BEYOND LOCAL ERROR: FUNCTIONAL RECONSTRUCTION FOR VISION-LANGUAGE MODEL QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training quantization (PTQ) aims to reduce the cost of vision-language models while retaining their capabilities. Reconstruction-based methods pursue this goal by matching quantized intermediate outputs to their full-precision counterparts, with modality-aware approaches further balancing visual and textual errors. Yet an error's impact depends on what happens next: subsequent operations can respond differently to perturbations of similar magnitude. A smaller local error therefore need not better preserve the computation that follows. We introduce Modality-Native Functional Equalization (MNFE), a PTQ method based on *multimodal functional reconstruction*. MNFE evaluates quantization errors at module-specific states formed by downstream operations and balances their contributions across visual evidence, textual context, and answer information. A structured search refines channel scales through bounded group-wise adjustments. Extensive experiments on Qwen2.5-VL-7B and InternVL2.5-8B show that MNFE achieves the highest nine-benchmark average among evaluated quantized methods under W3A16, W4A8, and W4A6. Notably, under W4A6, MNFE retains 94.8% of the full-precision average score on InternVL2.5-8B. Our code will be made publicly available upon acceptance.

1 INTRODUCTION

Vision-language models (VLMs) support visual question answering, document understanding, and reasoning over visual and textual information (Bai et al., 2025; Chen et al., 2024b). Their storage and inference costs, however, limit deployment. Post-training quantization (PTQ) reduces these costs by lowering numerical precision without retraining the full model (Frantar et al., 2023; Xiao et al., 2023). Its effectiveness depends on preserving model capabilities, making the choice of calibration objective central to quantization.

Reconstruction-based methods use intermediate-output error as a tractable surrogate for preserving model behavior. Given a small calibration set, they select quantization parameters that minimize discrepancies between quantized and full-precision layer or block outputs (Frantar et al., 2023; Shao et al., 2024). For VLMs, MBQ (Li et al., 2025) further weights these errors to account for differences in the quantization sensitivity of visual and textual tokens.

At a given reconstruction boundary, however, error magnitude need not reflect the perturbation's effect on subsequent computation. Normalization can attenuate certain magnitude changes, whereas multiplicative gating can amplify the effect of others. Consequently, a scale choice that reduces local output error may still introduce larger deviations in downstream states. Modality-specific weighting addresses differences between visual and textual information, but does not by itself account for how later operations transform an error.

We address this limitation with *multimodal functional reconstruction* and introduce Modality-Native Functional Equalization (MNFE). Rather than applying a uniform local-output criterion, MNFE selects reconstruction states according to the downstream computation associated with each projection group. These targets incorporate residual addition, normalization, and multiplicative gating into channel-scale selection. The objective also distinguishes visual evidence, textual context, and answer information: it normalizes reconstruction errors within each role and allocates their weights

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.

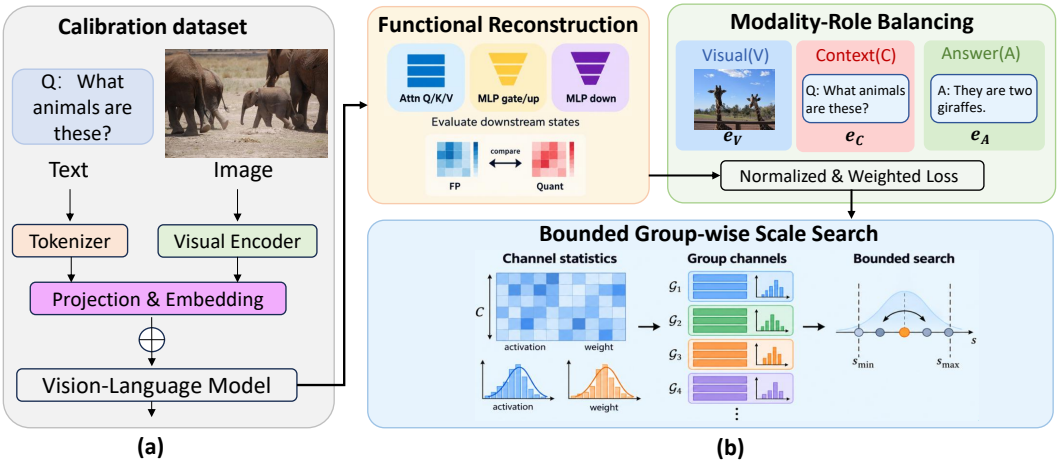

Figure 1: **Overview of MNFE.** (a) Multimodal calibration inputs and the shared language backbone. (b) Module-specific functional reconstruction and hierarchical role balancing define the objective for bounded group-wise scale search.

hierarchically, first between vision and text, then between context and answers. Calibration thus accounts for both the downstream effects of quantization and the information roles they affect.

