RONA: ROTATION AND NOISE-AWARE ABSORPTION FOR LOW-BIT UNIVERSAL MULTIMODAL EMBEDDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Universal multimodal embedding (UME) encodes text, images, video, and visually rich documents into a shared space for cross-modal retrieval. Although reasoning-enhanced methods improve embedding quality, their additional inference-time computation limits low-latency, high-throughput retrieval, leaving direct encoders as the more deployment-friendly choice. Their large backbones, however, still incur substantial serving cost. We study whether post-training quantization (PTQ) can compress UME without sacrificing retrieval accuracy. Across 78 MMEB-V2 tasks, strong PTQ methods remain near full precision at W4A16, W8A16, W4A8, and W8A8, yet collapse at W4A4, a failure not explained by layer-wise reconstruction error. We show that low-bit UME must preserve retrieval geometry and that directions safe under clean activations become harmful when coupled with activation quantization noise. We therefore propose RONA, which combines function-preserving rotations with a group-aligned, quantization-noise-aware safe subspace for weight reparameterization. A per-layer absorption coefficient, calibrated against grouped-GPTQ weight reconstructions and query–target embedding geometry, controls how strongly each layer uses these safe directions while the backbone remains frozen. RONA achieves 73.55 at W4A4, exceeding the strongest PTQ baseline by 3.65 points and retaining 94.4% of full-precision performance. Gains of 6.20 and 2.29 points on two 2B backbones demonstrate its effectiveness across model scales and architectures.

1 INTRODUCTION

Universal multimodal embedding (UME) maps text, images, video, and visually rich documents into a shared vector space for cross-modal retrieval. Representative UME systems include VLM2Vec (Jiang et al., 2025), VLM2Vec-V2 (Meng et al., 2026), and UEmbed (Song et al., 2026). Recent work further improves retrieval embeddings by introducing reasoning before representation readout, including Think Then Embed (Cui et al., 2026), UME-R1 (Lan et al., 2026), Embed-RL (Jiang et al., 2026), and retrieval-centric CoT (Sun et al., 2026); however, explicit reasoning requires multiple autoregressive decoding steps and incurs substantial latency. More efficient variants reduce this overhead through retrieval-oriented rewriting (Wu et al., 2026a), on-demand CoT (Zhang et al., 2026), TTE-Flash (Cheng et al., 2026), latent reasoning with LaME (Wu et al., 2026b), or recurrent computation with ReLoop-UME (Wang et al., 2026a). Even so, their gains still rely on extra inference-time computation, making low-latency, high-throughput retrieval difficult. Practical systems therefore continue to favor direct encoders such as Qwen3-VL-Embedding (Li et al., 2026a), DME (Chen et al., 2026a), and WeMM-Embedding (Zhou et al., 2026). Yet their weights and activations still dominate serving cost. Low-bit quantization is a mature, architecture-orthogonal way to reduce storage and computation while preserving the model architecture and inference path. This raises a natural question: **can low-bit quantization reduce UME serving cost while maintaining high retrieval accuracy and minimizing the performance loss introduced by quantization?**

Post-training quantization (PTQ) is a widely used method for low-bit quantization because it requires no model retraining and typically needs only a small calibration set. Among PTQ methods, GPTQ (Frantar et al., 2023) uses approximate second-order information to compensate weight-quantization error, whereas SmoothQuant (Xiao et al., 2023) alleviates activation outliers by migrating quantization difficulty into the weights through channel scaling. To further improve joint weight–activation quantization, recent methods reshape the quantization space itself:

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.

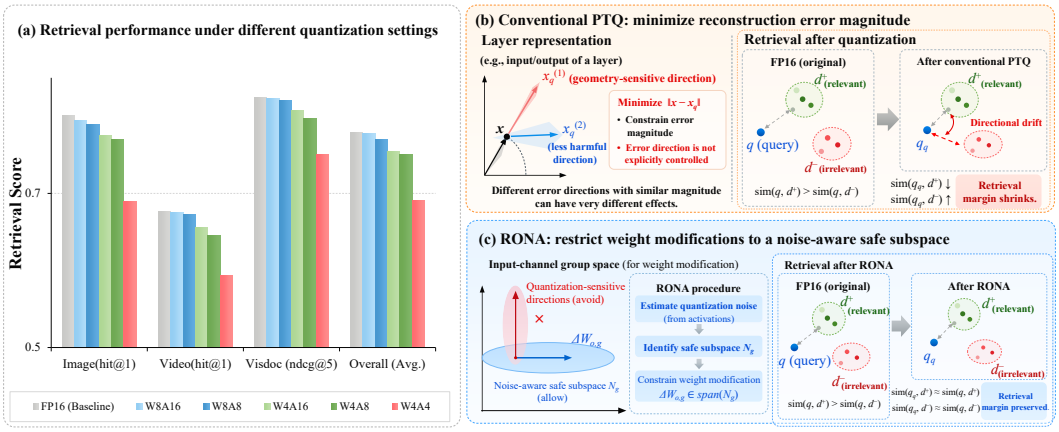

Figure 1: Motivation for RONA. (a) QuaRot exemplifies the W4A4 retrieval drop on MMEB-V2; additional precision comparisons are organized in Appendix E. (b) Similar reconstruction errors can induce different retrieval-margin distortions. (c) RONA restricts weight modifications to a quantization-noise-aware safe subspace to preserve retrieval geometry.

QuaRot (Ashkboos et al., 2024) and SpinQuant (Liu et al., 2025) introduce rotations, DuQuant (Lin et al., 2024) combines rotations with channel permutation, and FlatQuant (Sun et al., 2025) learns affine transforms for weights and activations. Despite these advances, a clear low-bit bottleneck remains on MMEB-V2 (Meng et al., 2026): across its 78 tasks, conventional methods based on layer-wise reconstruction error remain near full-precision performance at W4A16, W8A16, W4A8, and W8A8, yet collapse at **W4A4**, as illustrated by QuaRot in Figure 1(a).

Existing PTQ mainly constrains error *magnitude*, not *direction*: it keeps local numerical error small without controlling how perturbations propagate. This is often tolerable for token prediction, but an embedding compresses an input into a *single point* in a shared space. Small perturbations shift query–candidate similarity margins and reorder nearest neighbors (Figure 1(b)). Low-bit UME must therefore preserve retrieval geometry, not merely layer-wise numerical fidelity.

Direction alone is insufficient. Directions identified as safe under full-precision activations become harmful after activation quantization when a weight modification acts on both the clean signal and the quantizer's error. Grouped quantizers round each group independently, and this independent rounding breaks cross-group cancellation after discretization. Safe reparameterization must therefore be identified under *real activation quantization noise* and aligned with quantization-group boundaries.

Based on this observation, we propose **RONA**. Function-preserving rotations first select a coordinate system more amenable to activation quantization. Within each quantization group, RONA constructs a *quantization-noise-aware safe subspace* from the rotated activations and the target quantizer's errors, and restricts weight modifications to it (Figure 1(c)). Rotation, subspace estimation, and reparameterization share the same group boundaries, and all modifications are absorbed into the final weights, requiring no high-precision residual branch at inference.

The safe subspace determines a layer's admissible modification directions, but not *how far* it moves. RONA therefore assigns each target Linear layer a continuous *absorption coefficient* interpolating between the rotation-only and safe-subspace parameterizations. The coefficients are calibrated against grouped-GPTQ weight reconstructions using query–target embedding drift and positive-pair geometry rather than reconstruction error; only these scalars are optimized while the backbone remains frozen. We summarize our contributions as follows:

- We present the first systematic study of low-bit UME quantization and identify a W4A4 failure mode: directions insensitive under clean activations become highly sensitive when coupled with real activation quantization noise, showing why layer-wise reconstruction alone cannot preserve retrieval geometry.
- We propose RONA, a *group-aligned, quantization-noise-aware* PTQ framework combining function-preserving rotations, safe weight reparameterization, and task-adaptive absorption cal-

ibrated against grouped-GPTQ weight reconstructions and query–target embedding geometry, without a high-precision residual branch.

- On all 78 MMEB-V2 tasks, RONA outperforms strong PTQ baselines at **W4A4** while retaining high retrieval accuracy, and remains near full precision at W4A16, W8A16, W4A8, and W8A8.

2 Related Work

2.1 Universal Multimodal Embedding

Universal multimodal embedding uses a shared encoder for queries and targets combining text and visual content. VLM2Vec (Jiang et al., 2025) introduces instruction-conditioned embedding tasks, and VLM2Vec-V2 (Meng et al., 2026) extends them to videos and visual documents through MMEB-V2. Qwen3-VL-Embedding (Li et al., 2026a), DME (Chen et al., 2026a), and WeMM-Embedding (Zhou et al., 2026) further advance direct encoders. A complementary line enriches embeddings through generation: Think Then Embed (Cui et al., 2026) supplies reasoning context, while UME-R1 (Lan et al., 2026) supports discriminative and generative embeddings through supervised fine-tuning and reinforcement learning with ranking and similarity-gap rewards. RONA compresses a trained direct encoder and calibrates its query–target geometry while retaining direct embedding readout.

2.2 Post-Training Quantization for Large Models

PTQ methods reshape weights and activations using different transformations and calibration signals. GPTQ (Frantar et al., 2023) compensates weight-rounding errors with approximate second-order information; SmoothQuant (Xiao et al., 2023) redistributes quantization difficulty through channel scaling. QuaRot (Ashkboos et al., 2024) applies invariant rotations, SpinQuant (Liu et al., 2025) learns them, and DuQuant (Lin et al., 2024) combines block rotations with channel permutation. OmniQuant (Shao et al., 2024) learns clipping and equivalent transformations, whereas FlatQuant (Sun et al., 2025) learns affine transformations. OSAQ (Li et al., 2026b) suppresses weight outliers through closed-form updates in an approximate Hessian null space, also complementing W4A4 PTQ. RONA instead targets W4A4 retrieval through group-local clean/noise second moments and learned absorption calibrated by bilateral embedding geometry (Appendix F).

Error propagation and model-level calibration are also important: QEP (Arai & Ichikawa, 2025) incorporates upstream errors into layer calibration, and SpinQuant (Liu et al., 2025) uses a model-level objective. A mechanistic study (Chen et al., 2026b) connects inherited and newly introduced hidden errors to LM-head geometry and token predictions. Retrieval compares two encoded inputs, so query- and target-side errors jointly affect similarity margins. RONA controls their combined effect through noise-aware, group-aligned weight directions and grouped-GPTQ-consistent calibration.

3 Method

3.1 Problem Formulation and Overview

Let f_θ be a pretrained direct UME encoder. Its normalized embedding $z(x) = f_\theta(x)/\|f_\theta(x)\|_2$ induces the retrieval score $s(q,t) = z(q)^\top z(t)$. Given matched calibration pairs $\mathcal{D} = \{(q_i, t_i^+)\}_{i=1}^N$, RONA constructs a low-bit encoder while keeping the pretrained backbone frozen. Both query and target embeddings guide calibration.

Figure 2 connects the encoder-level objective to local layer design. For $Y = XW^\top + \mathbf{1}b^\top$, we write $X \in \mathbb{R}^{n\times d_{\text{in}}}$, $W \in \mathbb{R}^{d_{\text{out}}\times d_{\text{in}}}$, and $b \in \mathbb{R}^{d_{\text{out}}}$, with n activation rows and an all-ones vector $\mathbf{1}$. RONA learns a rotation, constructs safe weight modifications, and calibrates their strengths against quantized embeddings. Rotation, subspace estimation, activation quantization, and GPTQ share the same input-channel partition. Weight groups use symmetric quantization with scale $s_w = \max(\|w\|_\infty, \varepsilon_w)/q_w^{\max}$; $q_w^{\max}$ is the largest positive code. The activation quantizer $Q_A(u;c)$ uses a token-wise, zero-inclusive asymmetric range scaled by group clipping ratio c, with a dynamic integer zero-point. Appendix B gives the complete quantizer definitions; Appendix A develops the equations below.

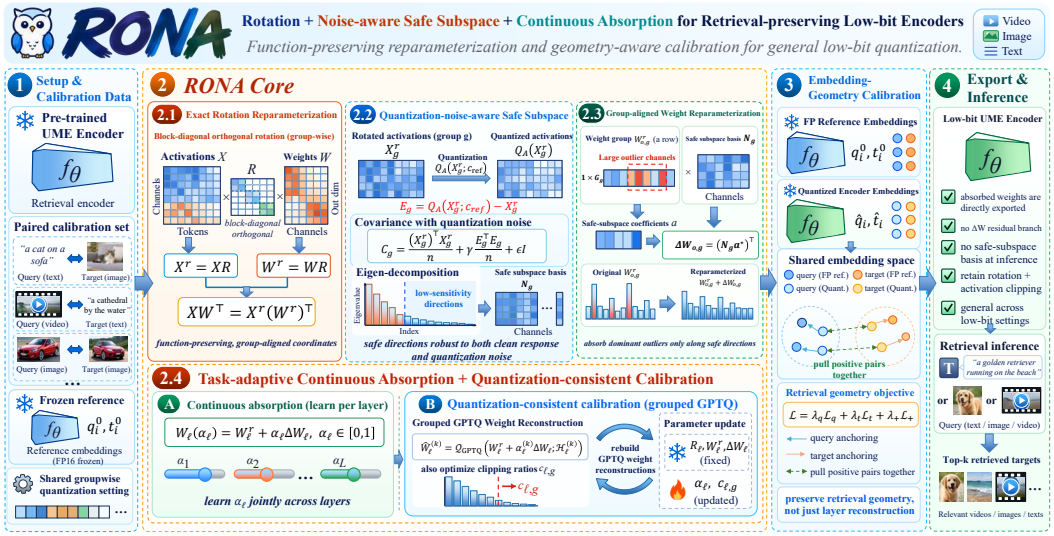

Figure 2: Overview of RONA. Group-aligned rotation and noise-aware weight reparameterization are followed by continuous absorption and embedding-geometry calibration with grouped GPTQ. The absorbed weights are exported directly.

3.2 Exact Rotation Reparameterization

We first choose quantization coordinates while preserving the unquantized function. For groups of width $G_g \le G$, let $R_g \in \mathbb{R}^{G_g \times G_g}$ be orthogonal and $R = \mathrm{diag}(R_1, \dots, R_m)$. Applying the same rotation to activations and weights gives

$$X^r = XR, \qquad W^r = WR, \qquad X^r (W^r)^\top = X \underbrace{RR^\top}_{I} W^\top = XW^\top,$$

with unchanged bias. We fit this transformation using activation quantization and a symmetric round-to-nearest (RTN) weight proxy:

$$\mathcal{L}_{\mathrm{rot}} = \frac{\| Q_A(XR; c_{\mathrm{ref}}) Q_W^{\mathrm{RTN}}(WR)^\top + \mathbf{1}b^\top - Y \|_F^2}{\|Y\|_F^2 + \varepsilon_{\mathrm{rot}}}. \tag{1}$$

Here $c_{\mathrm{ref}} = 1$ is the unclipped reference and $\varepsilon_{\mathrm{rot}} > 0$ stabilizes normalization. Each block uses a Cayley parameterization,

$$R_g = (I - A_g)^{-1}(I + A_g)H_g, \qquad A_g = S_g - S_g^\top,$$

where S_g is learned. The orthogonal initializer H_g is signed Hadamard for power-of-two widths and QR-based otherwise. Straight-through rounding supplies gradients; the best reconstruction among identity, initialization, and fitted iterates is retained. These layer-specific rotations are then fixed. Because every block acts only within its input group, the transform preserves the partition used for activation scales and weight reconstruction. This provides a common coordinate system for the subsequent subspace estimate and absorption solve.

