSC mechanism to precisely retain selected activation residuals, followed by MPO strategy to maximize the representational fidelity of the inlier distribution.

3.1 Preliminaries: Uniform Quantization and Its Limitations

For a standard linear quantization process, given a full-precision activation tensor $\mathbf{X}$ and weight matrix $\mathbf{W}$, the uniform quantization maps the continuous values into a k-bit integer grid. The quantized activation $\mathbf{X}_Q$ and weight $\mathbf{W}_Q$ are formulated as:

$$\begin{aligned} \mathbf{X}_Q &= \text{clip}\left(\left\lfloor \frac{\mathbf{X}}{s_x} \right\rfloor, Q_{min}, Q_{max}\right) \cdot s_x, \\ \mathbf{W}_Q &= \text{clip}\left(\left\lfloor \frac{\mathbf{W}}{s_w} \right\rfloor, Q_{min}, Q_{max}\right) \cdot s_w, \end{aligned} \tag{1}$$

where s_x and s_w denote the quantization scaling factors, and $\lfloor \cdot \rceil$ represents the round-to-nearest operator and $[Q_{min}, Q_{max}]$ defines the representational boundaries. The above equations describe uniform quantization. In our implementation, weights are quantized using NF4 through bitsandbytes. In the following formulation, W_Q denotes the reconstructed weight matrix represented by the quantized weights. The fundamental quantization error $\Delta\mathbf{X}$ consists of two mutually exclusive components: clipping error $\mathcal{E}_{clip}$ and rounding error $\mathcal{E}_{round}$:

$$\Delta\mathbf{X} = \mathbf{X} - \mathbf{X}_Q = \underbrace{\mathbf{X} - \mathbf{X}_c}_{\mathcal{E}_{\text{clip}}\text{: Clipping error}} + \underbrace{\mathbf{X}_c - \mathbf{X}_Q}_{\mathcal{E}_{\text{round}}\text{: Rounding error}} . \tag{2}$$

Here, $\mathbf{X}_c$ represents the clipped activation, and α_c is the clipping threshold derived from the scaling factor s_x. The objective of conventional PTQ is typically formulated to minimize the output reconstruction error with ℓ_2-norm:

$$\min_{s_x, s_w} \|\mathbf{W}\mathbf{X} - \mathbf{W}_Q\mathbf{X}_Q\|_2^2 . \tag{3}$$

Specifically, this ℓ_2-norm objective is inherently sensitive to the **disparate outlier patterns and distribution characteristics** prevalent in OLLMs, which precipitates a critical dilemma for selection of quantization threshold α_c under extreme low-bit constraints. Expanding the quantization range to accommodate outliers inflates the step size, thereby amplifying rounding errors $\mathcal{E}_{\text{round}}$ for dense inliers; conversely, tightening the range to preserve inlier precision results in severe clipping errors $\mathcal{E}_{\text{clip}}$ for critical outliers. Both scenarios substantially increase the quantization error $\Delta\mathbf{X}$.

3.2 Motivation: Morphological Deformation and Displacement

To elucidate the root causes of the threshold selection dilemma, we conduct a comprehensive empirical analysis of the activation distributions within omni-modal architectures. Unlike conventional VLMs Xie et al. (2024); Wang et al. (2024a); Yu et al. (2025) that process bimodal inputs, omni-modal systems integrate heterogeneous sensory data into a unified backbone:

$$\begin{aligned} \mathbf{X}_{omni} &= \text{Concat}\left(\text{Proj}_m(\text{Enc}_m(\mathbf{I}_m))\right), \\ m &\in \{\text{Audio, Video, Image, Text}\}. \end{aligned} \tag{4}$$

This architectural convergence leads to a catastrophic representation collapse under extreme low-bit quantization. Specifically, our empirical analysis reveals three critical phenomena (see the supplementary materials for image visualizations):

- **Significant Outlier Tails:** In model activation histograms, specific channels exhibit extreme outlier values that dominate the dynamic range, complicating optimal threshold selection.
- **Omni-Modal Morphological Disparity:** Dissecting the unified distribution into isolated modalities reveals fundamentally disparate magnitudes and outlier morphologies. This forces a scaling dilemma: global uniform scaling factors either severely clip crucial high-magnitude features or drown subtle features in rounding noise.
- **Modal-Boundary Morphological Degradation:** Attention map visualizations corroborate that prominent quantization degradation occurs prominently at boundaries between modalities, disrupting omni-modal reasoning.