To optimize this objective, we combine statistics-based initialization with bounded group-wise scale refinement. Channel rescaling preserves the unquantized mapping while changing the distributions presented to the weight and activation quantizers (Xiao et al., 2023; Lin et al., 2024). Whereas MBQ's scale family couples channel adjustments through a shared parameter, MNFE allows groups of channels to make bounded adjustments around an initial solution.

We evaluate MNFE on two VLM families at W3A16, W4A8, and W4A6. On Qwen2.5-VL-7B and InternVL2.5-8B, it achieves the highest nine-benchmark average among the evaluated quantized methods in all six model–precision settings, exceeding MBQ, QIG, and MASQuant by 2.0, 1.1, and 1.0 points on average, respectively. On Qwen2.5-VL-32B and InternVL2.5-38B, it leads the seven-benchmark average in five of six settings. Fixed-candidate ablations further examine the contributions of the reconstruction targets and role weighting independently of the structured search.

In summary, our contributions are:

- We formulate multimodal functional reconstruction using module-specific downstream targets and hierarchical weighting of visual, context, and answer errors.
- We develop a bounded group-wise search that selects channel scales under this objective without adding inference-time operations.
- We demonstrate improvements in average benchmark performance across two VLM families and multiple precisions, with evaluations spanning nine benchmarks on 7B/8B models and seven benchmarks on Qwen2.5-VL-32B and InternVL2.5-38B.

2 Related Work

2.1 Large Vision-Language Models

Large vision-language models connect visual encoders to language backbones through learned interfaces, as exemplified by BLIP-2 (Li et al., 2023) and LLaVA (Liu et al., 2023a). Qwen2.5-VL (Bai et al., 2025) and InternVL2.5 (Chen et al., 2024b) further develop this paradigm for multimodal understanding and generation. For the models studied here, Figure 1(a) illustrates how visual features and text embeddings form a shared input sequence for the language backbone that we quantize.

2.2 POST-TRAINING QUANTIZATION

PTQ reduces numerical precision without retraining the full model. GPTQ (Frantar et al., 2023) uses approximate second-order information for layer-output reconstruction, while BRECQ (Li et al., 2021) and OmniQuant (Shao et al., 2024) use block-level reconstruction. Equivalent transformations rescale channels in SmoothQuant (Xiao et al., 2023) and AWQ (Lin et al., 2024), or suppress activation outliers through rotations in QuaRot (Ashkboos et al., 2024). QEP (Arai & Ichikawa, 2025) compensates for errors propagated from preceding layers. These approaches address complementary aspects of PTQ: reconstruction granularity, reparameterization, and error propagation.

2.3 QUANTIZATION OF VISION-LANGUAGE MODELS

VLM quantization addresses computational dependencies and multimodal heterogeneity. Q-VLM (Wang et al., 2024) uses entropy-guided block partitioning to account for cross-layer dependencies. MBQ (Li et al., 2025) weights reconstruction errors by modality sensitivity, while QIG (Xiang et al., 2026) uses quantization-aware integrated gradients for token-level weighting. MQuant (Yu et al., 2025) introduces modality-specific static quantization, and MASQuant (Hu et al., 2026) combines modality-specific smoothing with cross-modal compensation. Quant Experts (Jia et al., 2026) uses shared and routed low-rank experts for error compensation; VLM-PTQ (Deng & Huang, 2026) corrects weight-rounding targets under asymmetric calibration and incorporates modality-dependent channel importance. MNFE couples module-specific downstream reconstruction targets with balanced visual, context, and answer errors to select channel scales without adding inference-time operations.

3 METHOD

Figure 1(b) summarizes MNFE, which combines three components: (i) module-specific functional reconstruction (Section 3.2), (ii) hierarchical role balancing (Section 3.3), and (iii) bounded group-wise scale search (Section 3.4). The first two define the calibration objective, while the third uses this objective to select channel scales. We first introduce channel equalization in Section 3.1.

