COTQUANT: CROSS-MODAL ORTHOGONAL TRANSFORM FOR LARGE VISION-LANGUAGE MODEL QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training quantization (PTQ) is essential for efficient deployment of vision-language models (VLMs), yet low-bit quantization remains challenging due to heterogeneous multimodal activation outliers. We propose COTQuant, a PTQ framework with two complementary components. Cross-Modal Transform (CMT) jointly redistributes text and vision activation outliers through hierarchical orthogonal transformations, while Quantization-Aware Group Search (QAS) reorganizes weight channels to minimize group-wise quantization error. Both transformations leave the model output unchanged and introduce only a small amount of additional inference-time overhead. Across three VLMs and five multimodal benchmarks, COTQuant consistently outperforms existing PTQ methods while retaining 98.03%–98.73% of the corresponding BF16 performance under W4A4 quantization. Notably, we demonstrate effective 4-bit PTQ on Qwen3.5-122B, which is the first 4-bit PTQ attempt on over 100B-scale VLM.

1 INTRODUCTION

Vision-language models (VLMs) (Alayrac et al., 2022; Liu et al., 2023; Bai et al., 2023) have demonstrated strong capabilities in multimodal understanding and reasoning, but their rapidly growing model sizes impose substantial memory and computational costs(Dong et al., 2024; Heo et al., 2024; Wan et al., 2024). Post-training quantization (PTQ) provides an efficient solution by reducing numerical precision without expensive retraining(Han et al., 2016; Liu et al., 2024a; Xiao et al., 2023). In particular, low-bit quantization can substantially reduce inference cost, yet remains challenging because activation outliers enlarge the quantization range and thus greatly increase the activation quantization loss(Wei et al., 2023; Bondarenko et al., 2023).

Existing PTQ methods have extensively studied activation outliers in large language models (LLMs). SmoothQuant(Xiao et al., 2023) mitigates outliers through equivalent channel-wise scaling, while OmniQuant(Shao et al., 2023) learns transformation and clipping parameters from calibration data. More recent methods such as QuaRot and DuQuant(Ashkboos et al., 2024; Lin et al., 2024a), employ orthogonal transformations to redistribute activation energy while preserving the original computation. However, directly applying these methods to VLMs overlooks an important multimodal characteristic: text and vision activations exhibit heterogeneous(Li et al., 2024) yet partially overlapping outlier channels, whose locations and magnitudes can differ substantially. Meanwhile, group-wise quantization is widely adopted in low-bit weight quantization (Frantar et al., 2022; Lin et al., 2024b; Ashkboos et al., 2024). However, we observe that its quantization error also depends on how weight channels are assigned to groups, providing an additional opportunity for weight-side optimization.

Motivated by these observations, we propose **COTQuant**, a W4A4 PTQ framework that jointly improves activation and weight quantization through two complementary components. **Cross-Modal Transform**

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.

(CMT) explicitly exploits the cross-modal structure of activation outliers and hierarchically redistributes their energy through orthogonal transformations. **Quantization-Aware Group Search (QAS)** further reorganizes transformed weight channels by directly minimizing the actual group-wise quantization error. The resulting channel permutation is fused into the CMT transformation, avoiding additional standalone reordering during inference.

We evaluate COTQuant on three VLMs ranging from 7B to 122B parameters across five multimodal benchmarks. Extensive experiments show that COTQuant consistently outperforms existing PTQ methods under W4A4 quantization while preserving near-full-precision performance, and remains effective when scaling to the 122B-parameter model.

Our main contributions are summarized as follows:

- We propose **Cross-Modal Transform (CMT)**, which exploits the heterogeneous yet partially overlapping outlier structures of text and vision activations to improve low-bit activation quantization.
- We introduce **Quantization-Aware Group Search (QAS)**, which directly optimizes channel grouping according to the actual group-wise weight quantization error and can be fused into CMT without additional inference-time reordering.
- We conduct extensive evaluations across multiple VLMs and five multimodal benchmarks, including Qwen3.5-122B, demonstrating the effectiveness and scalability of COTQuant.

2 PRELIMINARIES

2.1 LINEAR LAYER QUANTIZATION

Consider a linear layer in a vision-language model (VLM),

$$Y = XW, \tag{1}$$

where $X \in \mathbb{R}^{T \times C_{\text{in}}}$ denotes the input activations and $W \in \mathbb{R}^{C_{\text{in}} \times C_{\text{out}}}$ denotes the weight matrix. We focus on uniform W4A4 post-training quantization, with per-token quantization for activations and channel-wise or group-wise quantization for weights(Bondarenko et al., 2023; Nagel et al., 2020; Yao et al., 2022).