3.3 Quantization-Noise-Aware Safe Subspace

In the rotated coordinates, define $E_g = Q_A(X_g^r; c_{\mathrm{ref}}) - X_g^r$. A weight direction v produces both a clean response $X_g^r v$ and a noise response $E_g v$. We select directions with low joint energy through the uncentered second moment

$$C_g = \underbrace{\frac{(X_g^r)^\top X_g^r}{n}}_{\textit{Clean response}} + \underbrace{\gamma \frac{E_g^\top E_g}{n}}_{\textit{Noise awareness}} + \underbrace{\epsilon I}_{\textit{Regularization}}, \qquad N_g = [v_{g,1}, \dots, v_{g,r}]. \tag{2}$$

where $N_g \in \mathbb{R}^{G_g \times r}$ contains the eigenvectors of the r smallest eigenvalues, γ weights noise sensitivity, and $\epsilon > 0$ regularizes the spectrum. For $v = N_g a$, $a \in \mathbb{R}^r$, its joint energy is $a^\top N_g^\top C_g N_g a$. For

$\gamma > 0$, this energy controls the quantized response:

$$\frac{\|(X_g^r + E_g)v\|_2^2}{n} \le (1 + \gamma^{-1})\left(v^\top C_g v - \epsilon\|v\|_2^2\right). \tag{3}$$

Thus, the retained eigenspace provides low-sensitivity coordinates for weight modification. Each basis is estimated at c_{ref} and held fixed during subsequent absorption and clipping calibration. The two moments use the same calibration tokens, so the noise term measures sensitivity to the actual activation quantizer in the learned coordinates. Setting $\gamma = 0$ recovers the clean-only construction used in the ablation and directional probe.

3.4 Group-Aligned Weight Reparameterization

Within these coordinates, RONA reduces a quadratic surrogate that emphasizes weight outliers. For output row o and input group g, let $w = (W_{o,g}^r)^\top$, $D_w = \operatorname{diag}(\omega)$, and $\omega = \operatorname{softmax}(|w|/\tau)$ with temperature $\tau > 0$. We solve

$$a^\star = \arg\min_a \underbrace{(w + N_g a)^\top D_w (w + N_g a)}_{\textit{Outlier suppression}} + \underbrace{\mu\|a\|_2^2}_{\textit{Update size}} + \underbrace{\nu(\mathbf{1}^\top N_g a)^2}_{\textit{Shift control}}. \tag{4}$$

The first term suppresses originally large coordinates; $\mu > 0$ controls modification size and $\nu \ge 0$ penalizes group-wide shift. With $v_g = N_g^\top \mathbf{1}$, the solution is

$$a^\star = -\left(N_g^\top D_w N_g + \mu I + \nu v_g v_g^\top\right)^{-1} N_g^\top D_w w, \qquad \Delta W_{o,g} = (N_g a^\star)^\top. \tag{5}$$

The outlier weights are computed from the original rotated row and remain fixed during the solve. The row-group updates are assembled into a fixed ΔW. Sharing group boundaries throughout the procedure keeps each modification aligned with the scales and error compensation used by the quantizers.

3.5 Task-Adaptive Continuous Absorption

The local solve determines a direction; encoder-level calibration determines how strongly each layer uses it. For layer ℓ, we set

$$W_\ell(\alpha_\ell) = \underbrace{W_\ell^r}_{\textit{Rotated weights}} + \underbrace{\alpha_\ell \Delta W_\ell}_{\textit{Adaptive absorption}}, \qquad \alpha_\ell \in [0, 1]. \tag{6}$$

A single coefficient controls all groups of each Linear layer, interpolating from rotation alone to full absorption. We jointly optimize all α_ℓ and group clipping ratios $c_{\ell,g}$, while W_ℓ^r, ΔW_ℓ, and R_ℓ remain fixed. This joint optimization accounts for their combined effect on the final retrieval embedding. The resulting coefficient is shared across inputs but specific to its layer, allowing different parts of the encoder to use different amounts of the same precomputed reparameterization.

3.6 Quantization-Consistent GPTQ Calibration

To expose calibration to the deployed weight grid, iteration k materializes the grouped-GPTQ reconstruction

$$\widehat{W}_\ell^{(k)} = \mathcal{Q}_{\text{GPTQ}}\left(W_\ell^r + \alpha_\ell^{(k)} \Delta W_\ell; \mathcal{H}_\ell^{(k)}\right), \tag{7}$$

where $\mathcal{Q}_{\text{GPTQ}}$ includes dequantization and $\mathcal{H}_\ell^{(k)}$ is a damped, group-block-diagonal activation second moment. Let $A_\ell = Q_A(X_\ell R_\ell; c_\ell)$, with $c_\ell = \{c_{\ell,g}\}_g$. We use the following surrogate for the layer output:

$$\widetilde{Y}_\ell = \underbrace{A_\ell(\widehat{W}_\ell^{(k)})^\top + \mathbf{1}b_\ell^\top}_{\textit{GPTQ forward}} + \underbrace{(\alpha_\ell - \alpha_\ell^{(k)})\,\mathrm{sg}\left[A_\ell \Delta W_\ell^\top\right]}_{\textit{Absorption-direction gradient}}. \tag{8}$$

At $\alpha_\ell = \alpha_\ell^{(k)}$, the correction vanishes and the forward pass uses the actual GPTQ reconstruction. The stop-gradient operator sg supplies the local absorption-direction derivative. Clipping is differentiated through the first term with straight-through activation rounding and a detached integer zero-point.

After each update, we project absorption and clipping to their feasible ranges, collect all quantized layer inputs from the current encoder, and rebuild the group Hessians and GPTQ weights. Collection precedes reconstruction so that all Hessians describe the same pre-refresh encoder. The initial Hessians use frozen-encoder inputs at reference clipping, and the final post-update reconstruction is exported. Appendix B.3 details this alternating procedure.

3.7 Embedding-Geometry Calibration Objective

The calibration signal acts on both endpoints of a retrieval pair. Let $q_i^0, t_i^0 \in \mathbb{R}^d$ be normalized full-precision embeddings and $\widehat{q}_i, \widehat{t}_i$ their quantized counterparts. We combine two-sided anchoring with positive-pair attraction:

$$\mathcal{L}_q = \frac{1}{Nd}\sum_i \|\widehat{q}_i - q_i^0\|_2^2, \qquad \mathcal{L}_t = \frac{1}{Nd}\sum_i \|\widehat{t}_i - t_i^0\|_2^2,$$
$$\mathcal{L}_+ = \frac{1}{N}\sum_i \left[1 - \cos(\widehat{q}_i, \widehat{t}_i)\right], \tag{9}$$
$$\mathcal{L} = \underbrace{\lambda_q \mathcal{L}_q}_{\textit{Query anchoring}} + \underbrace{\lambda_t \mathcal{L}_t}_{\textit{Target anchoring}} + \underbrace{\lambda_+ \mathcal{L}_+}_{\textit{Pair attraction}}.$$

Anchoring retains the reference coordinate system, while attraction brings matched endpoints together. The positive-pair term directly increases the similarity of quantized matches; reference anchoring supplies the complementary constraint that keeps both endpoints aligned with the pretrained embedding space. We balance their magnitudes using a calibration-only RTN trajectory: $\lambda_+ = 1$, $\lambda_q = \overline{\mathcal{L}}_+/\overline{\mathcal{L}}_q$, and $\lambda_t = \overline{\mathcal{L}}_+/\overline{\mathcal{L}}_t$, where bars denote trajectory averages. Query- and target-side gradients are accumulated at unchanged parameters before each optimizer step, yielding both derivatives of the same pair objective.

Export and inference. The final weight reconstruction incorporates absorption directly. Inference retains the online rotation and activation clipping, with the subspace basis and calibration-only correction removed. Appendix B.5 gives the resulting layer expression and implementation details.

4 Experiments

4.1 Experimental Setup

Quantization and calibration. Our main experiments use Qwen3-VL-Embedding-8B (Li et al., 2026a), quantizing 368 Linear layers to W4A4 with 128-channel groups. We apply symmetric grouped GPTQ (Frantar et al., 2023) to weights and dynamic asymmetric quantization to activations. Calibration uses 25 training pairs from distinct subsets, with both query and target inputs. Rotations are fitted for 16 Adam steps at learning rate 10^{-2}; each group retains 32 safe directions with $\gamma = 1$. Absorption coefficients and clipping ratios are jointly optimized for four updates at learning rate 0.05, refreshing GPTQ after each update. Full hyperparameters, calibration sampling, and evaluation settings are in Appendix B.

Evaluation and baselines. We evaluate all 78 MMEB-V2 tasks (Meng et al., 2026): 36 image, 18 video, and 24 visual-document tasks. We use Hit@1 for images and videos and nDCG@5 for visual documents; modality and All scores are task-wise arithmetic means. PTQ comparisons include SmoothQuant (Xiao et al., 2023), QuaRot (Ashkboos et al., 2024), DuQuant (Lin et al., 2024), SpinQuant (Liu et al., 2025), and FlatQuant (Sun et al., 2025). The 7B baseline group includes CAFe (Yu et al., 2025), UME-R1 (Lan et al., 2026), TTE$_s$ (Cui et al., 2026), RzenEmbed-V2 (Jian et al., 2025), Embed-RL (Jiang et al., 2026), MMEmb-R1 (Wang et al., 2026b), and UEmbed-dense (Song et al., 2026), with model sizes shown individually. Full-precision reference scores are our reproductions. Task-wise results and additional precision settings are organized in Appendices D and E.

4.2 Main Results

Table 1 compares RONA with multimodal embedders and W4A4 PTQ baselines. RONA achieves **73.55**, retaining **94.4%** of the full-precision score (77.93), and exceeds UEmbed-dense (9B) and DuQuant by **1.75** and **3.65** points, respectively.

Table 1: Comparison with recent multimodal embedders and W4A4 PTQ baselines on MMEB-V2. CLS: classification; QA: question answering; RET: retrieval; GD: grounding; MRET: moment retrieval; VDR: ViDoRe; VR: VisRAG; OOD: out-of-domain. Best non-reference results are bold.

	Image					Video					VisDoc					All
Model	CLS	QA	RET	GD	Overall	CLS	QA	RET	MRET	Overall	VDRv1	VDRv2	VR	OOD	Overall	Overall
# Datasets	10	10	12	4	36	5	5	5	3	18	10	4	6	4	24	78
Full-Precision Reference																
Qwen3-VL-Embedding-8B	74.19	81.23	80.11	92.15	80.11	79.15	71.53	59.07	55.75	67.55	87.28	70.05	88.71	73.37	82.45	77.93
7B baseline																
CAFe (7B)	63.60	61.70	69.10	87.60	67.60	35.80	58.70	34.40	39.50	42.40	70.70	49.60	79.50	38.10	63.90	60.60
UME-R1 (7B)	67.10	69.20	71.90	84.90	71.30	48.60	60.70	38.20	39.30	47.50	75.70	50.50	83.70	37.60	67.10	64.50
TTE_s (7B)	69.70	72.40	74.00	90.60	74.20	49.10	60.60	36.40	37.20	46.80	84.10	**62.70**	**91.90**	47.60	76.40	68.60
RzenEmbed-V2 (7B)	**70.60**	71.70	**78.50**	92.10	**75.90**	58.80	63.50	51.00	45.50	55.70	**89.70**	60.70	88.70	34.70	75.50	71.10
Embed-RL (4B)	63.70	70.50	71.30	91.40	70.10	57.60	58.40	45.10	49.50	53.00	80.20	53.40	84.90	67.10	74.70	68.10
MMEmb-R1 (4B)	67.70	74.20	74.50	**94.90**	74.80	60.70	61.10	51.60	50.60	56.60	83.00	60.60	83.70	66.60	76.70	71.20
UEmbed-dense (9B)	64.70	73.90	73.90	90.60	73.20	63.00	65.60	51.80	**53.50**	59.00	85.50	59.10	87.90	**70.40**	**79.20**	71.80
W4A4 PTQ Baselines																
SmoothQuant	59.98	68.06	67.73	75.02	66.48	64.80	61.08	47.83	44.14	55.61	79.95	57.95	80.70	66.05	74.15	66.33
QuaRot	64.01	70.93	69.77	81.82	69.83	69.50	63.86	51.07	49.46	59.48	81.47	58.48	82.02	66.70	75.32	69.13
DuQuant	63.30	72.27	71.25	83.05	70.64	70.58	64.11	51.81	49.68	60.09	82.33	60.23	82.30	67.42	76.16	69.90
SpinQuant	63.14	71.43	70.08	81.88	69.84	67.79	64.28	50.79	48.74	58.92	81.60	59.22	81.60	65.74	75.23	68.97
FlatQuant	63.40	71.03	69.03	82.25	69.49	67.87	64.24	50.89	48.57	58.93	80.94	55.20	81.63	64.80	74.13	68.48
Ours																
RONA (W4A4)	68.45	**75.03**	75.21	88.58	74.77	**74.17**	**68.13**	**55.27**	53.48	**63.79**	85.81	60.65	86.44	69.49	79.06	**73.55**

Compared with DuQuant, RONA improves Image, Video, and VisDoc by 4.13, 3.70, and 2.90 points. These gains span all three modalities while both weights and activations use 4-bit representations, supporting retrieval-preserving compression of the direct encoder.

4.3 Ablation Study

Contribution of RONA components. Table 2 isolates noise awareness, group alignment, learned absorption, and the embedding-geometry objective under the same W4A4 backbone and evaluation protocol. The variants replace one component at a time; their settings are specified in Appendix B.4.

Table 2: Ablation study on MMEB-V2 under W4A4. Red downward arrows report the absolute decrease from full RONA.

#	Setting	Image		Video		VisDoc		All	
1	RONA	**74.77**		**63.79**		**79.06**		**73.55**	
2	w/o Noise Awareness	73.61	↓1.16	62.17	↓1.62	77.77	↓1.29	72.25	↓1.30
3	w/o Group Alignment	74.09	↓0.68	62.84	↓0.95	78.41	↓0.65	72.82	↓0.73
4	Fixed Absorption	74.26	↓0.51	63.10	↓0.69	78.64	↓0.42	73.03	↓0.52
5	Local Reconstruction	73.85	↓0.92	62.59	↓1.20	78.08	↓0.98	72.55	↓1.00

Removing noise awareness yields the largest overall drop (1.30 points), including 1.62 points on video, highlighting the role of activation noise in safe-direction selection. Local reconstruction loses 1.00 point, supporting calibration against encoder-level retrieval geometry. Removing group alignment and fixing absorption reduce the overall score by 0.73 and 0.52 points, respectively. Together, the

ablations show the value of matching both the modification directions and their layer-wise strengths to the quantized retrieval encoder.