3.1 CHANNEL EQUALIZATION

Let $X \in \mathbb{R}^{C\times N}$ contain N token activations and $W \in \mathbb{R}^{D\times C}$ denote a projection matrix. For a positive diagonal matrix $E = \mathrm{diag}(s_1, \ldots, s_C)$,

$$WX = (WE)(E^{-1}X), \qquad \widehat{Y}(E) = Q_W(WE)\, Q_A(E^{-1}X). \tag{1}$$

Here Q_W and Q_A return the quantized–dequantized weights and activations, respectively; Q_A is the identity for weight-only quantization. The transformation preserves the unquantized mapping but changes the numerical ranges presented to the quantizers. Biases are omitted from the notation.

We calibrate three projection groups in each language block: attention Q/K/V, MLP gate/up, and MLP down. Projections within a group share an input and therefore use the same channel scales. For group g, let $F_{g,k}$ and $\widehat{F}_{g,k}(E)$ denote the full-precision and candidate states at reconstruction target k, respectively. We seek scales that minimize a role-balanced reconstruction loss $\mathcal{L}_g(E)$ using the structured search in Section 3.4. During each trial, quantization is confined to the active group; its reference inputs, residual context, and the parameters of all other operations used to form its targets remain fixed.

3.2 FUNCTIONAL RECONSTRUCTION

The reconstruction target determines which effects of a local perturbation enter the objective. We place targets at states formed by the downstream operations relevant to each projection group, rather than applying the same local-output criterion to every group.

For a pre-normalized Transformer block, let h be the block input and a the attention output, including the attention output projection. The full-precision states are

$$\begin{aligned} r &= h + a, \qquad z = N_{\mathrm{MLP}}(r), \\ m &= \sigma(W_{\mathrm{gate}}z) \odot (W_{\mathrm{up}}z), \qquad d = W_{\mathrm{down}}m, \end{aligned} \tag{2}$$

where N_{MLP} is the MLP-input RMSNorm, σ is the gating activation, and $\odot$ denotes element-wise multiplication. The block output is $r + d$.

These states define the group-specific reconstruction targets:

$$\begin{aligned} F_{\mathrm{attn},1} &= z, \\ (F_{\mathrm{gu},1}, F_{\mathrm{gu},2}) &= (m, r + d), \\ F_{\mathrm{down},1} &= N_{\mathrm{next}}(r + d). \end{aligned} \tag{3}$$

Attention and down each have one target ($K_g = 1$), whereas gate/up has two ($K_g = 2$).

Attention Q/K/V. Reconstructing z evaluates the attention perturbation after residual addition and normalization, at the state consumed by the MLP. This incorporates the effect of these downstream operations into scale selection rather than measuring the attention output alone.

MLP gate/up. The product target m preserves features formed by multiplicative gating, while $r + d$ evaluates their projected contribution to the residual stream. The shared residual cancels from the latter discrepancy, so this target measures down-projected feature error relative to the reference energy of $r + d$. The two target errors receive equal weight after normalization, preserving both the gated features and their contribution to the state passed forward.

MLP down. Reconstructing $N_{\mathrm{next}}(r + d)$ evaluates the residual update in the normalized representation passed to the subsequent computation. Here N_{next} is the next block's input RMSNorm, or the model's final RMSNorm in the last block.

Candidate targets follow the same computations with only the active group quantized. All targets are computed on complete multimodal sequences, with token-role masks applied only when aggregating their errors.

3.3 Hierarchical Role-Balanced Reconstruction

Pooling all token errors assigns influence according to both token count and representation magnitude. To control these effects, we first normalize reconstruction error within each information role, then allocate explicit weights across modalities and within text.

Reference-normalized errors. Let $\mathcal{R} = \{V, A, C\}$ denote visual, answer, and context roles, with disjoint token sets $\mathcal{T}_\rho$ pooled over the calibration set. Visual positions are identified by the input adapter, answer positions by non-ignored supervision labels from the calibration responses, and context comprises all remaining valid positions. The pooled role sets are nonempty and exclude padding.