Based on these observation, we conclude that conventional methods appear substantial discretization errors for omni-modal models under extreme low-bitwidths, as the uniform quantization functions fail to accommodate both the long tails and the dense inlier bodies simultaneously.

3.3 Morphology-Driven Sparse Compensation

To address the pivotal challenge of uniform quantization, we propose the MDSC mechanism. The core principle is to strategically shift the clipping loss of activations into channel-wise sparse compensation. This preserves critical outlier information while maintaining accurate quantization for the majority of inliers.

Specifically, we first evaluate the dispersion metric for each channel to identify the minuscule fraction of activations demanding protection. The dispersion score $\mathcal{D}_c$ of the input $\mathbf{X}_c$ in channel c is formulated as:

$$D_c = \frac{\max(X_c) - \min(X_c)}{\log\left(1 + \frac{1}{N}\|X_c\|_1 + \epsilon\right)}, \qquad \epsilon > 0, \tag{5}$$

Here, X_c contains the N calibration activation values for channel c. The numerator measures the activation range, while $\|X_c\|_1/N$ measures the mean absolute activation magnitude. We apply $\log(1 + \cdot)$ to compress the scale of the denominator and add a positive ϵ to ensure that it remains strictly positive. The resulting score is used as a heuristic for selecting channels with large activation ranges relative to their average magnitudes.

Once the dispersion metric for all channels is computed, we evaluate their statistical distributions to partition the channels, striking an optimal accuracy-efficiency trade-off. We identify salient channels based on whether their dispersion metric passes 2σ (a heuristic threshold based on the mean and standard deviation of the channel scores) either at the local layer level or the global network level. For these selected channels, their severe truncation residuals are shielded from the destructive k-bit grid and instead rerouted into the sparse compensation term. Specifically, we generate a channel-wise binary mask $\mathbf{m} \in \{0, 1\}^{C_{in}}$ to pinpoint the unbalanced channels with long tails:

$$m_c = \mathbb{I}(\mathcal{D}_c > \tau_l \ \vee \ \mathcal{D}_c > \tau_g), \tag{6}$$

where $\tau_l = \mu_l + 2\sigma_l$ and $\tau_g = \mu_g + 2\sigma_g$ serve as the local and global thresholds, respectively. Here, (μ_l, σ_l) are the mean and standard deviation of the dispersion metrics within the current layer, and (μ_g, σ_g) are the global mean and standard deviation across all layers, and $\mathbb{I}(\cdot)$ denotes the indicator function for binary masks. During actual operation, the mask is fixed during the offline calibration phase and will not change during the inference phase. By decoupling long-tailed activations, MDSC safeguards the vital magnitude of outliers within the high-precision sparse compensation while maintaining high-precision discretization.

To restrict residual correction to large-magnitude activations within the selected channels, we additionally apply an element-wise threshold. For the identified channels, only elements whose magnitudes fall significantly outside the quantization bounds are included in the sparse branch. The final sparse residual tensor ΔX_{sparse} is formulated as:

$$\mathbf{\Delta X_{sparse}} = (\mathbf{X} - \mathbf{X}_Q) \odot \mathbf{m} \odot \mathbb{I}\Big(|X| > \gamma \cdot \alpha_c\Big), \tag{7}$$

where α_c is the clipping boundary of the current channel. The correction retains the full quantization residual at selected positions, including both clipping and rounding errors. The channel mask is determined during calibration and fixed during inference, while element-wise selection depends on the input activation values. The morphology-decoupled forward pass is thus:

$$\mathbf{Y} = \underbrace{\mathbf{W}_Q\mathbf{X}_Q}_{\text{Dense Inliers}} + \underbrace{\mathbf{W}_Q\mathbf{\Delta X}_{sparse}}_{\text{Sparse Compensation}} + \mathbf{b}.$$

Computational Structure. The forward computation consists of a dense quantized term and an additional residual correction term. Both terms use the same quantized weight representation. This decomposition specifies the numerical computation, while its runtime cost depends on the residual density, data representation, and kernel implementation. We therefore do not infer hardware acceleration from the decomposition alone.