4.4 Generalization Across Model Scale and Architecture

We next evaluate Qwen3-VL-Embedding-2B to test transfer across parameter scales. Table 3 reports the full-precision reference and the same five W4A4 PTQ baselines.

Table 3: W4A4 results on Qwen3-VL-Embedding-2B.

Method	Image	Video	VisDoc	All
FP16	74.98	62.15	79.16	73.31
SmoothQuant	50.56	42.60	58.83	51.27
QuaRot	57.80	47.36	66.51	58.07
DuQuant	55.70	44.65	67.29	56.72
SpinQuant	60.11	49.14	66.56	59.56
FlatQuant	58.13	49.56	65.12	58.30
RONA	**66.43**	**55.16**	**72.70**	**65.76**

RONA obtains **65.76**, outperforming the strongest PTQ baseline by **6.20** points overall. Relative to the best baseline in each modality, gains are 6.33, 5.60, and 5.41 points on Image, Video, and VisDoc. We further evaluate the distinct WeMM-Embedding-2B backbone (Zhou et al., 2026) in Table 4.

Table 4: W4A4 results on WeMM-Embedding-2B.

Method	Image	Video	VisDoc	All
FP16	79.63	70.74	80.80	77.94
SmoothQuant	71.19	60.97	75.51	70.16
QuaRot	74.63	65.81	77.05	73.34
DuQuant	74.51	64.70	77.46	73.16
SpinQuant	74.91	66.22	76.67	73.44
FlatQuant	75.10	66.34	77.10	73.69
RONA	**77.14**	**68.64**	**79.73**	**75.98**

On WeMM, RONA reaches **75.98**, improving the strongest baseline by **2.29** points and retaining **97.5%** of the FP16 score. The corresponding modality gains are 2.04, 2.30, and 2.27 points. Alongside the 8B results, these improvements establish the effectiveness of RONA across model scales and backbones.

4.5 Deep Analysis

Bilateral retrieval geometry. We compare normalized Qwen3-VL-Embedding-8B embeddings across 78 tasks, with QuaRot–GPTQ labeled Vanilla W4A4. For each query, full precision fixes the highest-scoring relevant and irrelevant targets; their similarity difference defines margin m. Figure 3(a) plots mean drift $1 - \cos(z^{\mathrm{FP}}, z^{M})$ horizontally for queries and targets. Panel (b) plots mean $|m_M - m_{\mathrm{FP}}|$ horizontally by modality. Points are task-wise means with paired-bootstrap 95% intervals (Appendix C.1). These measures test endpoint alignment and relative retrieval geometry.

RONA reduces query/target drift from 0.157/0.132 to 0.113/0.093 and overall margin deviation from 0.045 to 0.037, with improvements across all three modalities. Thus, closer endpoint alignment is accompanied by better preservation of relative retrieval margins.

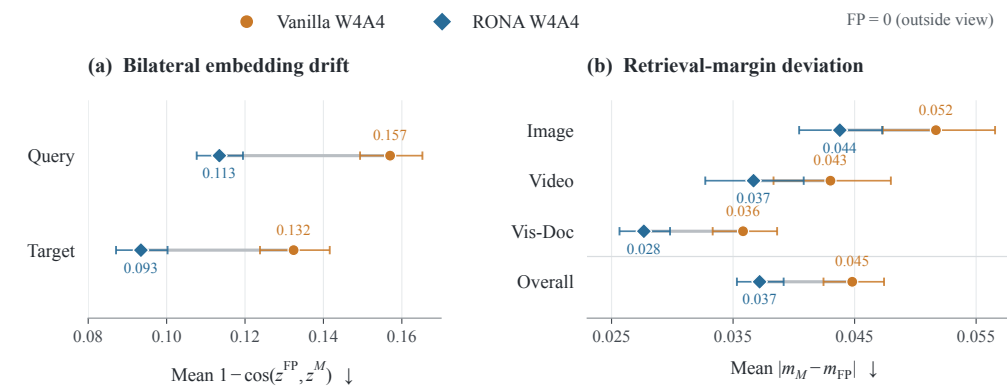

Figure 3: Bilateral embedding drift and retrieval-margin deviation under W4A4. Lower is better.

Why noise-aware directions matter. We inject a local perturbation into the FP32 encoder at decoder block 6, 18, or 30 and run the complete remaining network, using up to 128 queries per task. Both bases use the same 25 calibration pairs and rotations. Figure 4(a) uses test-time clean sensitivity $\|X_g^r v\|_2$ horizontally and noise sensitivity $\|E_g v\|_2$ vertically, both on logarithmic axes, for unit safe directions v; lower positions indicate weaker noise response. For a fixed clean-only update, we compare $\delta_o = (X^r + E)\Delta W^\top$ with a random direction and the direction after removing $E\Delta W^\top$, matching every token's norm. Panels (b–c) place interventions on the horizontal axis and margin deviation/initially correct ranking flips on the vertical axes. Appendix C.2 gives the protocol.

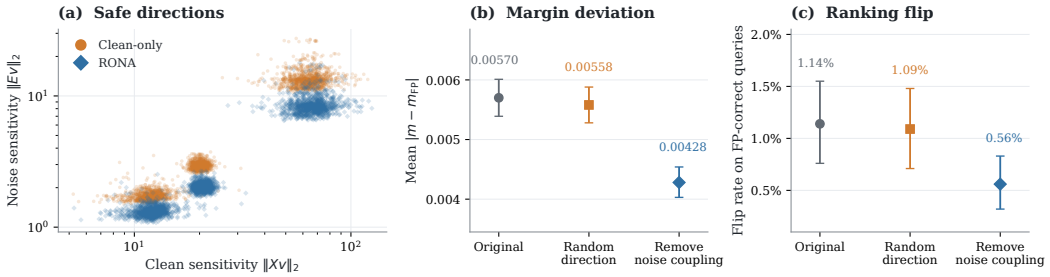

Figure 4: Noise-aware directions and norm-matched interventions on retrieval margins and rankings.

In the displayed comparison, randomization changes margin deviation from 0.00570 to 0.00558 and flips from 1.14% to 1.09%. Removing noise coupling reduces them to 0.00428 and 0.56%. This fixed-magnitude contrast identifies activation-noise coupling as a source of retrieval distortion, motivating RONA's directional constraint.

Together, the two analyses connect the local design to its retrieval objective. The full-encoder comparison measures preservation of both endpoints and their relative margins; the norm-matched intervention isolates sensitivity to the noise-coupling direction. These complementary observations motivate choosing noise-aware weight directions and calibrating their layer-wise strengths against the geometry of the complete encoder.

5 Conclusion

We present RONA, a retrieval-aware PTQ method that combines exact rotations, noise-aware safe subspaces, and continuous absorption calibrated against grouped-GPTQ reconstructions and bilateral embedding geometry. Across 78 MMEB-V2 tasks, it improves the strongest W4A4 baseline by 3.65, 6.20, and 2.29 points on Qwen3-VL-Embedding-8B, Qwen3-VL-Embedding-2B, and WeMM-Embedding-2B, respectively. Ablations and norm-matched interventions connect these gains to controlling activation-noise coupling and preserving retrieval margins. RONA establishes retrieval geometry as a practical calibration target for low-bit multimodal encoders. Integrating packed low-bit kernels is a natural direction for translating these accuracy gains into efficient deployment.

AI USE STATEMENT

We used generative AI tools to polish the manuscript, assist in drafting the Method section and its mathematical exposition, and support the development of experimental code. The research idea, method design, and experimental conception originated with the authors, who conducted the experiments and produced and interpreted the reported results. AI assistance was used to express and implement the authors' research rather than to generate experimental observations. The authors reviewed the AI-assisted text and code and take full responsibility for the final manuscript, technical claims, implementation, and results.

ETHICS STATEMENT

This work studies post-training quantization of pretrained multimodal retrieval encoders using established benchmarks and calibration pairs from their training subsets. Our experiments involve no new human-subject study or collection of user data. We recognize that compressed encoders can inherit biases and privacy risks from their pretrained models and data, and that broader access to retrieval systems requires responsible use. Released materials will respect the licenses and usage terms of the underlying models and datasets. We encourage downstream deployments to assess retrieval quality across user groups and input modalities and to apply appropriate access controls when indexing sensitive content.

REPRODUCIBILITY STATEMENT

We will publicly release the RONA implementation, the calibration set used in our experiments, and the configurations for quantization and evaluation. Section 3 specifies the algorithm, and Appendix A provides the notation and derivations. Section 4.1 and Appendix B describe the quantizers, calibration sampling, hyperparameters, and grouped-GPTQ refresh procedure. The protocols for the geometry and direction analyses are detailed in Appendix C, while Appendices D and E provide task-wise and cross-precision results. These materials document the experimental pipeline and support independent reproduction and further development.

REFERENCES

Yamato Arai and Yuma Ichikawa. Quantization Error Propagation: Revisiting Layer-Wise Post-Training Quantization. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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.

Haonan Chen, Chu Li, Zhicheng Wang, Yuanwei Liu, Yuanjiang Wang, Shaohua Jiang, and Zhicheng Dou. Douyin Multimodal Embedding Model Technical Report. *arXiv preprint arXiv:2608.02148*, 2026a.

Yuxiang Chen, Michael Beyer, Jun Zhu, and Jianfei Chen. Why Does Post-Training Quantization Work? *arXiv preprint arXiv:2609.11716*, 2026b.

Jianpeng Cheng, Xian Wu, Jiangfan Zhang, Wentao Bao, Chaitanya Ahuja, Shlok Kumar Mishra, Hanchao Yu, Yang Gao, Fan Xia, Qi Guo, Shaodan Zhai, Xiangjun Fan, and Jun Xiao. TTE-Flash: Accelerating Reasoning-based Multimodal Representations via Think-Then-Embed Tokens. *arXiv preprint arXiv:2605.16638*, 2026.

Xuanming Cui, Jianpeng Cheng, Hong-You Chen, Satya Narayan Shukla, Abhijeet Awasthi, Xichen Pan, Chaitanya Ahuja, Shlok Mishra, Taipeng Tian, Qi Guo, Ser-Nam Lim, Aashu Singh, and Xiangjun Fan. Think Then Embed: Generative Context Improves Multimodal Embedding. In *International Conference on Learning Representations*, 2026.

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.

Weijian Jian, Yajun Zhang, Dawei Liang, Chunyu Xie, Yixiao He, Dawei Leng, and Yuhui Yin. RzenEmbed: Towards Comprehensive Multimodal Retrieval. *arXiv preprint arXiv:2510.27350*, 2025.

Haonan Jiang, Yuji Wang, Yongjie Zhu, Xin Lu, Wenyu Qin, Meng Wang, Pengfei Wan, and Yansong Tang. Embed-RL: Reinforcement Learning for Reasoning-Driven Multimodal Embeddings. In *Computer Vision – ECCV 2026*, Lecture Notes in Computer Science, pp. 39–56. Springer Nature Switzerland, 2026.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: Training Vision-Language Models for Massive Multimodal Embedding Tasks. In *International Conference on Learning Representations*, 2025.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. UME-R1: Exploring Reasoning-Driven Generative Multimodal Embeddings. In *International Conference on Learning Representations*, 2026.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A Unified Framework for State-of-the-Art Multimodal Retrieval and Ranking. *arXiv preprint arXiv:2601.04720*, 2026a.

Zhikai Li, Zhen Dong, Xuewen Liu, Jing Zhang, and Qingyi Gu. OSAQ: Outlier Self-Absorption for Accurate Low-bit LLM Quantization. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026b.

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. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. SpinQuant: LLM Quantization with Learned Rotations. In *International Conference on Learning Representations*, 2025.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. VLM2Vec-V2: Advancing Multimodal Embedding for Videos, Images, and Visual Documents. *Transactions on Machine Learning Research*, 2026.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Gao Peng, Yu Qiao, and Ping Luo. OmniQuant: Omnidirectionally Calibrated Quantization for Large Language Models. In *International Conference on Learning Representations*, 2024.

Tingyu Song, Mingxin Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Zhijie Nie, Yilun Zhao, and Shu Wu. UEmbed: Unified Sparse and Dense Multimodal Embeddings. *arXiv preprint arXiv:2608.02583*, 2026.

Yuxuan Sun, Ruikang Liu, Haoli Bai, Han Bao, Kang Zhao, Yuening Li, Jiaxin Hu, Xianzhi Yu, Lu Hou, Chun Yuan, Xin Jiang, Wulong Liu, and Jun Yao. FlatQuant: Flatness Matters for LLM Quantization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 57587–57613. PMLR, 2025.

Zelong Sun, Jun Wang, Kaicheng Yang, Tiancheng Gu, Ziyong Feng, and Zhiwu Lu. Learning from Failures: Retrieval-Centric CoT via Hard Negatives for Unified Multimodal Retrieval. *arXiv preprint arXiv:2608.06060*, 2026.

Shijie Wang, Xiangzhao Hao, Yueti Li, Guangyu Cao, Xinyu Tang, and Haiyun Guo. ReLoop-UME: Recurrent Depth with Learnable Retrieval Registers for Universal Multimodal Embedding. *arXiv preprint arXiv:2607.28751*, 2026a.

Yuchi Wang, Dingkang Yang, Haiyang Yu, Weikang Bian, Jiefeng Long, Xiao Liang, Chao Feng, and Hongsheng Li. MMEmb-R1: Reasoning-Enhanced Multimodal Embedding with Pair-Aware Selection and Adaptive Control. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026b.

Peixi Wu, Ke Mei, Feipeng Ma, Bosong Chai, Zhibin Lan, Chenxi Zhao, Shannan Yan, Jie Chen, Zhangchi Hu, Yansong Peng, Bo Lin, Junjie Zhou, Dacheng Yin, Tianyi Wang, Fengyun Rao, Jing Lyu, Hebei Li, and Xiaoyan Sun. Beyond Chain-of-Thought: Rewrite as a Universal Interface for Generative Multimodal Embeddings. *arXiv preprint arXiv:2604.22280*, 2026a.

Peixi Wu, Biao Yang, Feipeng Ma, Bosong Chai, Bo Lin, Wei Yuan, Fan Yang, Tingting Gao, Hebei Li, and Xiaoyan Sun. LaME: Learning to Think in Latent Space for Multimodal Embedding via Information Bottleneck. *arXiv preprint arXiv:2606.13061*, 2026b.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 38087–38099. PMLR, 2023.

Hao Yu, Zhuokai Zhao, Shen Yan, Lukasz Korycki, Jianyu Wang, Baosheng He, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, and Hanchao Yu. CAFe: Unifying Representation and Generation with Contrastive-Autoregressive Finetuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 6345–6356, 2025.