For target k with D_k features, let $e_{g,k,\rho}$ denote its mean squared full-precision value over tokens and features within role ρ, floored by a positive constant. We measure

$$R_{g,k,\rho}(E) = \frac{\sum_{t \in \mathcal{T}_\rho} \|\widehat{F}_{g,k,t}(E) - F_{g,k,t}\|_2^2}{|\mathcal{T}_\rho| D_k \, e_{g,k,\rho}}. \tag{4}$$

Averaging within a role removes direct scaling with its token count, while division by reference energy measures relative reconstruction error. The energies remain fixed across candidates.

Hierarchical allocation. Let $b_{g,\rho} > 0$ summarize the role's full-precision target-to-input energy ratios and normalized across roles. These reference-derived scores are independent of candidate errors. We first balance visual and textual information, using $\sqrt{b_{g,A} b_{g,C}}$ as the textual score, then divide the textual share between answer and context:

$$\begin{aligned} \pi_{g,V} &= \frac{b_{g,V}}{b_{g,V} + \sqrt{b_{g,A} b_{g,C}}}, \\ \pi_{g,A} &= (1 - \pi_{g,V}) \frac{b_{g,A}}{b_{g,A} + b_{g,C}}, \qquad \pi_{g,C} = (1 - \pi_{g,V}) \frac{b_{g,C}}{b_{g,A} + b_{g,C}}. \end{aligned} \tag{5}$$

The weights sum to one. Equal target-to-input energy ratios yield $(\pi_{g,V}, \pi_{g,A}, \pi_{g,C}) = (1/2, 1/4, 1/4)$, so distinguishing two textual roles does not give text twice the aggregate weight of vision in this reference case.

The complete functional reconstruction objective is

$$\mathcal{L}_g(E) = \frac{1}{K_g} \sum_{k=1}^{K_g} \sum_{\rho \in \mathcal{R}} \pi_{g,\rho} R_{g,k,\rho}(E). \tag{6}$$

Functional targets specify the states to preserve; role normalization and hierarchical weighting determine their contributions to scale selection. Both target energies and role weights remain fixed throughout a group's search.

3.4 Structured Scale Search

We search for scales in two stages. A statistics-based grid provides an initialization, and bounded channel-bin offsets refine it. Both stages use the same objective in Equation 6.

Statistics-based initialization. For input channel c, let a_c be the 99th percentile of its absolute activations and w_c the maximum absolute weight across all output channels and projections in the group. With both statistics floored to be positive, we construct scales in log space:

$$\begin{aligned} u_c(\alpha, \beta) &= \alpha \log a_c - \beta \log w_c, \\ s_c(u) &= \max\left\{ s_{\min}, \exp\left(u_c - \frac{\max_j u_j + \min_j u_j}{2} \right) \right\}. \end{aligned} \tag{7}$$

Here $s_{\min} > 0$ is a numerical floor and $E(u) = \mathrm{diag}(s_1(u), \ldots, s_C(u))$. Independent exponents control the activation and weight contributions, while midrange centering removes a common log-scale shift. We score a small exponent grid with $\mathcal{L}_g$ and retain the lowest-loss initialization $u^{(0)}$.

Bounded channel-bin refinement. We sort channels by $\log a_c - \log w_c$ and partition them into approximately equal-sized bins. Let $b(c)$ be the bin assigned to channel c. Keeping this assignment and the initialization fixed, we use

$$u_c(\delta) = u_c^{(0)} + \delta_{b(c)}, \qquad |\delta_b| \le d_g. \tag{8}$$

Channels with similar activation-to-weight ratios share an offset, allowing adjustments beyond the two-exponent initialization family. The bound constrains these offsets before the normalization in Equation 7.

We set the range using two isolated functional losses at $E(u^{(0)})$: $L_{W,g}$ quantizes only weights, and $L_{A,g}$ only activations. Both use the same targets, reference energies, and role weights as $\mathcal{L}_g$. We set

$$q_g = \frac{L_{A,g}}{L_{W,g} + L_{A,g} + \epsilon_{\mathrm{range}}}, \qquad d_g = \frac{d_{\max}}{2}(1 + q_g), \tag{9}$$

where $d_{\max}$ is the nominal bound and $\epsilon_{\mathrm{range}} > 0$ stabilizes the denominator. This rule permits larger offsets when the activation-only probe has a larger relative loss, with $d_g \in [d_{\max}/2, d_{\max}]$. The probes determine only the range, not an additive decomposition of joint quantization error.