For a tensor x, uniform quantization is defined as

$$Q(x) = \text{clamp}\left(\left\lfloor \frac{x}{\Delta} + z \right\rceil, 0, 2^b - 1\right), \tag{2}$$

where Δ and z denote the quantization step size and zero point, respectively. Large activation outliers enlarge the quantization range and consequently increase Δ, reducing the effective precision for the majority of activation values. Similarly, the error of group-wise weight quantization depends on the dynamic range within each group.

2.2 ORTHOGONAL TRANSFORMATION EQUIVALENCE

Let $F \in \mathbb{R}^{C_{\text{in}} \times C_{\text{in}}}$ be an orthogonal matrix satisfying $F^\top F = I$. We transform the activations and weights as

$$\widetilde{X} = XF, \qquad \widetilde{W} = F^\top W. \tag{3}$$

The full-precision computation remains unchanged:

$$\widetilde{X}\widetilde{W} = XFF^\top W = XW. \tag{4}$$

Therefore, an orthogonal transformation can reshape the activation distribution without altering the original model output. This property enables us to redistribute activation outliers into a more quantization-friendly representation while preserving computational equivalence(Tseng et al., 2024; Ashkboos et al., 2024).

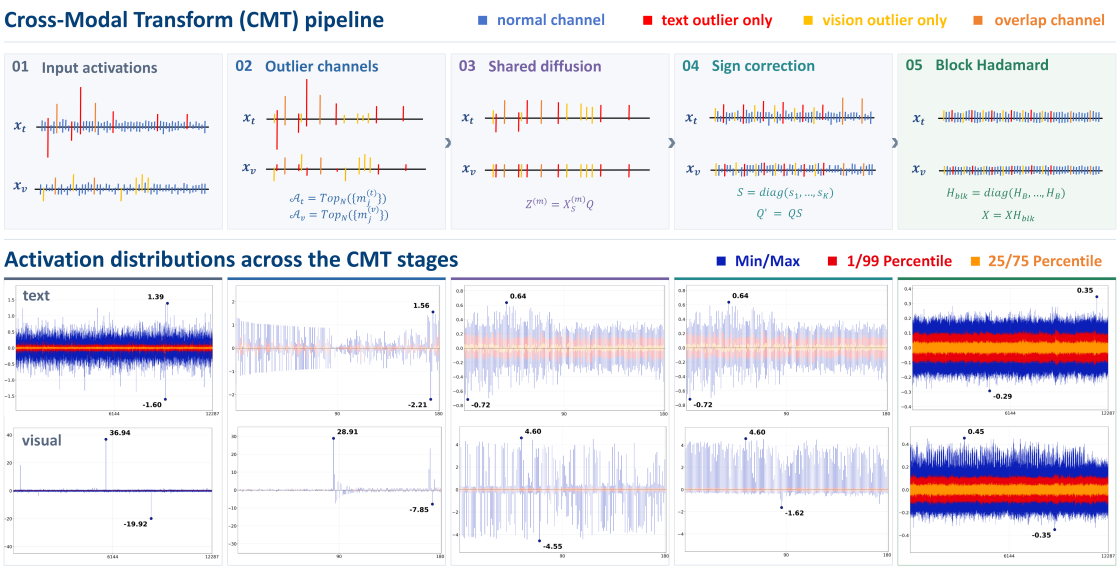


Figure 1: Overview of the proposed Cross-Modal Transform (CMT). The top row illustrates the CMT pipeline. The bottom row visualizes the corresponding activation distributions at different stages, showing how CMT progressively suppresses cross-modal outliers and redistributes activation energy for low-bit quantization.

3 METHODOLOGY

3.1 OVERVIEW

COTQuant jointly improves activation and weight quantization through two complementary components. **Cross-Modal Transform (CMT)** exploits the outlier structures of text and vision activations, combining shared orthogonal transform with block-wise Hadamard transformation to redistribute activation energy. **Quantization-Aware Group Search (QAS)** then reorganizes the transformed weight channels by minimizing the actual group-wise quantization error. Its resulting permutation is fused into the CMT transformation, introducing no additional standalone channel reordering during inference. Both components preserve the original full-precision computation, while producing more quantization-friendly activation and weight distributions.

3.2 CROSS-MODAL TRANSFORM

This module targets the severe activation outliers that affect low-bit activation quantization precision. Given text and vision calibration activations, CMT first schedules cross-modal outlier channels, then applies shared orthogonal transform followed by block-wise Hadamard transformation.