Longxiang Zhang, Weilong Dai, Guanghao Zhang, Hao Jiang, and Pipei Huang. Think When Needed: Adaptive Reasoning-Driven Multimodal Embeddings with a Dual-LoRA Architecture. *arXiv preprint arXiv:2605.14448*, 2026.

Junjie Zhou, Ke Mei, Lei Li, Tianyi Wang, Fengyun Rao, and Jing Lyu. WeMM-Embedding: WeChat Multi-Modal Embedding Technical Report. *arXiv preprint arXiv:2608.24053*, 2026.

A NOTATION AND DERIVATIONS

This section supplies the derivations underlying Section 3 and Figures 1–2. Activations are row vectors; a weight row is transposed when treated as a column direction. All norms are Euclidean or Frobenius unless a subscript specifies otherwise.

A.1 NOTATION

Table A.1: Symbols used in the method and its derivations.

Symbol	Definition
$f_\theta, z(x), d$	Pretrained encoder, normalized embedding, and embedding dimension.
$\mathcal{D}, N$	Matched calibration pairs and their number.
ℓ, o, g	Linear-layer, output-row, and input-group indices.
X, W, b, Y	Token activations, weights, bias, and output of a Linear layer.
$n, d_{\text{in}}, d_{\text{out}}$	Number of token rows and layer input/output widths.
G, G_g, m	Nominal group width, actual width of group g, and number of groups.
R_g, R, X^r, W^r	Orthogonal group rotation, block-diagonal rotation, XR, and WR.
E_g, C_g, N_g, r	Activation quantization error, joint second moment, safe basis, and rank.
γ, ϵ	Noise-energy coefficient and spectral regularizer.
$w, D_w, a, \Delta W$	Group weight column, outlier weights, basis coefficients, and assembled update.
τ, μ, ν	Outlier temperature, coefficient regularizer, and shift regularizer.
$\alpha_\ell, c_{\ell,g}$	Layer absorption coefficient and group activation clipping ratio.
$\widehat{W}_\ell^{(k)}, \mathcal{H}_\ell^{(k)}$	GPTQ reconstruction and groupwise Hessian approximation at step k.
$q_i^0, t_i^0, \widehat{q}_i, \widehat{t}_i$	Normalized reference and quantized embeddings of a matched pair.
$\text{sg}, I, \mathbf{1}$	Stop-gradient operator, identity matrix, and all-ones vector.

The dimensions of I and $\mathbf{1}$ follow the expression: $\mathbf{1} \in \mathbb{R}^n$ in a layer bias, and $\mathbf{1} \in \mathbb{R}^{G_g}$ in the shift penalty. The basis satisfies $N_g^\top N_g = I_r$. The symbol N counts calibration pairs, whereas n counts activation tokens.

A.2 ORTHOGONAL ROTATION AND RECONSTRUCTION

For a skew-symmetric A_g, define $P_g = I - A_g$ and $Q_g = I + A_g$. Then $P_g^\top = Q_g$, and P_g, Q_g commute because both are polynomials in A_g. Since the eigenvalues of a real skew-symmetric matrix are imaginary, P_g is invertible. The Cayley factor $T_g = P_g^{-1} Q_g$ satisfies

$$T_g^\top T_g = P_g Q_g^{-1} P_g^{-1} Q_g = I. \tag{A.1}$$

An orthogonal initializer H_g therefore yields $R_g = T_g H_g$ with $R_g R_g^\top = I$. Stacking these blocks preserves orthogonality and the group partition, giving

$$(XR)(WR)^\top + \mathbf{1}b^\top = XRR^\top W^\top + \mathbf{1}b^\top = Y. \tag{A.2}$$

This is the function-preserving identity in Figure 2. The quantized output differs by

$$\Delta Y_{\text{rot}} = Q_A(XR; c_{\text{ref}}) Q_W^{\text{RTN}}(WR)^\top - XW^\top. \tag{A.3}$$

Normalizing $\|\Delta Y_{\text{rot}}\|_F^2$ by $\|Y\|_F^2 + \varepsilon_{\text{rot}}$ gives Equation (1). Rotation fitting thus selects coordinates using their quantized reconstruction quality within an exactly equivalent unquantized family.

A.3 SAFE EIGENSPACE AND QUANTIZED-RESPONSE BOUND

For a weight direction $v \in \mathbb{R}^{G_g}$, define its regularized joint energy

$$J_g(v) = \frac{\|X_g^r v\|_2^2}{n} + \gamma \frac{\|E_g v\|_2^2}{n} + \epsilon \|v\|_2^2 = v^\top C_g v. \tag{A.4}$$

For an orthonormal basis $B \in \mathbb{R}^{G_g \times r}$, the sum of its directional energies is $\mathrm{tr}(B^\top C_g B)$. Write $C_g = U \,\mathrm{diag}(\lambda_1, \ldots, \lambda_{G_g}) U^\top$, with ascending eigenvalues, and $P = U^\top B$. Then

$$\mathrm{tr}(B^\top C_g B) = \sum_{j=1}^{G_g} \lambda_j \|P_{j,:}\|_2^2, \quad 0 \le \|P_{j,:}\|_2^2 \le 1, \quad \sum_j \|P_{j,:}\|_2^2 = r. \tag{A.5}$$

The minimum is obtained by assigning unit mass to the r smallest eigenvalues, yielding $B = N_g$. For $v = N_g a$, $J_g(v) = \sum_{j=1}^r \lambda_j a_j^2 \le \lambda_r \|a\|_2^2$.

To obtain Equation (3), set $u = X_g^r v$ and $e = E_g v$. For $\gamma > 0$, $0 \le \|\gamma^{-1/2} u - \gamma^{1/2} e\|_2^2$ gives $2u^\top e \le \gamma^{-1}\|u\|_2^2 + \gamma \|e\|_2^2$. Hence

$$\begin{aligned} \|u + e\|_2^2 &\le (1 + \gamma^{-1})\|u\|_2^2 + (1 + \gamma)\|e\|_2^2 \\ &= (1 + \gamma^{-1})\big(\|u\|_2^2 + \gamma \|e\|_2^2\big). \end{aligned} \tag{A.6}$$

Dividing by n and substituting Equation (A.4) proves the stated bound. For a group update ΔW_g, the corresponding output change decomposes exactly as

$$Q_A(X_g^r; c_{\mathrm{ref}}) \Delta W_g^\top = X_g^r \Delta W_g^\top + E_g \Delta W_g^\top. \tag{A.7}$$

The second term is the activation-noise coupling highlighted in Figures 1 and 4. Including $E_g^\top E_g$ selects directions with low response to this component as well as to the clean input.

A.4 Closed-Form Group Absorption

Fix w, its softmax weights D_w, and N_g. Let $v_g = N_g^\top \mathbf{1}$, $h = N_g^\top D_w w$, and $K = N_g^\top D_w N_g + \mu I + \nu v_g v_g^\top$. Expanding Equation (4) gives

$$J(a) = w^\top D_w w + 2h^\top a + a^\top K a, \qquad \nabla_a J = 2h + 2Ka. \tag{A.8}$$

For $\mu > 0$, K is positive definite. Setting the gradient to zero yields $a^\star = -K^{-1}h$ and $\Delta W_{o,g} = (N_g a^\star)^\top$, as in Equation (5). The shift penalty follows from $(\mathbf{1}^\top N_g a)^2 = a^\top v_g v_g^\top a$.

Continuous absorption scales these coefficients by α. Since $K a^\star = -h$,

$$J(\alpha a^\star) = J(0) - (2\alpha - \alpha^2) h^\top K^{-1} h, \qquad 0 \le \alpha \le 1. \tag{A.9}$$

This explains the interpolation in Equation (6): the local surrogate decreases along the chosen update, and retrieval calibration determines the strength used by the encoder. The actual weight scale is evaluated from the modified row,

$$s_w(\alpha) = \frac{\max(\|w + \alpha N_g a^\star\|_\infty, \varepsilon_w)}{q_w^{\max}}. \tag{A.10}$$

A.5 GPTQ Consistency and Absorption Gradients

For a quantized activation group A_g and weight-column error e_w, the mean reconstruction error is

$$\frac{1}{n}\|A_g e_w\|_2^2 = e_w^\top \left(\frac{A_g^\top A_g}{n} \right) e_w. \tag{A.11}$$

Its Hessian is twice the second moment; the constant factor leaves the quadratic minimizer unchanged. Independent group approximation and diagonal damping produce $\mathcal{H}_\ell^{(k)}$ in Equation (7). After fixing this reconstruction for differentiation, let $G_\ell = \partial \mathcal{L} / \partial \widetilde{Y}_\ell$. Equation (8) has the exact forward value and local derivative

$$\widetilde{Y}_\ell \Big|_{\alpha_\ell^{(k)}} = A_\ell (\widehat{W}_\ell^{(k)})^\top + \mathbf{1} b_\ell^\top, \qquad \frac{\partial \mathcal{L}}{\partial \alpha_\ell} = \big\langle G_\ell, A_\ell \Delta W_\ell^\top \big\rangle_F. \tag{A.12}$$

The correction's stopped activation factor keeps clipping gradients on the quantized forward term. Thus $\partial \mathcal{L} / \partial c_{\ell,g}$ is obtained through $A_\ell (\widehat{W}_\ell^{(k)})^\top$, using the activation-rounding surrogate. Re-estimating all group second moments after an optimizer step, followed by rebuilding all weights, advances the reconstruction point to $k + 1$.

A.6 Geometry Loss, Retrieval Margins, and Export

For unit vectors u, v, $\|u - v\|_2^2 = 2(1 - u^\top v)$. Consequently, the three terms in Equation (9) can be written as

$$\mathcal{L}_q = \frac{2}{Nd}\sum_i (1 - \widehat{q}_i^\top q_i^0), \quad \mathcal{L}_t = \frac{2}{Nd}\sum_i (1 - \widehat{t}_i^\top t_i^0), \quad \mathcal{L}_+ = \frac{1}{2N}\sum_i \|\widehat{q}_i - \widehat{t}_i\|_2^2. \tag{A.13}$$

The first two terms anchor each endpoint; the third attracts matched quantized endpoints. Choosing $\lambda_q = \overline{\mathcal{L}}_+/\overline{\mathcal{L}}_q$, $\lambda_t = \overline{\mathcal{L}}_+/\overline{\mathcal{L}}_t$, and $\lambda_+ = 1$ makes the three weighted reference means equal to $\overline{\mathcal{L}}_+$.

If $z = h/\|h\|_2$, its normalization Jacobian is $(I - zz^\top)/\|h\|_2$. For one positive pair and shared calibration parameters ϕ, the attraction gradient is

$$\nabla_\phi(1 - \widehat{q}^\top \widehat{t}) = -(\nabla_\phi \widehat{q})^\top \widehat{t} - (\nabla_\phi \widehat{t})^\top \widehat{q}. \tag{A.14}$$

Evaluating these two contributions separately at unchanged parameters and summing them gives the same joint gradient used in calibration.

To connect endpoint drift to Figure 1, let q, t_+, t_- denote unit reference embeddings and e_q, e_+, e_- their changes after quantization. Expanding $\widehat{m} = (q + e_q)^\top[(t_+ + e_+) - (t_- + e_-)]$ yields

$$\widehat{m} - m = e_q^\top(t_+ - t_-) + q^\top(e_+ - e_-) + e_q^\top(e_+ - e_-). \tag{A.15}$$

The projections onto the query and target-difference directions explain why perturbations with equal norm can affect rankings differently. Cauchy–Schwarz gives

$$|\widehat{m} - m| \le 2\|e_q\|_2 + \|e_+\|_2 + \|e_-\|_2 + \|e_q\|_2(\|e_+\|_2 + \|e_-\|_2). \tag{A.16}$$

Anchoring controls these endpoint deviations, while the noise-aware subspace acts earlier on the directional response of each weight modification.

At the final calibration step K, the correction in Equation (8) is zero and absorption is already incorporated into $\widehat{W}_\ell^{(K)}$. The exported layer is therefore

$$Y_\ell^{\text{export}} = Q_A(X_\ell R_\ell; c_\ell)\big(\widehat{W}_\ell^{(K)}\big)^\top + \mathbf{1}b_\ell^\top. \tag{A.17}$$

This is the inference computation depicted in Figure 2: the quantized weight, rotation, and clipping parameters suffice to evaluate the layer.

B Complete Experimental Settings

B.1 Quantizers and Target Layers

For an activation group $u \in \mathbb{R}^{G_g}$, clipping ratio c, and bit width b_A, the zero-inclusive range, scale, and integer zero-point are

$$\begin{aligned} l &= c\min(\min_j u_j, 0), & h &= c\max(\max_j u_j, 0), \\ s_A &= \frac{\max(h - l, 10^{-8})}{2^{b_A} - 1}, & z_A &= \text{clip}\big(\text{round}(-l/s_A), 0, 2^{b_A} - 1\big). \end{aligned} \tag{B.1}$$

Mapping a value to the affine grid, clamping its code, and reconstructing it gives

$$Q_A(u; c) = s_A\left[\text{clip}\big(\text{round}(u/s_A) + z_A, 0, 2^{b_A} - 1\big) - z_A\right]. \tag{B.2}$$

Thus A4 uses codes $0, \ldots, 15$. Both range endpoints are scaled toward zero by $c \in [0.5, 1]$. Quantizer statistics are recomputed per token and group; the zero-point is detached during calibration. Straight-through rounding passes the upstream gradient through the rounding operator. A shortened final group is zero-padded for grouped computation and trimmed after reconstruction.

For a weight row-group w, $q_W^{\min} = -2^{b_W - 1}$, $q_W^{\max} = 2^{b_W - 1} - 1$, and

$$s_W = \frac{\max(\|w\|_\infty, 10^{-8})}{q_W^{\max}}, \qquad Q_W^{\text{RTN}}(w) = s_W\,\text{clip}\big(\text{round}(w/s_W), q_W^{\min}, q_W^{\max}\big). \tag{B.3}$$

W4 therefore uses $[-8, 7]$ with zero-point zero and per-row, per-group scale. The rotation objective uses this RTN proxy; final weights use grouped GPTQ with the same grid. For the current quantized input A_g, damping is

$$H_g = \frac{A_g^\top A_g}{n}, \quad \eta_g = \max\big(0.01 \ \mathrm{mean}(\mathrm{diag}\, H_g), 10^{-5}\big), \quad \mathcal{H}_g = H_g + \eta_g I. \tag{B.4}$$

Within each group, GPTQ processes columns in order and compensates subsequent columns using the upper Cholesky factor of $\mathcal{H}_g^{-1}$. All group scales are determined from the corresponding pre-compensation weight rows.

The 8B model has 368 quantized Linear layers: 252 language-side projections (seven in each of 36 decoder blocks) and 116 visual-side projections. The latter comprise four projections in each of 27 vision blocks and eight merger projections. The language projections are the query, key, value, attention-output, gate, up, and down projections. Embedding lookup, normalization, and bias retain their reference representations.