Discrete coordinate search. Starting from zero offsets, we test discrete values for one bin at a time, holding the others fixed, and accept a change only if it strictly lowers $\mathcal{L}_g$.

Parameter folding. After calibrating a block, we fold the selected scales into its existing normalization and linear parameters, preserving the unquantized mapping in Equation 1. Inference therefore retains the original architecture without additional operations.

Table 1: Main results on Qwen2.5-VL-7B. Best and second-best scores among quantized methods within each precision are highlighted in **bold** and underline, respectively.

Method	MMMU	OCR	VizWiz	AI2D	TextVQA	ScienceQA	MMStar	ChartQA	DocVQA	**Avg.**
BF16/FP16	50.2	83.6	70.6	82.5	83.1	87.5	62.6	83.1	96.0	77.7
W3A16										
RTN	**46.0**	75.8	67.1	79.2	79.0	84.3	56.7	75.4	93.0	73.0
MBQ	45.3	80.4	66.3	79.4	80.2	85.0	56.2	76.7	93.4	73.7
QIG	43.9	**83.0**	65.3	79.2	80.7	85.0	57.6	75.7	93.5	73.8
MASQuant	45.9	81.4	66.2	**80.8**	80.8	**85.3**	57.5	76.9	**94.9**	74.4
MNFE (Ours)	45.3	81.5	**67.5**	80.3	**81.6**	85.2	**58.0**	**79.1**	94.4	**74.8**
W4A8										
RTN	45.6	67.1	63.5	79.2	77.2	84.0	58.8	78.4	90.6	71.6
MBQ	44.4	69.6	62.5	78.9	75.4	83.4	58.2	78.5	87.7	70.9
QIG	46.9	78.0	64.9	**79.7**	78.5	85.6	59.3	**80.4**	90.5	73.8
MASQuant	**47.6**	74.7	64.2	79.0	77.1	84.2	58.1	77.5	91.7	72.7
MNFE (Ours)	47.4	**79.7**	**66.3**	**79.7**	**80.4**	**86.4**	**59.7**	79.9	**93.8**	**74.8**
W4A6										
RTN	29.0	43.7	39.5	12.8	45.6	11.6	16.3	13.6	64.6	30.7
MBQ	38.3	68.3	57.5	73.5	70.8	75.5	49.2	71.5	87.5	65.8
QIG	39.8	70.3	58.8	75.0	75.7	77.8	52.4	71.7	89.4	67.9
MASQuant	40.4	74.5	61.6	**76.4**	73.4	77.1	**55.9**	**73.5**	91.3	69.3
MNFE (Ours)	**41.3**	**75.5**	**62.0**	75.8	**77.4**	**81.0**	55.4	71.2	**92.7**	**70.2**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Implementation Details. Our main comparison uses Qwen2.5-VL-7B (Bai et al., 2025) and InternVL2.5-8B (Chen et al., 2024b) at W3A16, W4A8, and W4A6. We also evaluate Qwen2.5-VL-32B and InternVL2.5-38B at the same three precisions. W3A16 uses 3-bit weights and 16-bit activations; W4A8 and W4A6 use 4-bit weights with 8-bit and 6-bit activations. All experiments are conducted on NVIDIA H100 GPUs.

Calibration and evaluation. Following prior work (Li et al., 2025), we randomly sample 128 image–caption pairs from the COCO subset of ShareGPT4V (Chen et al., 2023) for calibration. The main results use sampling seed 42. Each quantized model uses one set of calibrated scales across all evaluated tasks. The 7B/8B suite comprises MMMU (Yue et al., 2024), OCRBench (Liu et al., 2023b), VizWiz (Gurari et al., 2018), AI2D (Kembhavi et al., 2016), TextVQA (Singh et al., 2019), ScienceQA (Lu et al., 2022), MMStar (Chen et al., 2024a), ChartQA (Masry et al., 2022), and DocVQA (Mathew et al., 2021). The 32B/38B suite excludes MMStar and ChartQA, leaving seven benchmarks. We report the unweighted mean score over each suite.

Baselines. We compare with BF16/FP16, round-to-nearest quantization (RTN), MBQ (Li et al., 2025), QIG (Xiang et al., 2026), and MASQuant (Hu et al., 2026) across all four models and three precisions.