1. Modality-Aware Channel Scheduling

Given calibration activations $\mathcal{X}_t \in \mathbb{R}^{N_t \times C}$ and $\mathcal{X}_v \in \mathbb{R}^{N_v \times C}$ collected from the text and vision branches of the target layer, and a block size B with $C \bmod B = 0$, we first identify outlier channels per modality. Let

$N = C/B$ be the number of blocks. The per-channel magnitude of each modality is defined as

$$m_j^{(t)} = \max_i |\mathcal{X}_t[i,j]|, \qquad m_j^{(v)} = \max_i |\mathcal{X}_v[i,j]|, \tag{5}$$

and the top-N channels form the anomaly sets(Zhao et al., 2024)

$$\mathcal{A}_t = \mathrm{Top}_N\big(\{m_j^{(t)}\}\big), \qquad \mathcal{A}_v = \mathrm{Top}_N\big(\{m_j^{(v)}\}\big). \tag{6}$$

The two sets are then partitioned into three groups:

$$\mathcal{A}_{t\setminus v} = \mathcal{A}_t \setminus \mathcal{A}_v, \qquad \mathcal{A}_{v\setminus t} = \mathcal{A}_v \setminus \mathcal{A}_t, \qquad \mathcal{O} = \mathcal{A}_t \cap \mathcal{A}_v, \tag{7}$$

where $\setminus$ denotes the set difference operator, and O contains the outlier channels shared by both modalities and $M = |O|$. The number of distinct cross-modal outlier channels is

$$K = |A_t \cup A_v| = 2N - M \le \frac{2C}{B}. \tag{8}$$

We then construct a permutation Π_a that distributes modality-specific outliers across different blocks while placing a text only and visual only outlier or a overlap outlier only once. This avoids concentrating modality-specific outliers within the same block and collects all distinct cross-modal outliers into a compact K-dimensional subspace for subsequent orthogonal transform.

2. Shared Orthogonal Transform

After scheduling, a block-wise transform alone can only redistribute energy within individual blocks and cannot directly correct the imbalance among different blocks(Sanjeet et al., 2026). We therefore first learn a global orthogonal transformation over the compact outlier subspace. Let

$$Q \in \mathbb{R}^{K\times K}, \qquad Q^\top Q = I, \tag{9}$$

denote the shared orthogonal matrix. Importantly, the same Q is optimized jointly using text and vision calibration activations rather than learning two independent modality-specific transformations.

Let $X_{\mathcal{S}}^{(m)}$ denote the scheduled outlier channels for modality $m \in \{t, v\}$. The diffused activation is

$$Z^{(m)} = X_{\mathcal{S}}^{(m)} Q \tag{10}$$

Following the scheduled channel structure, we partition the output into text-specific, vision-specific, and overlapping components:

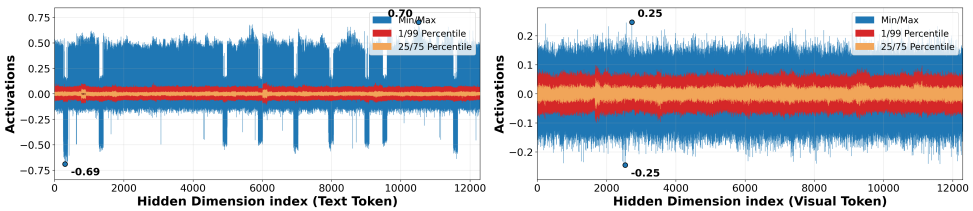
$$Z^{(m)} = \left[Z_t^{(m)}, Z_v^{(m)}, Z_o^{(m)}\right] \tag{11}$$

Since the subsequent block-wise Hadamard transform spreads each special channel over a block of size B, we construct a block-aware peak estimate,

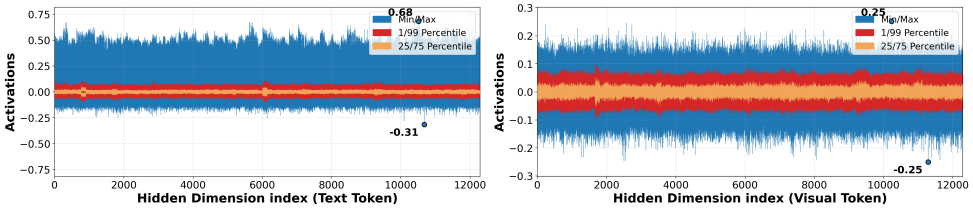
$$P^{(m)} = \left[\frac{|Z_t^{(m)}| + |Z_v^{(m)}|}{\sqrt{B}}, \ \frac{|Z_o^{(m)}|}{\sqrt{B}}\right] \tag{12}$$

To directly suppress extreme values, we use LogSumExp as a differentiable approximation to the maximum:

$$\mathrm{SoftPeak}(p;\beta) = \frac{1}{\beta}\log\sum_k \exp(\beta p_k) \tag{13}$$



(a) Before sign correction



(b) After sign correction

Figure 2: Activation distribution before and after sign correction. By flipping extreme outliers to the same side of zero, sign correction reduces the full dynamic range of text activations from 1.39 to 0.99, directly decreasing the quantization step size under asymmetric uniform quantization.