B.2 Calibration Data and Hyperparameters

The calibration set contains one training pair from each of 25 subsets. Query and target sides share a budget of 64 tokens per record, initially divided equally and redistributed when one side is shorter. Longer sequences are sampled at evenly spaced positions. The per-layer cap is 2,048 tokens; the paired-record budget yields at most 1,600 tokens. The recorded 8B layer caches contain 1,544–1,595 tokens. Rotation fitting uses at most 128 cached tokens per layer.

The subsets comprise ImageNet-1K, N24News, HatefulMemes, VOC2007, SUN397, OK-VQA, A-OKVQA, DocVQA, InfographicsVQA, ChartQA, Visual7W, VisDial, CIRR, both VisualNews directions, both MSCOCO caption-retrieval directions, NIGHTS, WebQA, MSCOCO grounding, two video-caption subsets, one video-question-answering subset, ColPali training data, and VisRAG in-domain training data. The first valid training pair from each subset is used.

Table B.1: Main calibration settings for Qwen3-VL-Embedding-8B.

Component	Setting
Precision / group width	W4A4 / 128 input channels.
Rotation	Signed Hadamard initialization; Cayley parameterization; Adam, 16 steps, learning rate 0.01, seed 0.
Safe subspace	Rank $r = 32$, noise coefficient $\gamma = 1$, spectral regularizer $\epsilon = 10^{-5}$.
Group absorption	$\tau = 0.1$, $\mu = 10^{-3}$, $\nu = 10^{-3}$.
Absorption / clipping	One $\alpha_\ell \in [0, 1]$ per Linear; one $c_{\ell,g} \in [0.5, 1]$ per activation group, initialized at 1.
Joint optimization	Adam; four updates; learning rate 0.05 for both parameter families.
Loss weights	$\lambda_q = 4463.467478$, $\lambda_t = 5250.259029$, $\lambda_+ = 1$.
GPTQ refresh	After every joint update; final reconstruction after update four.
Compute	Eight NVIDIA A800 GPUs for calibration and benchmark evaluation.

Initial absorption coefficients are layer-specific calibration values: 333 layers start at 0, 25 at 0.25, five at 0.5, three at 0.75, and two at 1. The initialization mean is 0.03533. The coefficient vector is fixed before joint optimization; rotations, safe bases, and Hessians are constructed for the current model. Appendix B.6 distinguishes this stored initialization from continuous refinement and the Fixed Absorption control.

Loss balancing uses four calibration-only RTN updates with mean raw losses $\overline{\mathcal{L}}_q = 8.583885 \times 10^{-5}$, $\overline{\mathcal{L}}_t = 7.297524 \times 10^{-5}$, and $\overline{\mathcal{L}}_+ = 0.3831389225$. Their ratios give the weights in Table B.1, with each weighted reference contribution equal to 0.3831389225. Calibration and model selection use training data.

For the 2B experiments, W4A4 and the 78-task evaluation are shared. Table B.2 records the model-specific calibration configuration.

Table B.2: Model-specific calibration settings for the two 2B backbones.

Field	Qwen3-VL-Embedding-2B	WeMM-Embedding-2B
Weight / activation bits	4 / 4	4 / 4
Evaluated tasks	78	78
Target Linear count	300 (vision 104, language 196)	284 (vision 98, language 186)
Group width / safe rank	128 / 32	128 / 32
Calibration pairs / token budget	25 / 64 per pair; 2048 per layer rotation fitting: 128 tokens	25 / 64 per pair; 2048 per layer rotation fitting: 128 tokens
Optimization steps / learning rate	Rotation: 16 / 0.01; α, c: 4 Adam updates / 0.05	Rotation: 16 / 0.01; α, c: 4 Adam updates / 0.05
Initial absorption / loss weights	$\alpha_0 \in \{0, 0.25, 0.5, 0.75, 1\}$ inherited from same-layer RONA geometry; $(\lambda_q, \lambda_t, \lambda_+) = (4463.467478286071, 5250.25902909722, 1)$	All $\alpha_0 = 0$; $(\lambda_q, \lambda_t, \lambda_+) = (4463.467478286071, 5250.25902909722, 1)$

B.3 Alternating Calibration Procedure

Starting from the frozen reference encoder, RONA first records paired reference embeddings and layer inputs. It then fits each group rotation using Equation (1), estimates E_g and N_g at $c_{\mathrm{ref}} = 1$, and solves Equation (5) for every row-group. The resulting R_ℓ, W_ℓ^r, and ΔW_ℓ remain fixed.

The iterative stage initializes α_ℓ and $c_{\ell,g}$, constructs quantized reference inputs, and builds the initial grouped-GPTQ reconstructions. Each update computes the full paired loss, accumulates query- and target-side gradients at the same parameters, takes one Adam step, and projects α_ℓ and $c_{\ell,g}$ onto their prescribed intervals. It subsequently recollects the current encoder's quantized inputs for every target layer, forms all group Hessians, and rebuilds all GPTQ weights. The fourth post-update reconstruction supplies the final model. This order keeps the forward weights, activation distribution, and absorption derivative synchronized across calibration iterations.

B.4 Evaluation and Ablation Settings

Image and video scores are Hit@1; visual-document scores are linear-gain nDCG@5. For task set $\mathcal{T}_M$ in modality M, $S_M = |\mathcal{T}_M|^{-1} \sum_{t \in \mathcal{T}_M} S_t$, and $S_{\mathrm{All}} = 78^{-1} \sum_{t=1}^{78} S_t$. Thus the modality weights in All are $36/78$, $18/78$, and $24/78$. VDRv1, VDRv2, and VR denote ViDoRe, ViDoRe-V2, and VisRAG; OOD combines the two ViDoSeek and two MMLongBench-Doc tasks.

The standard evaluation uses up to 64 frames per video and 8-frame clips for moment retrieval, with nine negative clips in the moment-retrieval configurations. It runs with eight processes and default per-device batch size 8. Image tasks use their test splits; video and document tasks follow their benchmark-specific evaluation splits. Appendix C.3 lists task coverage and candidate-pool conventions.

Table B.3: Component changes in Table 2.

Variant	Modified component
w/o Noise Awareness	Select the safe basis using clean activation sensitivity, omitting the noise second moment ($\gamma = 0$).
w/o Group Alignment	Keep group width 128 and rank 32, but cyclically shift the safe-subspace grouping by 64 channels.
Fixed Absorption	Fix each layer's α to its original initialization in $\{0, 0.25, 0.5, 0.75, 1\}$.
Local Reconstruction	Replace encoder-level geometry calibration with layer-local MSE over all tokens and output dimensions; accumulate the losses and divide by $25 \times 2 \times 368$, with unexecuted layers contributing zero.

All four variants retain the W4A4 backbone and the benchmark aggregation in the main experiment. The table records the isolated design change for each ablation.

B.5 Export and Inference

Weights are stored as packed 4-bit codes with group scales. The evaluation runtime reconstructs these weights before the PyTorch Linear operation and applies quantize–dequantize activation inputs. Online operations comprise the group rotation, dynamic activation quantization, and linear projection in Equation (A.17). Absorption is incorporated into the exported weight codes; safe bases and the calibration correction are used only during preparation and optimization. Theoretical packed-kernel memory traffic and efficiency are analyzed in Appendix E.2.

B.6 Absorption Initialization and Interpretation of the Ablation

Fixed, adaptive, and disabled absorption. For a frozen group-local update ΔW_ℓ, the weight candidates before quantization are:

$$W_\ell^{\text{fixed}} = W_\ell^r + \alpha_\ell^{(0)} \Delta W_\ell, \qquad W_\ell^{\text{adaptive}} = W_\ell^r + \alpha_\ell \Delta W_\ell, \qquad W_\ell^{\text{disabled}} = W_\ell^r. \tag{B.5}$$

The *Fixed Absorption* row in Table 2 uses the first setting: each coefficient remains at its original initialization throughout calibration, retaining the underlying safe-subspace weight modification. Full RONA instead learns the continuous coefficients through the bilateral embedding-geometry objective. We additionally evaluate *w/o Absorption*, which fixes every $\alpha_\ell = 0$ from initialization through export. This removes the safe-subspace weight update while retaining the same rotations, 25 calibration pairs, token budget, and embedding-geometry objective. Activation clipping still starts at 1 and receives four Adam updates at learning rate 0.05; the current-student GPTQ Hessians and weight reconstructions are refreshed after each update. The three settings therefore compare disabled, fixed, and adaptive absorption under the same remaining calibration procedure. Their results are reported in Table B.5.

Initial coefficients used by the main run. The asymmetric W4A4 implementation reads a fixed, layer-indexed vector of initial coefficients and then rebuilds the rotations, safe subspaces, and ΔW_ℓ for its activation quantizer. Table B.4 reports the vector used by the 8B experiment. Its values are included with the released configuration, so reproducing the main run starts from this explicit vector rather than requiring another coefficient-search run.

Table B.4: Initial absorption coefficients for Qwen3-VL-Embedding-8B.

Initial coefficient	0	0.25	0.5	0.75	1	Total
Number of Linear layers	333	25	5	3	2	368

Thus 333 layers start with zero absorption, while 35 start with a nonzero coefficient. These are starting values; zero initialization still permits subsequent continuous updates, whereas disabled absorption fixes the coefficient at zero throughout optimization.

Preparatory initialization criterion. The preparatory source implementation defines a layer-local quantization-in-the-loop initializer with candidate set $\mathcal{A} = \{0, 0.25, 0.5, 0.75, 1\}$. For its calibration inputs, let Y_ℓ be the reference output and $\widehat{Y}_\ell(a)$ the output after activation quantization and grouped-GPTQ reconstruction of the candidate weight. The candidate score is

$$J_\ell^{\text{init}}(a) = \frac{\|\widehat{Y}_\ell(a) - Y_\ell\|_F^2}{\max(\|Y_\ell\|_F^2, 10^{-12})} + \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \left[1 - \cos(\widehat{y}_{\ell,i}(a), y_{\ell,i})\right]. \tag{B.6}$$

Here n_ℓ counts the calibration tokens used to score candidates. The routine evaluates the fixed grid, selects the smallest score, and breaks ties toward the smaller coefficient; zero is the fallback whenever no nonzero candidate strictly improves its score. This preparatory routine uses symmetric A4. The reported asymmetric W4A4 run imports the stored coefficients instead of re-executing that initializer with its newly fitted geometry. The initializer's local criterion and the main run's end-to-end objective therefore serve different stages.

Continuous refinement and calibration scope. In the main run, all $\alpha_\ell \in [0, 1]$ and $c_{\ell,g} \in [0.5, 1]$ are refined for four Adam updates with learning rate 0.05, using the fixed 25 training pairs and token budgets above. GPTQ reconstructions are refreshed after every update, and the fourth reconstruction is exported. Fixed Absorption freezes the initial coefficients; it is not an alternative initialization trial. The 2B configurations are recorded separately in Table B.2: Qwen3 inherits its recorded same-layer initial values, whereas WeMM starts with all $\alpha_\ell^{(0)} = 0$. The latter reaches 75.98 in Table 4, providing a reported zero-start run on that backbone. The evaluation follows these specified configurations; same-backbone initialization and calibration-resampling sensitivity are discussed in Appendix G.

Effect of the absorption update and its adaptation. Table B.5 extends the fixed-strength contrast in the main text with the fully disabled control. All scores use the same 78-task MMEB-V2 evaluation. The arrows show absolute score decreases relative to full RONA.

Table B.5: Absorption controls on Qwen3-VL-Embedding-8B under W4A4.

Setting	Image	Video	VisDoc	All
Full RONA	**74.77**	**63.79**	**79.06**	**73.55**
Fixed Absorption	74.26 ↓0.51	63.10 ↓0.69	78.64 ↓0.42	73.03 ↓0.52
w/o Absorption	73.98 ↓0.79	62.62 ↓1.17	78.22 ↓0.84	72.66 ↓0.89

With absorption disabled, All is 72.66. Retaining the fixed safe-subspace updates raises it to 73.03, a 0.37-point gain; learning their layer-wise strengths further raises it to 73.55, adding 0.52 points. Full RONA thus improves by 0.89 points over the matched rotation–clipping–GPTQ system without absorption. These two successive contrasts distinguish the benefit of the weight update from the additional benefit of coefficient adaptation under this protocol. Removing absorption lowers Image, Video, and VisDoc by 0.79, 1.17, and 0.84 points, respectively, with the largest absolute decrease on video.

B.7 Matched Protocol for the Five PTQ Baselines

Shared calibration and evaluation. The SmoothQuant, QuaRot, DuQuant, SpinQuant, and FlatQuant results in Table 1 use matched implementations on the same Qwen3-VL-Embedding-8B checkpoint as RONA. All six methods use the same calibration records, input preprocessing, token allocation, target Linear layers, W4/A4 definitions, and grouped-GPTQ backend. They also share the final clipping-calibration objective and update schedule. Table B.6 specifies these common settings; each method retains its own weight/activation reparameterization.

Transform fitting and tuning budget. Transformation fitting uses the same calibration inputs and a 128-token-per-Linear budget. Learned rotations or transformations use 16 Adam steps at learning rate 0.01 with seed 0; fixed transformations retain their prescribed form. RONA's group rotations use the signed-Hadamard initialization and Cayley parameterization in Table B.1. SmoothQuant retains channel scaling, QuaRot its invariant rotations, DuQuant its block rotations and permutations, SpinQuant its learned rotations, and FlatQuant its learned affine transformations. Thus matching the protocol preserves the transformation being evaluated rather than replacing every baseline with RONA's rotation.

The shared final calibration consists of four full-pair updates, with the same clipping initialization, bounds, objective, learning rate, and GPTQ-refresh schedule. RONA updates its absorption coefficients during these same four iterations. Its additional safe-direction construction, closed-form update, and recorded coefficient preparation are method-specific preprocessing, detailed in Sections 3.3–3.5 and Appendix B.6. Calibration examples and final checkpoints are fixed by the stated protocol, with no evaluation-set selection. This separates the common calibration budget from the algorithmic changes responsible for the comparison.

Table B.6: Shared 8B W4A4 protocol for the five PTQ baselines and RONA.