4.2 MAIN RESULTS

Tables 1 and 2 report results on Qwen2.5-VL-7B and InternVL2.5-8B. MNFE achieves the highest nine-benchmark average among the evaluated quantized methods in all six model–precision settings. Averaged across these settings, its score of 73.7 exceeds MBQ, QIG, and MASQuant by 2.0, 1.1, and 1.0 points, respectively.

Tables 1 and 2 report results on Qwen2.5-VL-7B and InternVL2.5-8B. MNFE achieves the highest nine-benchmark average among the evaluated quantized methods in all six model–precision settings.

Table 2: Main results on InternVL2.5-8B. Best and second-best scores among quantized methods within each precision are highlighted in **bold** and underline, respectively.

Method	MMMU	OCR	VizWiz	AI2D	TextVQA	ScienceQA	MMStar	ChartQA	DocVQA	**Avg.**
BF16/FP16	53.9	80.2	62.2	83.3	79.3	98.1	60.7	84.1	93.2	77.2
W3A16										
RTN	49.1	76.9	56.8	**81.6**	77.6	97.2	56.8	81.7	91.5	74.4
MBQ	50.3	**78.2**	57.9	81.5	77.5	96.9	57.7	81.7	91.5	74.8
QIG	49.2	77.7	**59.4**	81.0	77.6	97.1	57.6	81.8	**91.7**	74.8
MASQuant	49.4	75.9	58.0	80.4	77.7	97.0	56.8	82.3	91.7	74.3
MNFE (Ours)	**50.6**	77.8	58.7	81.4	**77.9**	**97.6**	**58.0**	**82.5**	91.4	**75.1**
W4A8										
RTN	48.8	76.1	54.5	80.3	76.5	**97.7**	55.8	81.4	91.4	73.6
MBQ	**50.1**	76.4	56.4	80.5	76.3	97.6	**58.2**	80.8	91.0	74.1
QIG	48.2	76.8	56.3	80.0	73.8	96.7	57.1	80.6	91.6	73.5
MASQuant	48.9	75.7	56.2	**80.9**	75.9	97.5	58.0	81.4	**92.3**	74.1
MNFE (Ours)	49.4	**77.5**	**57.2**	80.7	**76.8**	97.5	56.7	**82.0**	91.4	**74.4**
W4A6										
RTN	44.4	72.8	53.3	76.5	73.7	95.2	52.2	77.7	87.8	70.4
MBQ	46.3	71.4	54.2	78.6	72.9	95.5	55.7	78.6	88.7	71.3
QIG	46.7	73.2	56.2	78.7	74.8	**97.1**	55.5	80.2	89.6	72.4
MASQuant	45.6	**73.7**	55.5	78.7	72.0	95.4	**56.9**	79.2	88.8	71.8
MNFE (Ours)	**48.8**	**73.7**	**57.7**	**79.6**	**75.3**	96.4	56.5	**80.5**	**90.2**	**73.2**

Table 3: Seven-benchmark averages on Qwen2.5-VL-32B and InternVL2.5-38B. Best and second-best quantized results in each column are highlighted in **bold** and underline, respectively.

Method	Qwen2.5-VL-32B			InternVL2.5-38B		
	W3A16	W4A8	W4A6	W3A16	W4A8	W4A6
BF16/FP16	78.4			81.7		
RTN	**76.5**	75.2	3.3	80.0	75.9	72.0
MBQ	76.3	72.8	70.8	80.4	77.4	73.1
QIG	75.5	72.9	72.1	80.6	77.9	74.4
MASQuant	76.1	74.3	70.9	80.4	78.0	75.5
MNFE (Ours)	76.3	**75.4**	**73.1**	**80.7**	**78.4**	**76.2**

Averaged across these settings, its score of 73.7 exceeds MBQ, QIG, and MASQuant by 2.0, 1.1, and 1.0 points, respectively.

4.3 Results on Larger Models

MNFE achieves the highest seven-benchmark average among the evaluated quantized methods in five of six 32B/38B settings. Table 3 summarizes results on Qwen2.5-VL-32B and InternVL2.5-38B at all three precisions. Averaged across these settings, MNFE improves over MBQ, QIG, and MASQuant by 1.5, 1.1, and 0.8 points, respectively. It achieves the highest quantized average in all four W4A8/W4A6 settings and on InternVL2.5-38B/W3A16.

Appendix A.6 reports the complete task-level results for both models at every precision.