3.4 Morphology-Preserving Optimization

Building on the compact, zero-symmetric distribution established by MDSC, we propose the MPO strategy to refine the quantization grid. Following the previous section, data with outliers removed exhibits a highly regular zero-symmetric distribution, enabling us to aggressively tighten the clipping boundary α_c optimization strategy based on the intrinsic morphological traits of the activation data: *zero-symmetric inlier* and *long-tailed outlier*.

We use a symmetric clipping interval $[-\alpha_c, \alpha_c]$ and select the boundary through grid search. An ℓ_p reconstruction loss controls the relative emphasis on errors of different magnitudes:

$$\mathcal{L}_p = \left(\frac{1}{N} \sum_{i=1}^{N} \left| X_{FP}^{(i)} - \hat{X}_Q^{(i)}(\alpha_c) \right|^p \right)^{\frac{1}{p}}, \tag{8}$$

where $\hat{\mathbf{X}}_Q = \mathbf{X}_Q + \Delta \mathbf{X}_{sparse}$ is the dequantized value together with compensation, and N is the total number of elements. Furthermore, as the directional alignment of high-dimensional feature vectors is crucial for omni-modal semantic interaction in OLLMs, we introduce a Cosine Similarity Loss $\mathcal{L}_{cos}$ to strictly preserve the spatial angle between the full-precision and quantized activations:

$$\mathcal{L}_{cos} = 1 - \frac{1}{M} \sum_{j=1}^{M} \frac{\mathbf{X}_{FP}^{(j)} \cdot \hat{\mathbf{X}}_Q^{(j)}(\alpha_c)}{\left\| \mathbf{X}_{FP}^{(j)} \right\|_2 \left\| \hat{\mathbf{X}}_Q^{(j)}(\alpha_c) \right\|_2}, \tag{9}$$

where M denotes the sequence length, and $\mathbf{X}^{(j)}$ represents the feature vector of the j-th token. Then, the overall optimization objective is formulated as:

$$\min_{\alpha_c} \mathcal{L}_{total} = \mathcal{L}_p + \lambda_{cos} \mathcal{L}_{cos}, \tag{10}$$

Here, $\lambda_{\cos}$ balances the reconstruction and cosine terms. We select α_c using the per-channel grid search described below. Since the residual threshold is $\gamma\alpha_c$, each candidate boundary also changes the element-wise residual selection. We set $p = 3$ in our experiments, placing greater relative emphasis on larger reconstruction errors. The cosine-similarity term complements this magnitude-based objective by encouraging directional agreement between the original and reconstructed activations. Since the residual selection threshold is $\gamma\alpha_c$, each candidate clipping boundary also determines which activation residuals are compensated.

Per-Channel Clipping-Boundary Search. We optimize clipping boundaries through independent per-channel grid search over a shared set of clipping ratios $\mathcal{R}$. For each channel c, a candidate ratio $r \in \mathcal{R}$ defines the clipping boundary $\alpha_c(r) = r a_c$, where a_c denotes the reference magnitude of that channel. For each candidate, we recompute the quantized activations and the residual correction with the threshold $\gamma\alpha_c(r)$. Each channel independently selects the ratio that minimizes its calibration objective. The channel-selection mask remains fixed throughout the search.

4 Experiments

In this section, we evaluate MorphoQuant on Qwen2.5-Omni and InternVL2.5 under low-bit quantization settings. We describe the experimental setup, compare accuracy across the evaluated configurations, and report a progressive ablation of the method components.

4.1 Experimental Setup

Models and Benchmarks. We deploy our PTQ framework on the pre-trained OLLM/VLM models. To rigorously validate the robustness of our method across the entire omni-modal spectrum, we deliberately depart from the conventional practice of evaluating solely on vision-language tasks. Instead, our evaluation suite spans highly diverse modalities: **ScienceQA** Lu et al. (2022) and **MMMU** Yue et al. (2024) for complex vision-language reasoning, **Video-MME** Fu et al. (2025) for long-context video comprehension, and **AIR-Bench** Yang et al. (2024) for dense audio signal understanding. These benchmarks evaluate vision-language, video-language, and audio-language understanding. Standalone text-only capabilities are not evaluated in this study.