Controlled item	Common setting
Reference / layer scope	Same original BF16 checkpoint; 368 Linear layers (252 language, 116 vision). Embedding lookup, normalization, and biases retain their reference representations.
Calibration examples	Same fixed 25 training pairs, one from each calibration subset listed above; both query and target sides are used.
Token allocation	At most 64 tokens shared by both sides of each pair; 2,048-token cap per Linear and at most 1,600 from the 25-pair budget; identical token-selection rule.
Channel grouping	128 input channels per quantization group; weight scales are per output row and group, activation statistics per token and group.
Transform fitting	At most 128 tokens per Linear; learned transforms use 16 Adam steps at learning rate 0.01, seed 0. Fixed transforms retain their prescribed form.
W4 definition	Symmetric codes $[-8, 7]$, zero-point 0, and scale $\max(\|w\|_\infty, 10^{-8})/7$; common grouped-GPTQ reconstruction.
A4 definition	Asymmetric codes $[0, 15]$ with dynamic integer zero-point; zero-inclusive min–max range and quantize–dequantize rule in Equations (B.1)–(B.2).
GPTQ backend	Same independent-group CUDA implementation; sequential in-group compensation, damping ratio 0.01 and damping floor 10^{-5}.
Activation clipping	One $c_{\ell,g} \in [0.5, 1]$ per group, initialized at 1; four Adam updates at learning rate 0.05.
Final calibration target	Same paired frozen-reference embeddings and bilateral objective in Equation (9); fixed loss weights in Table B.1.
GPTQ refresh / export	Recollect current-student A4 inputs and rebuild group Hessians after every update; export the post-fourth-update reconstruction.
Evaluation / hardware	Same 78 tasks, preprocessing, candidate pools, and metrics; task-wise means. Eight A800 GPUs, eight evaluation processes, default batch size 8 per device.

C Protocols for Deep Analysis

The analyses in Section 4.5 examine complementary levels of the same retrieval process: full-encoder embedding geometry and controlled local perturbation directions. Both use the 78-task MMEB-V2 suite and positive/negative target identities selected by the full-precision encoder.

C.1 Bilateral Retrieval Geometry

Models and sample coverage. We compare normalized embeddings from Qwen3-VL-Embedding-8B in full precision, its QuaRot–GPTQ W4A4 baseline, and RONA W4A4. The baseline is labeled Vanilla W4A4 in Figure 3. The analysis covers 83,530 queries and 265,088 target entries deduplicated within each task. Query drift uses all queries; margin analysis uses 83,525 queries with valid positive and negative candidates. Five MVBench queries have no negative candidate and contribute only to drift. Corresponding queries are aligned by evaluation order and metadata; target embeddings are aligned by target identity.

Fixed targets and metrics. Let C_i be a query's candidate set, P_i its relevant set, and $N_i = C_i \setminus P_i$. For graded labels, relevance is strictly positive. Using unit full-precision embeddings, select

$$t_i^+ = \arg\max_{t \in P_i \cap C_i} (q_i^{\mathrm{FP}})^\top t^{\mathrm{FP}}, \qquad t_i^- = \arg\max_{t \in N_i} (q_i^{\mathrm{FP}})^\top t^{\mathrm{FP}}. \tag{C.1}$$

The same target identities are used for every method. For quantized method M, query and target drift are

$$d_{q,i}^M = 1 - (q_i^M)^\top q_i^{\mathrm{FP}}, \qquad d_{t,j}^M = 1 - (t_j^M)^\top t_j^{\mathrm{FP}}. \tag{C.2}$$

Each unique target contributes once within its task. Fixed-pair margins and their deviations are

$$m_i^M = (q_i^M)^\top t_{i,+}^M - (q_i^M)^\top t_{i,-}^M, \qquad D_i^M = |m_i^M - m_i^{\mathrm{FP}}|. \tag{C.3}$$

The mean of D_i^M uses all valid fixed pairs. For $\mathcal{I}_+ = \{i : m_i^{\mathrm{FP}} > 0\}$, the fixed-pair flip rate is

$$F^M = \frac{1}{|\mathcal{I}_+|} \sum_{i \in \mathcal{I}_+} \mathbf{1}[m_i^M < 0]. \tag{C.4}$$

Both comparisons are strict; zero margin contributes no flip. An additional full-pool statistic replaces the fixed pair by the highest-scoring relevant and irrelevant candidates under each method, using the same $\mathcal{I}_+$ for its failure-rate denominator.

Aggregation and uncertainty. Metrics are first averaged within each task, then equally across tasks in each modality; Overall equally weights all 78 tasks. Figure 3(a) separates query and target drift, while panel (b) separates modalities and Overall. Horizontal error bars are percentile 95% intervals from 5,000 paired task-bootstrap replicates with seed 20260913. Overall resampling is stratified by modality, preserving the 36/18/24 task counts. Methods share resampling indices. Normalization and candidate-search dot products use FP32 with TF32 disabled; selected-pair margins and aggregate statistics are recomputed in FP64.

C.2 Noise-Aware Direction and Intervention Protocol

Encoder and inputs. The local probe uses the original Qwen3-VL-Embedding-8B weights promoted from BF16 to FP32. It perturbs one MLP down-projection at a time in decoder blocks 6, 18, or 30 (zero-based indices 5, 17, 29), followed by the remaining decoder, final normalization, and last-valid-token pooling. The model has 36 decoder blocks, hidden width 4,096, MLP width 12,288, and 32 attention heads with 8 key/value heads. TF32 is disabled, attention uses scaled dot-product attention, and inference runs without gradients or key/value caching.

For each task, we sample up to 128 queries without replacement from the valid fixed-pair population, with seed $20260914 + k$ for zero-based task index k. Queries are then sorted by their original order. Target identities follow Equation (C.1). The selected population contains 9,703 queries and 24,575 input items, including queries and task-deduplicated positive/negative targets. All three layer positions use these same inputs. Clean reference embeddings are recomputed under the probe preprocessing: maximum sequence length 8,192, at most 8 uniformly sampled video frames, image pixel limit 524,288, and total video pixel limit 7,864,320.

Fixed subspaces. The down-projection input has 96 groups of 128 channels; each group retains 32 basis directions. Clean-only and RONA use the same fitted group rotations and the 25 training pairs described in Appendix B. With reference clipping fixed at 1, form X_g^r and $E_g = Q_A(X_g^r; 1) - X_g^r$. Their bases use

$$C_g(\gamma) = \frac{(X_g^r)^\top X_g^r}{n} + \gamma \frac{E_g^\top E_g}{n} + 10^{-5} I, \qquad \gamma = 0 \text{ (clean-only)}, \quad \gamma = 1 \text{ (RONA)}. \tag{C.5}$$

Second moments and eigendecompositions use FP64; bases are stored in FP32. Each layer supplies 3,072 directions per method, shared across tasks. The local update ΔW is solved once in the clean-only subspace using $\tau = 0.1$, $\mu = \nu = 10^{-3}$, and absorption gain 1. This same update is used in every intervention condition.

Direction sensitivity. For test activations, at most 64 evenly spaced tokens per input are used for direction statistics. Query and deduplicated target token energies are accumulated before taking square roots. For a unit basis direction $v_{g,j}$, Figure 4(a) uses

$$S_X(g,j) = \|X_g^r v_{g,j}\|_2, \qquad S_E(g,j) = \|E_g v_{g,j}\|_2. \tag{C.6}$$

The two axes are logarithmic and measure clean and quantization-noise response, respectively. Per-task/layer summaries average these sensitivities over groups and retained directions. Normalized directional-energy ratios are

$$f_X = \operatorname*{mean}_{g,j} \frac{\|X_g^r v_{g,j}\|_2^2}{\|X_g^r\|_F^2/128}, \qquad f_E = \operatorname*{mean}_{g,j} \frac{\|E_g v_{g,j}\|_2^2}{\|E_g\|_F^2/128}. \tag{C.7}$$

These ratios compare each retained direction with the average energy of a full-space direction in its group.

Equal-norm interventions. For each token, let x be its rotated input, $e = Q_A(x; 1) - x$, $\delta_o = (x + e)\Delta W^\top$, and $\rho = \|\delta_o\|_2$. The three perturbations are

$$\begin{aligned}
\delta_{\text{original}} &= \delta_o, \\
\delta_{\text{random}} &= \rho\, g/\|g\|_2, \qquad g \sim \mathcal{N}(0, I), \\
\delta_{\text{remove}} &= \rho\, \frac{\delta_o - e\Delta W^\top}{\|\delta_o - e\Delta W^\top\|_2} = \rho\, \frac{x\Delta W^\top}{\|x\Delta W^\top\|_2}.
\end{aligned} \tag{C.8}$$

All three have the same token-wise norm ρ. Norms are calculated in FP64, and the relative matching tolerance is 10^{-6}. The random control uses five seeds, 3101–3105. Each perturbation is added to the clean block output for all tokens; query, positive target, and negative target receive the same intervention type before the complete suffix computation. The surrounding encoder remains FP32, isolating the response to the local A4-induced weight-update perturbation.

Retrieval readout and statistics. Margins use the normalized probe embeddings and the same fixed target identities. Deviation is averaged over all selected queries, while flips use only strictly positive clean-probe margins as the denominator, as in Equations (C.3)–(C.4). Random-control deviations and binary flip indicators are averaged over the five seeds within each query before query aggregation. Each task/layer combination uses 5,000 paired query-bootstrap replicates for 95% intervals, with shared indices across conditions. Flip resampling uses the clean-correct query subset. For unit-level diagnostic reporting, a zero observed flip rate uses interval $[0, 1 - 0.025^{1/n_+}]$, where n_+ is the clean-correct count. Paired Original-minus-Remove differences use the same resampling indices. Unit-level one-sided paired sign-permutation tests use 5,000 draws, and Benjamini–Hochberg correction is applied separately to margin and flip comparisons across task/layer combinations.

Figure 4(b–c) aggregates the complete 78-task, three-layer evaluation with equal task/layer weights. The unit-level bootstrap statistics are combined replicate by replicate before constructing the plotted macro-mean intervals, as specified in Appendix C.4.

C.3 Task Inventory and Candidate Pools

Local candidate pools use the query-specific candidate set; global pools use all target identities in the task. All relevant targets are removed when constructing the negative set. The 36 image tasks use local pools; nine video tasks and all 24 document tasks use global pools. Table C.1 gives the evaluation-query and sampled-probe counts. Synthetic document task names are abbreviated by domain.

Table C.1: Task coverage, candidate pools, and query counts.

Task	Pool	Evaluation	Probe
Image: 36 tasks			
ImageNet-1K	local	1,000	128
N24News	local	1,000	128
HatefulMemes	local	1,000	128
VOC2007	local	1,000	128
SUN397	local	1,000	128
Place365	local	1,000	128
ImageNet-A	local	1,000	128
ImageNet-R	local	1,000	128
ObjectNet	local	1,000	128
Country211	local	1,000	128
OK-VQA	local	1,000	128
A-OKVQA	local	1,000	128
DocVQA	local	1,000	128
InfographicsVQA	local	1,000	128
ChartQA	local	1,000	128
Visual7W	local	1,000	128
ScienceQA	local	1,000	128

Task	Pool	Evaluation	Probe
VizWiz	local	1,000	128
GQA	local	1,000	128
TextVQA	local	1,000	128
MSCOCO_i2t	local	1,000	128
VisualNews_i2t	local	1,000	128
VisDial	local	1,000	128
MSCOCO_t2i	local	1,000	128
VisualNews_t2i	local	1,000	128
WebQA	local	1,000	128
EDIS	local	1,000	128
Wiki-SS-NQ	local	1,000	128
CIRR	local	1,000	128
NIGHTS	local	1,000	128
OVEN	local	1,000	128
FashionIQ	local	1,000	128
MSCOCO	local	1,000	128
RefCOCO	local	1,000	128
RefCOCO-Matching	local	1,000	128
Visual7W-Pointing	local	1,000	128
Video: 18 tasks			
SmthSmthV2	local	1,000	128
HMDB51	global	1,000	128
UCF101	global	1,000	128
K700	global	1,000	128
Breakfast	global	433	128
MSR-VTT	global	1,000	128
MSVD	global	670	128
DiDeMo	global	1,004	128
YouCook2	global	3,179	128
VATEX	global	4,478	128
QVHighlight	local	1,083	128
Charades-STA	local	727	128
MomentSeeker	local	1,602	128
Video-MME	local	2,700	128
NExTQA	local	8,564	128
EgoSchema	local	500	128
MVBench	local	4,000	128
ActivityNetQA	local	1,000	128
VisDoc: 24 tasks			
ViDoRe_arxivqa	global	500	128
ViDoRe_docvqa	global	451	128
ViDoRe_infovqa	global	494	128
ViDoRe_tabfquad	global	280	128
ViDoRe_tatdqa	global	1,646	128
ViDoRe_shiftproject	global	100	100
ViDoRe synthetic: AI	global	100	100
ViDoRe synthetic: energy	global	100	100
ViDoRe synthetic: government	global	100	100
ViDoRe synthetic: healthcare	global	100	100
ViDoRe-V2: ESG human labels	global	52	52
ViDoRe-V2: biomedical lectures	global	640	128
ViDoRe-V2: economics reports	global	232	128
ViDoRe-V2: ESG reports	global	228	128
VisRAG_ArxivQA	global	816	128
VisRAG_ChartQA	global	63	63
VisRAG_MP-DocVQA	global	591	128
VisRAG_SlideVQA	global	556	128
VisRAG_InfoVQA	global	718	128

Task	Pool	Evaluation	Probe
VisRAG_PlotQA	global	863	128
ViDoSeek-page	global	1,142	128
ViDoSeek-doc	global	1,142	128
MMLongBench-doc	global	838	128
MMLongBench-page	global	838	128

C.4 Full-Task Aggregation and Confidence Intervals for Figure 4

Analysis population. The aggregation uses all 78 tasks and all three specified probe layers. Let

$$\mathcal{U} = \{1, \dots, 78\} \times \{5, 17, 29\}, \qquad |\mathcal{U}| = 234. \tag{C.9}$$

Each unit $u = (t, \ell)$ is one task evaluated with a single perturbed layer. The 9,703 sampled task-indexed queries are reused at the three layer positions, yielding 29,109 query–layer evaluations. Query identities and fixed positive/negative targets are shared across conditions. The task/layer inventory is fixed, with every unit contributing equally to the summaries in panels (b–c); no diagnostic task/layer filter enters these averages. Table C.1 gives the query counts for individual tasks.

Direction scatter versus retrieval averages. Panel (a) visualizes the clean/noise sensitivity coordinates defined by Equation (C.6). Within a unit, the directional summary averages the responses over its 96 groups and 32 retained directions per group, after accumulating token energies and taking square roots. The number 234 describes the task/layer coverage, not the number of individual basis directions. Sensitivities retain the unnormalized response-norm convention of Appendix C.2; changing the axes to squared or token-normalized energy would define a different statistic. Panels (b–c), in contrast, display one macro-averaged retrieval statistic for each intervention condition.

Within-unit means and flip counts. Let n_t be the sampled query count for task t, and let $m^c_{i,u}$ denote its normalized-embedding margin under condition c. Write

$$d^c_{i,u} = |m^c_{i,u} - m^{\mathrm{FP}}_{i,u}|, \qquad \mathcal{I}^+_u = \{i : m^{\mathrm{FP}}_{i,u} > 0\}, \qquad n^+_u = |\mathcal{I}^+_u|. \tag{C.10}$$

For Original and Remove, $k^c_u = \sum_{i \in \mathcal{I}^+_u} \mathbf{1}[m^c_{i,u} < 0]$ is the integer count of fixed-pair reversals in that unit. The two unit-level statistics are

$$D^c_u = \frac{1}{n_t} \sum_{i=1}^{n_t} d^c_{i,u}, \qquad F^c_u = \frac{k^c_u}{n^+_u}. \tag{C.11}$$

The deviation uses all sampled valid pairs, whereas the flip denominator uses only positive clean-probe margins. A zero intervention margin contributes no reversal. These readouts concern the fixed positive/negative comparison after running the complete encoder suffix.