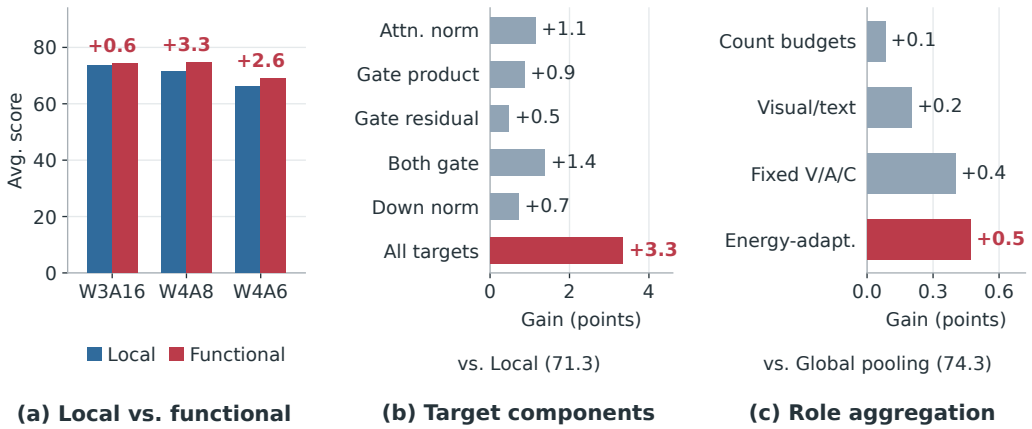

Figure 2: **Ablation of the reconstruction objective.** Nine-benchmark results on Qwen2.5-VL-7B with a shared fixed candidate pool. (a) Local and functional targets across precisions. (b) W4A8 gains from partial and complete target replacement. (c) W4A8 gains from alternative role-weighting schemes. Gains in (b) and (c) are relative to local reconstruction and global pooling, respectively.

Table 4: Target-component ablations on Qwen2.5-VL-7B at W4A8. Δ is relative to Local.

Control	Attn. norm	Gate product	Gate residual	Down norm	Avg.	Δ
Local	–	–	–	–	71.3	0.0
Attention norm	✓	–	–	–	72.5	+1.1
Gate product	–	✓	–	–	72.2	+0.9
Gate residual	–	–	✓	–	71.8	+0.5
Both gate targets	–	✓	✓	–	72.7	+1.4
Down next-norm	–	–	–	✓	72.1	+0.7
All targets	✓	✓	✓	✓	**74.7**	**+3.3**

4.4 Analysis of the Reconstruction Objective

We examine reconstruction targets and role weighting on Qwen2.5-VL-7B using 128 calibration examples (seed 42). All variants select from the same 128-candidate scale pool per projection group, replacing the structured search to isolate objective design. Protocols and task-level results are provided in Appendix A.5.

Reconstruction targets. With role-normalized squared error and fixed visual/answer/context weights $(1/2, 1/4, 1/4)$, functional targets raise the nine-benchmark average from 73.7 to 74.4 at W3A16, 71.3 to 74.7 at W4A8, and 66.3 to 69.0 at W4A6 (Figure 2(a)). All nine benchmarks improve at both W4A8 and W4A6. The larger gains under weight-and-activation quantization indicate that the choice of reconstruction state is particularly consequential in these settings.

Target composition. Table 4 compares functional-target combinations at W4A8. Checkmarks indicate selected targets; unmodified groups retain local reconstruction. Every partial replacement improves on Local (Figure 2(b)). Within gate/up, combining the gated-product and residual-stream targets gives 72.7, compared with 72.2 and 71.8 individually. The complete target set reaches 74.7, exceeding the strongest partial configuration by 2.0 points. These results support using both gate/up targets and applying functional reconstruction across projection groups.

Role balancing. With functional targets fixed, Table 5 compares five W4A8 aggregation rules. The average rises from 74.3 under global pooling to 74.4 with count-proportional budgets and 74.5 with equal visual/text budgets. Fixed visual/answer/context (V/A/C) weights reach 74.7, supporting

Table 5: Role-aggregation ablations on Qwen2.5-VL-7B at W4A8 with functional targets fixed. Δ is relative to global pooling.