Baselines for Comparison. We benchmark our framework against leading quantization methodologies. To fully demonstrate the efficacy of our approach on extremely low-bit activations, we evaluate both W4A16, W8A8 and W4A4 configurations. For the W4A16 setting, we compare with widely adopted baselines including bitsandbytes NF4 and AWQ Lin et al. (2024). For W8A8/W4A4 setting, we compare against recent multimodal quantization techniques such as MBQ Li et al. (2025) and MQuant Yu et al. (2025).

Model	Bit-width	Method	Vision-Language		Video	Audio
			ScienceQA	MMMU	Video-MME	AIR-Bench
Qwen2.5-Omni-3B	W4A16	bitsandbytes NF4	75.88	43.56	54.78	64.12
		SmoothQuant	49.68	24.41	0.00	49.99
		AWQ	78.00	36.70	57.07	63.91
	W8A8	MBQ	79.01	44.56	57.67	**65.94**
		MQuant	78.52	44.33	0.37	0.00
		Ours	77.98	**45.44**	**58.10**	65.66
		Ours + HiF8	**79.32**	43.22	56.93	65.42
	W4A4	MBQ	9.88	0.00	0.00	0.00
		MQuant	0.00	0.00	0.00	0.00
		Ours	72.93	42.44	50.11	62.33
		Ours + HiF4	**76.68**	**43.67**	**56.15**	**65.11**
Qwen2.5-Omni-7B	W4A16	bitsandbytes NF4	84.41	49.44	59.59	68.37
		SmoothQuant	72.62	30.00	33.38	43.46
		AWQ	84.01	47.80	58.78	66.85
	W4A4	MBQ	12.26	0.00	9.63	0.00
		MQuant	0.00	0.00	0.00	0.00
		Ours	81.84	**51.11**	54.11	65.06
		Ours + HiF4	**83.73**	49.56	**57.93**	**67.54**
InternVL2.5-8B	W4A4	MBQ	54.85	25.33	–	–
		QuaRot	52.58	42.19	–	–
		FreeAct	18.49	**42.44**	–	–
		Ours	82.79	**42.44**	–	–
		Ours + HiF4	**83.40**	**42.44**	–	–

Table 1: Main results of OLLMs/VLMs quantization across various multimodal benchmarks. We report accuracy (%) for ScienceQA, MMMU, Video-MME and AIR-Bench. **Bold** indicates the highest score within each model and bit-width group, with ties highlighted. The "0.00" in the table indicates that the model produced garbled characters.

Implementation Details. Our method uses NF4 weight quantization implemented with bitsandbytes, and Morphology-Driven Sparse Compensation (MDSC) and Morphology-Preserving Optimization (MPO) are applied to activation tensors, for vision encoder and the OLLM backbone. During the calibration and search phase, we sample a minuscule calibration set comprising 256 multimodal samples. In addition, we tested the compatibility of our method with different data types using HiF4 Luo et al. (2026) and HiF8 Luo et al. (2024) for activation as examples. For more detailed information, see the supplementary materials.

4.2 Results and Analysis

Performance under W4A4. Table 1 reports results on the evaluated omni-modal and vision-language tasks. On Qwen2.5-Omni-3B, MorphoQuant achieves scores of 72.93, 42.44, 50.11, and 62.33 on ScienceQA, MMMU, Video-MME, and AIR-Bench, respectively. The reproduced MBQ and MQuant configurations frequently produce degenerate outputs under W4A4. On InternVL2.5-8B, MorphoQuant improves ScienceQA accuracy over the evaluated baselines and matches FreeAct on MMMU. These comparisons are specific to the evaluated configurations and baseline adaptations. Failures are also observed for MQuant on Video-MME and AIR-Bench under W8A8, indicating that the observed failures are not restricted to W4A4.

Comparison with W4A16 Baselines. The results show task-dependent differences between MorphoQuant and the W4A16 configurations. On Qwen2.5-Omni-3B, MorphoQuant with HiF4 exceeds the bitsandbytes NF4 baseline by 0.80, 1.37, and 0.99 percentage points on ScienceQA, Video-MME, and AIR-Bench, respectively. However, its ScienceQA and Video-MME scores remain below the corresponding AWQ entries. On Qwen2.5-Omni-7B, standard MorphoQuant achieves 51.11 on MMMU compared with 49.44 for bitsandbytes NF4, while scoring lower on the other three benchmarks. Thus, W4A4 is competitive with selected W4A16 configurations, but does not consistently outperform them across tasks.