Five-seed random control. For Random, compute each query's absolute deviation and binary flip indicator separately for seeds $s = 1, \dots, 5$, then average:

$$d^{\mathrm{Random}}_{i,u} = \frac{1}{5} \sum_{s=1}^{5} |m^{\mathrm{Random}}_{i,u,s} - m^{\mathrm{FP}}_{i,u}|, \quad k^{\mathrm{Random}}_u = \frac{1}{5} \sum_{s=1}^{5} \sum_{i \in \mathcal{I}^+_u} \mathbf{1}[m^{\mathrm{Random}}_{i,u,s} < 0]. \tag{C.12}$$

The seed-averaged count k^{Random}_u can be fractional. The same query remains the sampling unit: seed runs are averaged before the unit statistics and before bootstrap resampling.

Equal-weight macro-aggregation. The plotted point estimates are

$$\overline{D}^c = \frac{1}{234} \sum_{u \in \mathcal{U}} D^c_u, \qquad \overline{F}^c = \frac{1}{234} \sum_{u \in \mathcal{U}} F^c_u. \tag{C.13}$$

This gives each task and each probe layer equal weight. Figure 4(b) reports 0.00570, 0.00558, and 0.00428 for Original, Random, and Remove; panel (c) reports $100\overline{F}^c$, namely 1.14%, 1.09%, and 0.56%. The flip percentages summarize the unit-level ratios k^c_u/n^+_u, rather than the pooled ratio $\sum_u k^c_u / \sum_u n^+_u$. Accordingly, the relevant numerators and denominators are the per-unit counts defined above; multiplying the displayed macro-rate by 29,109 would not recover a reversal count.

Bootstrap of the reported macro-mean. We use $B = 5{,}000$ paired query-bootstrap replicates. In replicate b, each unit resamples n_t queries with replacement for deviation and n_u^+ clean-correct queries with replacement for flips. Within the unit and metric, Original, Random, and Remove share the same sampled query indices. After computing $D_u^{c,(b)}$ and $F_u^{c,(b)}$, the corresponding replicate is averaged across the complete, fixed unit inventory:

$$\overline{D}^{c,(b)} = \frac{1}{234} \sum_{u \in \mathcal{U}} D_u^{c,(b)}, \qquad \overline{F}^{c,(b)} = \frac{1}{234} \sum_{u \in \mathcal{U}} F_u^{c,(b)}. \tag{C.14}$$

For either statistic T, the plotted interval is

$$\mathrm{CI}_{95\%}(\overline{T}^c) = \left[Q_{0.025}\left(\{\overline{T}^{c,(b)}\}_{b=1}^{B} \right), Q_{0.975}\left(\{\overline{T}^{c,(b)}\}_{b=1}^{B} \right) \right], \tag{C.15}$$

where Q_p is the empirical p-quantile. Thus we average replicate statistics first and take quantiles afterward, rather than averaging unit-level interval endpoints. Paired Original-minus-Remove differences are formed within each macro replicate. These intervals summarize within-unit query-resampling variation for the fixed task/layer inventory and the five evaluated random seeds. Bootstrap reuses the existing outputs and requires no additional model inference.

D Supplementary Task-Wise Results

Tables D.1–D.3 provide all 78 task scores for Qwen3-VL-Embedding-8B corresponding to Table 1. Values are reported on a 0–100 scale with two decimal places. SQ, QR, DQ, SP, and FQ denote SmoothQuant, QuaRot, DuQuant, SpinQuant, and FlatQuant, respectively. The full-precision encoder is a reference; the best W4A4 score in each task row is bold. Modality means are calculated from the unrounded task scores.

Table D.1: Image task results on Qwen3-VL-Embedding-8B.

Task	FP16	SQ	QR	DQ	SP	FQ	RONA
Classification							
VOC2007	93.30	79.10	84.20	80.90	83.30	84.30	**88.40**
N24News	80.50	62.80	63.60	65.80	66.80	68.10	**72.60**
SUN397	82.30	76.30	76.60	76.70	75.70	75.00	**81.00**
ObjectNet	80.40	70.80	73.20	69.90	73.60	73.00	**77.00**
Country211	27.30	19.70	20.60	21.60	16.10	17.30	**23.90**
Place365	48.10	43.10	42.40	42.20	41.70	42.00	**45.40**
ImageNet-1K	81.70	63.80	73.90	72.30	72.70	70.10	**76.00**
HatefulMemes	77.50	64.70	62.80	62.80	61.40	60.20	**69.90**
ImageNet-A	77.00	34.20	55.80	54.80	54.80	58.10	**59.90**
ImageNet-R	93.80	85.30	87.00	86.00	85.30	85.90	**90.40**
Image-level retrieval							
MSCOCO_I2T	79.00	69.20	68.00	70.90	70.30	68.30	**74.60**
MSCOCO_T2I	81.00	73.30	71.00	74.50	73.20	71.90	**78.40**
VisDial	87.60	73.80	74.80	76.30	77.00	76.50	**82.20**
CIRR	75.00	53.80	56.30	57.10	55.90	53.90	**67.90**
VisualNews_I2T	84.90	70.00	74.90	76.30	75.00	73.50	**79.10**
VisualNews_T2I	81.10	63.80	71.10	72.20	68.50	66.70	**74.70**
NIGHTS	72.90	64.90	65.60	64.90	65.70	64.40	**69.20**
WebQA	91.70	88.10	87.30	87.90	87.40	87.40	**90.70**
EDIS	96.40	86.60	90.00	90.70	89.80	88.50	**93.00**
OVEN	79.40	60.70	67.10	67.90	65.10	64.90	**70.90**
WIKI-SS-NQ	88.10	77.00	77.10	80.10	80.00	80.70	**82.60**
FashionIQ	44.20	31.50	34.10	36.20	33.00	31.70	**39.20**
Question answering							
OK-VQA	77.60	64.70	69.40	70.30	70.20	67.40	**73.30**
A-OKVQA	72.90	55.60	59.70	63.00	61.00	61.70	**64.60**
DocVQA	96.20	93.40	91.60	92.50	91.80	92.10	**95.30**
InfoVQA	87.90	74.50	77.30	80.20	77.30	78.40	**83.10**

Continued on next page

Table D.1 (continued)

Task	FP16	SQ	QR	DQ	SP	FQ	RONA
ChartQA	74.80	58.30	63.60	65.80	62.80	63.70	**68.00**
Visual7W	71.20	51.00	58.00	57.90	60.70	57.90	**65.00**
ScienceQA	81.80	59.90	65.30	68.70	65.90	66.40	**69.60**
GQA	92.50	82.60	**84.40**	83.40	83.60	81.10	83.00
TextVQA	92.70	87.00	86.40	87.00	86.50	86.00	**90.00**
VizWiz	64.70	53.60	53.60	53.90	54.50	55.60	**58.40**
Grounding							
MSCOCO	86.10	68.80	75.50	76.80	77.00	76.40	**83.00**
RefCOCO	95.70	82.90	89.40	89.80	87.50	90.20	**94.70**
RefCOCO-Matching	91.00	76.50	81.30	81.00	80.90	79.20	**85.60**
Visual7W-Pointing	95.80	71.90	81.10	84.60	82.10	83.20	**91.00**
Overall	80.11	66.48	69.83	70.64	69.84	69.49	74.77

Table D.2: Video task results on Qwen3-VL-Embedding-8B.

Task	FP16	SQ	QR	DQ	SP	FQ	RONA
Classification							
UCF101	95.50	84.00	88.00	89.30	86.70	86.00	**92.60**
HMDB51	87.00	68.60	78.70	78.80	74.30	74.90	**81.90**
Kinetics-700	67.60	53.90	57.60	57.90	58.70	57.10	**62.30**
Breakfast	63.74	50.81	54.51	**55.89**	48.73	51.27	55.43
Something-Something V2	81.90	66.70	68.70	71.00	70.50	70.10	**78.60**
Question answering							
Video-MME	62.59	53.37	55.96	56.89	55.44	55.30	**60.07**
MVBench	67.35	59.25	60.20	60.68	60.53	60.00	**63.48**
NExT-QA	76.59	67.46	71.16	69.79	69.61	69.08	**73.90**
EgoSchema	70.20	50.40	55.80	56.20	58.80	60.40	**63.60**
ActivityNetQA	80.90	74.90	76.20	77.00	77.00	76.40	**79.60**
Video-level retrieval							
MSR-VTT	58.40	50.60	53.20	53.20	51.20	52.40	**56.00**
MSVD	75.82	70.15	70.13	72.22	70.73	69.99	**74.33**
DiDeMo	66.04	49.50	55.47	55.57	56.16	55.67	**60.86**
VATEX	55.11	44.62	46.62	47.80	46.80	46.55	**51.85**
YouCook2	39.98	24.28	29.93	30.25	29.05	29.84	**33.31**
Moment retrieval							
QVHighlights	79.13	66.21	74.47	74.65	72.62	72.07	**78.21**
Charades-STA	32.05	17.33	25.34	24.65	24.24	24.65	**29.85**
MomentSeeker	56.05	48.88	48.56	49.75	49.37	49.00	**52.37**
Overall	67.55	55.61	59.48	60.09	58.92	58.93	63.79

Table D.3: Visual-document task results on Qwen3-VL-Embedding-8B.

Task	FP16	SQ	QR	DQ	SP	FQ	RONA
ViDoRe							
ViDoRe_arxivqa	86.90	79.50	80.14	82.03	81.08	79.57	**85.07**
ViDoRe_docvqa	54.10	44.27	46.64	47.33	47.46	47.49	**51.92**
ViDoRe_infovqa	91.07	84.85	86.14	85.71	85.71	85.27	**89.80**
ViDoRe_shiftproject	84.51	75.74	76.88	80.01	78.17	80.22	**83.95**
ViDoRe synthetic: AI	99.26	94.20	94.29	94.78	94.79	93.22	**99.26**
ViDoRe synthetic: energy	94.41	86.97	89.54	90.43	88.85	86.60	**93.43**
ViDoRe synthetic: government	98.02	90.68	93.18	94.15	93.47	90.04	**95.92**
ViDoRe synthetic: healthcare	97.79	94.05	93.85	95.26	94.21	94.49	**98.19**
ViDoRe_tabfquad	96.84	92.20	91.98	92.19	91.96	91.56	**95.84**
ViDoRe_tatdqa	69.84	57.03	62.09	61.45	60.33	60.90	**64.74**

Continued on next page

Table D.3 (continued)

Task	FP16	SQ	QR	DQ	SP	FQ	RONA
			ViDoRe-V2				
ViDoRe-V2: biomedical lectures	71.73	59.21	59.59	60.33	60.79	57.92	**64.90**
ViDoRe-V2: economics reports	66.93	56.72	54.73	57.80	57.28	56.95	**58.72**
ViDoRe-V2: ESG human labels	71.15	59.58	61.38	**66.12**	61.11	52.60	57.50
ViDoRe-V2: ESG reports	70.38	56.30	58.21	56.68	57.70	53.30	**61.47**
			VisRAG				
VisRAG_ArxivQA	88.24	77.90	82.11	81.67	80.75	80.07	**86.02**
VisRAG_ChartQA	88.34	83.94	79.97	83.85	83.06	82.43	**86.96**
VisRAG_InfoVQA	94.81	90.06	89.89	89.45	89.35	89.74	**93.44**
VisRAG_MP-DocVQA	89.30	77.74	82.32	84.26	82.02	81.34	**86.79**
VisRAG_PlotQA	74.83	62.79	65.85	62.04	63.00	65.03	**70.10**
VisRAG_SlideVQA	96.76	91.75	92.00	92.55	91.43	91.15	**95.35**
			ViDoSeek				
ViDoSeek-doc	86.07	83.50	82.11	81.85	82.02	82.01	**85.69**
ViDoSeek-page	88.46	80.35	81.05	**83.48**	78.29	78.53	81.87
			MMLongBench-Doc				
MMLongBench-doc	58.49	51.20	51.89	51.16	51.59	50.53	**55.79**
MMLongBench-page	60.46	49.15	51.76	53.18	51.05	48.15	**54.61**
Overall	82.45	74.15	75.32	76.16	75.23	74.13	79.06

The image, video, and visual-document partitions contain 36, 18, and 24 tasks, respectively. Within the visual-document partition, the OOD category in Table 1 combines ViDoSeek and MMLongBench-Doc with equal task weights. These task-wise results complement the aggregated comparison by retaining the individual benchmark outcomes.

E Additional Precision and Efficiency Experiments

E.1 Precision Comparisons

Figure 1(a) uses QuaRot to illustrate the change in retrieval quality across precision settings. Table E.1 reports the corresponding comparison for the five PTQ baselines and RONA. RONA reaches 77.50, 77.11, 75.58, and 75.31 overall at W8A16, W8A8, W4A16, and W4A8, respectively, while its W4A4 score is 73.55. All entries follow the task-mean aggregation defined in Appendix B.

Table E.1: Precision comparison on Qwen3-VL-Embedding-8B across different quantization settings.

Method	Precision	Image	Video	VisDoc	All
Reference	FP16	80.11	67.55	82.45	77.93
SmoothQuant	W8A16	79.30	67.48	82.28	77.49
SmoothQuant	W8A8	78.64	67.18	82.04	77.04
SmoothQuant	W4A16	76.65	64.66	80.41	75.04
SmoothQuant	W4A8	75.99	63.19	79.25	74.04
SmoothQuant	W4A4	66.48	55.61	74.15	66.33
QuaRot	W8A16	79.50	67.50	82.30	77.59
QuaRot	W8A8	79.00	67.30	82.10	77.25
QuaRot	W4A16	77.50	65.60	80.70	75.74
QuaRot	W4A8	77.00	64.60	79.70	74.97
QuaRot	W4A4	69.83	59.48	75.32	69.13
DuQuant	W8A16	79.55	67.50	82.32	77.62
DuQuant	W8A8	79.09	67.32	82.14	77.31
DuQuant	W4A16	77.71	65.75	80.91	75.93
DuQuant	W4A8	77.25	64.82	80.02	75.23
DuQuant	W4A4	70.64	60.09	76.16	69.90

Continued on next page

Table E.1 (continued)

Method	Precision	Image	Video	VisDoc	All
SpinQuant	W8A16	79.50	67.50	82.30	77.59
SpinQuant	W8A8	79.00	67.28	82.10	77.25
SpinQuant	W4A16	77.50	65.46	80.68	75.70
SpinQuant	W4A8	77.00	64.40	79.67	74.91
SpinQuant	W4A4	69.84	58.92	75.23	68.97
FlatQuant	W8A16	79.48	67.50	82.27	77.57
FlatQuant	W8A8	78.96	67.28	82.04	77.21
FlatQuant	W4A16	77.41	65.47	80.41	75.58
FlatQuant	W4A8	76.90	64.40	79.24	74.73
FlatQuant	W4A4	69.49	58.93	74.13	68.48
RONA	W8A16	79.62	67.21	82.05	77.50
RONA	W8A8	79.07	66.81	81.88	77.11
RONA	W4A16	77.10	65.63	80.75	75.58
RONA	W4A8	76.80	64.81	80.95	75.31
RONA	W4A4	74.77	63.79	79.06	73.55

E.2 Theoretical Memory and Efficiency Analysis

We account for the complete RONA inference path: independently learned dense group rotations, activation quantization, and the quantized Linear projection. Absorption is incorporated into the exported weights, so the safe bases, additive updates, and calibration gradients are discarded. The rotations remain online. The analysis below separates a prospective packed implementation from the current accuracy-evaluation runtime; all storage figures use decimal GB (10^9 bytes).