Aggregation	Budget / normalization rule	Avg.	Δ
Global pooling	Pooled tokens and reference energy	74.3	0.0
Count budgets	Role normalization; token-count budgets	74.4	+0.1
Visual/text	Equal V/text; count-proportional A/C	74.5	+0.2
Fixed V/A/C	Role budgets $(1/2, 1/4, 1/4)$	74.7	+0.4
Energy-adaptive	Hierarchical, reference-energy weights	**74.8**	**+0.5**

explicit allocation within the textual budget. Reference-energy weighting reaches 74.8, a modest additional gain of 0.1 points. Figure 2(c) summarizes these gains over global pooling.

5 Conclusion

We presented MNFE, a PTQ method that evaluates quantization errors through module-specific downstream states and balances their contributions across visual, context, and answer roles. Bounded group-wise scale search optimizes this objective without adding inference-time operations. MNFE achieves the highest benchmark average among the evaluated quantized methods in all six 7B/8B settings and five of six 32B/38B settings. Fixed-candidate ablations support functional targets over local reconstruction, with larger gains under weight-and-activation quantization. These findings highlight reconstruction-state selection and multimodal error aggregation as complementary considerations in VLM quantization.

AI Use Statement

Generative AI tools assisted code development, documentation, and language editing. The authors are responsible for checking the implementation, experimental results, references, and final manuscript.

References

Yamato Arai and Yuma Ichikawa. Quantization error propagation: Revisiting layer-wise post-training quantization. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=a3l3K9khbL.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. QuaRot: Outlier-free 4-bit inference in rotated LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b5b939436789f76f08b9d0da5e81af7c-Abstract-Conference.html.

Shuai Bai et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. ShareGPT4V: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. URL https://arxiv.org/abs/2311.12793.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL https://papers.nips.cc/paper_files/paper/2024/hash/2f8ee6a3d766b426d2618e555b5aeb39-Abstract-Conference.html.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b. URL https://arxiv.org/abs/2412.05271.

Juncan Deng and Kejie Huang. VLM-PTQ: Efficient post-training quantization for large vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24696–24705, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Deng_VLM-PTQ_Efficient_Post-Training_Quantization_for_Large_Vision-Language_Models_CVPR_2026_paper.html.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.17323.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz grand challenge: Answering visual questions from blind people. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Gurari_VizWiz_Grand_Challenge_CVPR_2018_paper.html.

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. *arXiv preprint arXiv:2603.04800*, 2026. URL https://arxiv.org/abs/2603.04800.

Chenwei Jia, Baoting Li, Xuchong Zhang, Mingzhuo Wei, Bochen Lin, and Hongbin Sun. Quant experts: Token-aware adaptive error reconstruction with mixture of experts for large vision-language models quantization. *arXiv preprint arXiv:2602.24059*, 2026. URL https://arxiv.org/abs/2602.24059.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, pp. 235–251, 2016. doi: 10.1007/978-3-319-46493-0_15. URL https://arxiv.org/abs/1603.07396.

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, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

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

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. BRECQ: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=POWv6hDd9XH.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Conference on Machine Learning and Systems*, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/hash/42a452cbafa9dd64e9ba4aa95cc1ef21-Abstract-Conference.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: On the hidden mystery of OCR in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023b. URL https://arxiv.org/abs/2305.07895.

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

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

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

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. OmniQuant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2308.13137.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Singh_Towards_VQA_Models_That_Can_Read_CVPR_2019_paper.html.

Changyuan Wang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. Q-VLM: Post-training quantization for large vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/cffbaf4f47546ece96bb42c0edda40ee-Abstract-Conference.html.

Ziwei Xiang, Fanhu Zeng, Hongjian Fang, Rui-Qi Wang, Renxing Chen, Yanan Zhu, Yi Chen, Peipei Yang, and Xu-Yao Zhang. Fine-grained post-training quantization for large vision language models with quantization-aware integrated gradients. *arXiv preprint arXiv:2603.17809*, 2026. URL https://arxiv.org/abs/2603.17809.

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*, 2023. URL https://proceedings.mlr.press/v202/xiao23c.html.

JiangYong Yu, Sifan Zhou, Dawei Yang, Shuo Wang, Shuoyu Li, Xing Hu, Chen Xu, Zukang Xu, Changyong Shu, and Zhihang Yuan. MQuant: Unleashing the inference potential of multimodal large language models via full static quantization. In *ACM International Conference on Multimedia*, 2025. URL https://arxiv.org/abs/2502.00425.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.16502.