Effect of Activation Formats. Using HiF4 increases Video-MME scores from 50.11 to 56.15 on Qwen2.5-Omni-3B and from 54.11 to 57.93 on the 7B model. The improvement is not uniform: on the 7B model, MMMU accuracy decreases from 51.11 to 49.56. These results indicate that the effect of the activation format depends on the model and task.

4.3 Ablation Study

We report a progressive ablation of the evaluated configurations on ScienceQA and MMMU under W4A4 quantization, as detailed in Table 2.

Variant	MDSC	Comp. Loss	Accuracy (%)	
			ScienceQA	MMMU
Baseline	–	–	54.73	30.44
+ MDSC and clipping search	✓	–	60.62	32.11
Ours (Full)	✓	✓	72.93	42.44

Table 2: **Ablation study on Qwen2.5-Omni-3B**. Progressive ablation under W4A4 quantization. The intermediate configuration combines MDSC with clipping-boundary search, and the full configuration additionally uses the composite objective.

Table 2 reports a progressive ablation on ScienceQA and MMMU. Adding MDSC together with clipping-boundary search increases the scores from 54.73 to 60.62 on ScienceQA and from 30.44 to 32.11 on MMMU. Using the composite objective further increases the scores to 72.93 and 42.44, respectively. These results support the effectiveness of the combined configurations, although this ablation does not isolate residual compensation from clipping search or the individual terms of the composite objective.

5 Conclusion

We presented a novel and efficient post-training quantization framework specifically tailored for OLLMs, MorphoQuant, evaluated on Qwen2.5-Omni and InternVL2.5. MDSC combines channel selection with element-wise thresholding to retain selected activation quantization residuals. MPO uses per-channel grid search over shared clipping ratios, coupling the clipping boundary with the residual selection threshold through a composite reconstruction objective. Experiments demonstrate improved W4A4 accuracy over the evaluated baselines on several tasks, with results competitive with selected W4A16 configurations. The effect of alternative activation formats varies across models and benchmarks. Future work will investigate explicit control of residual density, evaluate end-to-end inference efficiency, and extend the evaluation to standalone language tasks and additional omni-modal model families.

AI use statement

We used generative AI tools to assist with the implementation of experimental code and to improve the clarity and language of the manuscript. The authors reviewed and revised the AI-assisted code and text, and checked the code through the reported experiments. All methodological decisions, analyses, and conclusions were made and verified by the authors. We take responsibility for the final content of this paper, including AI-assisted code and writing.

Reproducibility statement

The quantization procedure and optimization objective are described in Section 3. Section 4 specifies the evaluated models, benchmarks, bit-width settings, and calibration procedure. We report the configurations and evaluation protocols used for the results in Tables 1 and 2.

REFERENCES

Saleh Ashkboos, Ilia Markov, Elias Frantar, Tingxuan Zhong, Xincheng Wang, Jie Ren, Torsten Hoefler, and Dan Alistarh. Quik: Towards end-to-end 4-bit inference on generative large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3355–3371, 2024.

Guangji Bai, Zheng Chai, Chen Ling, Shiyu Wang, Jiaying Lu, Nan Zhang, Tingwei Shi, Ziyang Yu, Mengdan Zhu, Yifei Zhang, et al. Beyond efficiency: A systematic survey of resource-efficient large language models. *arXiv preprint arXiv:2401.00625*, 2024.

Md Shalha Mucha Bhuyan, Eftekhar Hossain, Khaleda Akhter Sathi, Md Azad Hossain, and M Ali Akber Dewan. Bvqa: connecting language and vision through multimodal attention for open-ended question answering. *IEEE Access*, 2025.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey. *Findings of the association for computational linguistics: ACL 2024*, pp. 13590–13618, 2024.

Shaoxiang Chen, Zequn Jie, and Lin Ma. Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms. *arXiv preprint arXiv:2401.16160*, 2024.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, 35: 30318–30332, 2022.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023a.

Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. Spqr: A sparse-quantized representation for near-lossless llm weight compression. *arXiv preprint arXiv:2306.03078*, 2023b.

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.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*, 2024.

Wei Huang, Yangdong Liu, Haotong Qin, Ying Li, Shiming Zhang, Xianglong Liu, Michele Magno, and Xiaojuan Qi. Billm: Pushing the limit of post-training quantization for llms. *arXiv preprint arXiv:2402.04291*, 2024a.

Wei Huang, Xingyu Zheng, Xudong Ma, Haotong Qin, Chengtao Lv, Hong Chen, Jie Luo, Xiaojuan Qi, Xianglong Liu, and Michele Magno. An empirical study of llama3 quantization: From llms to mllms. *Visual Intelligence*, 2(1):36, 2024b.

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.

Sehoon Kim, Coleman Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W Mahoney, and Kurt Keutzer. Squeezellm: Dense-and-sparse quantization. *arXiv preprint arXiv:2306.07629*, 2023.

Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024.

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.

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.

Xiaohao Liu, Xiaobo Xia, Manyi Zhang, Ji-Fu Li, Xianzhi Yu, Fei Shen, Xiu Su, See-Kiong Ng, and Tat-Seng Chua. Freeact: Freeing activations for llm quantization. *arXiv preprint arXiv:2603.01776*, 2026.

Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. Datasets for large language models: A comprehensive survey. *Artificial Intelligence Review*, 58(12):403, 2025.

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, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Yuanyong Luo, Zhongxing Zhang, Richard Wu, Hu Liu, Ying Jin, Kai Zheng, Minmin Wang, Zhanying He, Guipeng Hu, Luyao Chen, et al. Ascend hifloat8 format for deep learning. *arXiv preprint arXiv:2409.16626*, 2024.

Yuanyong Luo, Jing Huang, Yu Cheng, Ziwei Yu, Kaihua Tang, Xinda Ma, Xin Wang, Anping Tong, Guipeng Hu, Yun Xu, et al. Hifloat4 format for language model inference. *arXiv preprint arXiv:2602.11287*, 2026.

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.

Gaurav Shinde, Anuradha Ravi, Emon Dey, Shadman Sakib, Milind Rampure, and Nirmalya Roy. A survey on efficient vision-language models. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(3):e70036, 2025.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

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.

Changyuan Wang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. Towards accurate post-training quantization for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16026–16035, 2024b.

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*, 2024c.

Xijun Wang, Junyun Huang, Rayyan Abdalla, Chengyuan Zhang, Ruiqi Xian, and Dinesh Manocha. Bi-vlm: Pushing ultra-low precision post-training quantization boundaries in vision-language models. *arXiv preprint arXiv:2509.18763*, 2025.

Xiuying Wei, Yunchen Zhang, Yuhang Li, Xiangguo Zhang, Ruihao Gong, Jinyang Guo, and Xianglong Liu. Outlier suppression+: Accurate quantization of large language models by equivalent and effective shifting and scaling. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1648–1665, 2023.

Di Wu, Qi Tang, Yongle Zhao, Ming Zhang, Ying Fu, and Debing Zhang. Easyquant: Post-training quantization via scale optimization. *arXiv preprint arXiv:2006.16669*, 2020.

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.

Jingjing Xie, Yuxin Zhang, Mingbao Lin, Liujuan Cao, and Rongrong Ji. Advancing multimodal large language models with quantization-aware scale learning for efficient adaptation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10582–10591, 2024.

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, 2025a. URL https://arxiv.org/abs/2503.20215.

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*, 2025b.

Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1979–1998, 2024.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18155–18165, 2022.

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.

Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22128–22136, 2025.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36:26650–26685, 2023.

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.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Jiaqi Zhao, Miao Zhang, Chao Zeng, Ming Wang, Xuebo Liu, and Liqiang Nie. Lrquant: Learnable and robust post-training quantization for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2240–2255, 2024a.

Yilong Zhao, Chien-Yu Lin, Kan Zhu, Zihao Ye, Lequn Chen, Size Zheng, Luis Ceze, Arvind Krishnamurthy, Tianqi Chen, and Baris Kasikci. Atom: Low-bit quantization for efficient and accurate llm serving. *Proceedings of Machine Learning and Systems*, 6:196–209, 2024b.