Persistent storage, including rotations. Let $P_q = \sum_\ell d_{\mathrm{out},\ell} d_{\mathrm{in},\ell}$ count quantized weights and P_u the unquantized parameters, including biases. If layer ℓ uses group widths $G_{\ell,g}$, independently storing its dense rotation blocks requires $P_{R,\ell} = \sum_g G_{\ell,g}^2$ scalar entries. Define $P_R = \sum_\ell P_{R,\ell}$, $N_s = \sum_\ell d_{\mathrm{out},\ell} \lceil d_{\mathrm{in},\ell}/G \rceil$ weight scales, and $N_c = \sum_\ell \lceil d_{\mathrm{in},\ell}/G \rceil$ clipping ratios. For a 16-bit reference and packed W4 weights,

$$\begin{aligned} M_{16} &= 2(P_q + P_u), \\ M_{\mathrm{packed}} &= \tfrac{1}{2} P_q + 2P_u + s_R P_R + s_s N_s + s_c N_c + M_{\mathrm{aux}}, \end{aligned} \tag{E.1}$$

where s_R, s_s, s_c are bytes per stored rotation entry, scale, and clipping ratio. M_{aux} covers alignment and any kernel-specific persistent metadata. For full groups, $P_{R,\ell} = d_{\mathrm{in},\ell} G$. These are stored matrix entries, rather than the intrinsic degrees of freedom of an orthogonal matrix. We budget each layer separately, with no cross-layer sharing or compression of learned rotations.

For the 8B language decoder, each of the 36 blocks has six projections with input width 4,096 and one down-projection with input width 12,288. Thus,

$$P_R^{\mathrm{lang}} = 36 \cdot 128\,(6 \cdot 4096 + 12288) = 169{,}869{,}312. \tag{E.2}$$

The visual branch has 27 blocks with three 1,152-input projections and one 4,304-input projection per block, plus eight 4,608-input merger projections. The 4,304-wide inputs contain 33 full groups and an 80-wide tail. Counting these actual widths adds 31,433,472 entries, giving Table E.2.

Table E.2: Dense-rotation storage for the 368 target Linear layers of the 8B encoder. Values are calculated from layer dimensions.

Branch	Linear	P_q (billions)	P_R (millions)	16-bit M_R (GB)	32-bit M_R (GB)
Language	252	6.945767	169.869312	0.339739	0.679477
Visual + mergers	116	0.571503	31.433472	0.062867	0.125734
Total	368	7.517270	201.302784	0.402606	0.805211

The current representation retains FP32 rotation blocks, weight scales, and clipping ratios. With $N_s = 58{,}740{,}336$ and $N_c = 12{,}303$, packed weights occupy 3.758635 GB, rotations 0.805211 GB, scales 0.234961 GB, and clipping ratios 0.000049 GB. Their sum is **4.798857 GB**, compared with 15.034540 GB for the same Linear weight matrices at 16 bits: a **68.1%** reduction in these persistent tensors, before adding $2P_u$ and M_{aux}. Storing rotations at 16 bits would reduce this subtotal to 4.396251 GB while retaining FP32 scales and clipping ratios; this is a separate storage scenario whose numerical behavior should be evaluated. The generic 75% W4 payload reduction therefore becomes a smaller method-specific saving once rotations and metadata are included.

Activation buffers and peak memory. For $A_q(t)$ live packed activation elements at time t, a deployment memory model is

$$M_{\text{peak}} = M_{\text{packed}} + \max_t \left\{ \tfrac{1}{2} A_q(t) + M_{\text{act,meta}}(t) + M_{\text{high}}(t) + M_{\text{work}}(t) \right\}. \tag{E.3}$$

Here M_{high} includes high-precision residuals, attention states, outputs, and any materialized rotated inputs; M_{work} includes kernel and framework workspaces. Only the packed A4 tensors receive the fourfold payload reduction. A separate rotation produces a high-precision intermediate before quantization, whereas a fused tiled implementation can keep that intermediate on chip. The current evaluator instead quantizes and dequantizes activations and caches reconstructed BF16 weights. If all target weight reconstructions are resident, their cache adds $2P_q = 15.034540$ GB to the packed-weight and rotation tensors. This explicitly distinguishes checkpoint storage, packed-device storage, and the resident tensors used for accuracy evaluation.

Dense-rotation arithmetic. For n_ℓ input tokens, the projection and rotation require, respectively,

$$\text{MAC}_{Q,\ell} = n_\ell d_{\text{in},\ell} d_{\text{out},\ell}, \qquad \text{MAC}_{R,\ell} = n_\ell \sum_g G_{\ell,g}^2. \tag{E.4}$$

For full groups their ratio is $G/d_{\text{out},\ell}$. At $G = 128$, it is 3.125% for 4,096-output projections, 12.5% for 1,024-output key/value projections, and 1.042% for 12,288-output MLP projections. Summed over the language-side projections with the same token count, rotations add 2.45% of their nominal MAC count. Their time cost depends on high-precision throughput, small-matrix utilization, and launch overhead, whereas the main GEMM may use faster low-bit arithmetic. The Cayley-fitted blocks are generally dense: the appropriate online cost is $O(n_\ell d_{\text{in},\ell} G)$ MACs, rather than the $O(n_\ell d_{\text{in},\ell} \log G)$ additions of a fixed fast Hadamard transform. Cayley inverses are computed during preparation, not during inference. Folding $W^r = WR$ offline still leaves XR before activation quantization.

Traffic and fusion assumptions. Let $W_\ell = d_{\text{in},\ell} d_{\text{out},\ell}$, $A_\ell = n_\ell d_{\text{in},\ell}$, and $O_\ell = n_\ell d_{\text{out},\ell}$ denote element counts. Consider an optimistic single-pass traffic model with 16-bit inputs and outputs, packed W4/A4 operands, and no accumulator spills. Let $B_{R,\ell}$ be rotation-coefficient traffic and $B_{\text{meta},\ell}$ quantization-metadata traffic. The bytes moved are

$$\begin{aligned} B_{16,\ell} &= 2W_\ell + 2A_\ell + 2O_\ell, \\ B_{\text{separate},\ell} &= \tfrac{1}{2}W_\ell + 7A_\ell + 2O_\ell + B_{R,\ell} + B_{\text{meta},\ell}, \\ B_{\text{prep-fused},\ell} &= \tfrac{1}{2}W_\ell + 3A_\ell + 2O_\ell + B_{R,\ell} + B_{\text{meta},\ell}, \\ B_{\text{fully-fused},\ell} &= \tfrac{1}{2}W_\ell + 2A_\ell + 2O_\ell + B_{R,\ell} + B_{\text{meta},\ell}. \end{aligned} \tag{E.5}$$

Separate rotation, quantization, and GEMM read/write $4A_\ell$, $2.5A_\ell$, and $0.5A_\ell$ activation bytes. Fusing rotation with quantization reduces their combined traffic; full fusion also avoids writing and rereading the packed intermediate. Reading each rotation once costs $s_R P_{R,\ell}$ bytes; cache reuse and tile scheduling determine its actual traffic. Fusion still performs the dense rotation and retains its coefficients. Groupwise asymmetric A4 additionally needs range reductions, zero-point handling, and rescaling; any precomputed weight sums belong in M_{aux}. Independent layer rotations are not assumed to merge through nonlinear or residual boundaries. These explicit costs replace the blanket assumption that all Linear-path traffic falls to one quarter.

End-to-end efficiency. Let f be the 16-bit runtime fraction spent in the target projections and s_Q their effective low-bit speedup excluding input transformations. Let η_R, η_A, and η_{sys} denote exposed

rotation, activation-preparation, and additional systems costs, normalized by the reference runtime and counted after overlap. Under an additive runtime model,

$$S = \frac{1}{(1-f) + f/s_Q + \eta_R + \eta_A + \eta_{\text{sys}}}, \qquad S > 1 \iff \eta_R + \eta_A + \eta_{\text{sys}} < f(1 - 1/s_Q). \quad \text{(E.6)}$$

For a bandwidth-limited stage, its attainable speedup is governed by the corresponding traffic budget; for a compute-limited stage, by effective arithmetic throughput. Long multimodal inputs change both token counts and the attention/projection balance. Consequently, RONA's deployment benefit is determined jointly by packed storage, dense-rotation execution, fusion, and workload shape. End-to-end latency and peak memory are left to a hardware implementation that measures these terms explicitly.

F Comparison with Existing Quantization Methods

This section distinguishes the shared ingredients of additive quantization from RONA's retrieval-specific design. OSAQ (Li et al., 2026b) identifies low-curvature directions of an approximate Hessian and suppresses weight outliers through a closed-form additive transformation. Its principal setting is weight-only LLM quantization; it also reports KV-cache quantization and W4A4 results by composing self-absorption with QuaRot, DuQuant, and SpinQuant. RONA directly calibrates joint weight–activation quantization of multimodal retrieval encoders, with a distinct direction criterion and retrieval-level objective. Table F.1 summarizes the differences in formulation and inference requirements. At the time of writing, OSAQ does not provide an official public implementation. We therefore do not include it as a numerical baseline in our matched W4A4 retrieval tables, since an independent reimplementation would introduce additional, unverified implementation choices; instead, we compare the published formulations here and leave a matched empirical comparison to a future release of the official implementation.

Table F.1: Formulation and deployment comparison between OSAQ and RONA.

Aspect	OSAQ	RONA
Evaluation setting	LLM quantization; W4A4 via composition	W4A4 multimodal retrieval
Direction criterion	Approximate Hessian low-curvature eigenspace	Groupwise clean/noise second moments in rotated coordinates
Additive solve	Softmax-weighted quadratic with ridge and anti-shift terms	Same quadratic family, solved within each quantization group
Adaptation	Closed-form weight adjustment	Learned layer absorption and group clipping with GPTQ refresh
Task alignment	Curvature-guided weight suppression	Query/target anchoring and positive-pair geometry
Online transforms	Absorption folds into weights; partner transforms remain	Absorption folds into weights; dense group rotations remain

Shared additive structure and the W4A4 distinction. Both methods express a weight correction in a low-sensitivity subspace and use a regularized, softmax-weighted quadratic solve. The additive parameterization and this solve are shared design elements. RONA changes the construction and use of the admissible directions: each basis is defined by Equation (2), which combines the uncentered clean-activation and actual activation-quantization-error moments within the same channel group. The matrix is a joint response-energy criterion, whereas OSAQ motivates its eigenspace through approximate task-loss curvature. This distinction becomes important when the input itself is quantized. For an update direction v, its response is $(X_g^r + E_g)v = X_g^r v + E_g v$; RONA explicitly penalizes both terms. The bound in Equation (3) links the selected joint energy to the quantized response, providing a criterion tailored to activation-noise coupling.

Retrieval-specific advantages. RONA further aligns rotations, safe directions, activation scales, and GPTQ with the same group boundaries, and learns how much of each precomputed update to apply through α_ℓ. Joint optimization of α_ℓ and $c_{\ell,g}$ uses both endpoints of a retrieval pair, while repeated GPTQ reconstruction exposes calibration to the updated weight grid. These choices address

the W4A4 retrieval problem at complementary levels: local noise sensitivity, quantizer alignment, and encoder-level similarity geometry. On the common 8B backbone, removing noise awareness, group alignment, or learned absorption reduces All by 1.30, 0.73, and 0.52 points, respectively; local reconstruction instead of embedding calibration loses 1.00 point (Table 2). These controlled comparisons support the individual design choices. The OSAQ comparison here concerns method structure; its own Hessian construction is distinct from RONA's clean-only ablation.

Relation to multiplicative PTQ and deployment. SmoothQuant (Xiao et al., 2023) rescales channels, while QuaRot (Ashkboos et al., 2024), SpinQuant (Liu et al., 2025), DuQuant (Lin et al., 2024), and FlatQuant (Sun et al., 2025) use rotational or affine changes of coordinates. RONA complements coordinate selection with constrained additive weight modification and bilateral retrieval calibration. The matched W4A4 comparisons yield overall gains of 3.65, 6.20, and 2.29 points over the strongest evaluated PTQ baseline on the three backbones. These are accuracy gains at a common precision, rather than an additional packing advantage over other W4A4 methods. Absorption adds no inference branch, but RONA's learned dense rotations have the storage and execution costs quantified in Appendix E.2. Its practical advantage is therefore improved retrieval accuracy under aggressive quantization, with deployment efficiency determined by the complete kernel implementation.

G LIMITATIONS

Hardware realization. This work establishes the algorithmic effectiveness of RONA for low-bit multimodal retrieval. The current evaluator reconstructs quantized weights before PyTorch Linear and applies activation quantize–dequantize, without a hardware-native fused W4A4 execution path. It also retains dense FP32 rotation blocks and may cache high-precision weight reconstructions. Accordingly, the persistent-storage calculations in Appendix E.2 describe specified packed deployment scenarios; measured end-to-end latency, throughput, and peak memory remain open implementation objectives. Efficient execution requires tiled dense rotations, groupwise asymmetric activation quantization, packed-weight projection, and appropriate metadata handling, together with profiling across token lengths, batch sizes, and modalities. These are infrastructure and hardware-adaptation tasks complementary to the calibration algorithm. We hope subsequent systems work will develop these kernels and complete RONA's accelerator-specific deployment.

Evaluation scope. Our experiments cover three encoder backbones and the 78-task MMEB-V2 suite, with a small paired calibration set. The safe directions are estimated from those calibration inputs, making robustness to shifts in domains and activation distributions an important direction for further study. Broader backbone coverage, repeated calibration samples, and matched W4A4 retrieval evaluations of additional additive methods such as OSAQ would further characterize the scope of the observed gains.

Initialization and calibration sensitivity. The reported results use the specified layer-wise starts and a fixed paired calibration set. The WeMM experiment includes a zero-initialized run, while the Qwen3 experiments use their recorded initial vectors. Appendix B.6 reports the fully disabled, fixed, and adaptive absorption controls under a matched remaining calibration procedure. Repeated same-backbone runs across alternative initializations and calibration draws remain directions for further evaluation.