The shared rotation is optimized jointly over both modalities:

$$\mathcal{L}_{\text{rot}} = \frac{1}{N_t}\sum_{i=1}^{N_t}\text{SoftPeak}\left(P_i^{(t)};\beta\right) + \frac{1}{N_v}\sum_{i=1}^{N_v}\text{SoftPeak}\left(P_i^{(v)};\beta\right) \tag{14}$$

The orthogonality of Q is maintained via orthogonal parameterization, with QR decomposition applied after each parameter update. In practice, the transformation is initialized from the identity matrix. During optimization, we adopt an annealing mechanism where β is gradually increased from 5 to 50, allowing the smooth objective to progressively approach the true peak value.

3. Sign Correction

Although the learned orthogonal transform reduces activation peaks, asymmetric quantization is determined by the full dynamic range

$$R(x) = \max(x) - \min(x), \tag{15}$$

rather than by the maximum absolute value alone. After orthogonal transform, extreme positive and negative outliers may still coexist on opposite sides of zero and unnecessarily enlarge this range. We therefore perform a single sign correction immediately after the shared orthogonal transform. For every output channel j of Q, we jointly collect the extrema of the two modalities:

$$P_j = \max\left(\max_i Z_{i,j}^{(t)}, \max_i Z_{i,j}^{(v)}\right), \tag{16}$$

$$N_j = \max\left(-\min_i Z_{i,j}^{(t)}, -\min_i Z_{i,j}^{(v)}\right) \tag{17}$$

The sign of channel j is selected as

$$s_j = \begin{cases} -1, & N_j > P_j, \\ +1, & \text{otherwise.} \end{cases} \tag{18}$$

$$S = \mathrm{diag}(s_1, \ldots, s_K), \tag{19}$$

we directly absorb the sign correction into the learned rotation:

$$Q' = QS \tag{20}$$

Since S is itself orthogonal,

$$Q'^{\top} Q' = S^{\top} Q^{\top} Q S = I \tag{21}$$

Therefore, sign correction introduces no magnitude scaling and does not require an additional transformation at inference time. It simply refines the orientation of the transformed outliers so that the dominant outliers are better aligned for asymmetric quantization. As shown in the bottom row of Figure 2, sign correction further compresses the range of the transformed activations.

4. Equivalent Transform Pipeline

After the shared orthogonal transform, the remaining activation energy is further redistributed within each block using a normalized Hadamard matrix

$$H_B = \frac{1}{\sqrt{B}} \mathcal{H}_B \tag{22}$$

$$H_{\mathrm{blk}} = \mathrm{diag}(H_B, \ldots, H_B) \tag{23}$$

Let Π_a denote the channel permutation and let $\overline{Q'} \in \mathbb{R}^{C \times C}$ denote the identity matrix whose outlier-channel subspace is replaced by Q'. The complete transform can then be written as

$$F_{\mathrm{rot}} = \Pi_a \overline{Q'} H_{\mathrm{blk}} \Pi_a^{\top} \tag{24}$$

All factors are orthogonal, and therefore

$$F_{\mathrm{rot}}^{\top} F_{\mathrm{rot}} = I \tag{25}$$

For an original linear layer $Y = XW$, we transform the activation and weight as

$$\widetilde{X} = X F_{\mathrm{rot}}, \qquad \widetilde{W} = F_{\mathrm{rot}}^{\top} W \tag{26}$$

$$\widetilde{X}\widetilde{W} = X F_{\mathrm{rot}} F_{\mathrm{rot}}^{\top} W = XW \tag{27}$$

Therefore, the **Cross-Modal Transform** can reshape the multimodal activation distribution into a more quantization-friendly state without changing the output of the linear layer.

3.3 Quantization-Aware Group Search

For the CMT-transformed weight $\widetilde{W} \in \mathbb{R}^{C_{\mathrm{in}} \times C_{\mathrm{out}}}$, QAS searches for a channel permutation that minimizes the actual group-wise quantization error. Given a group size G, the input channels are partitioned into C/G groups, and each group independently determines its quantization scale and zero point. Let $\Pi_q \in \mathbb{R}^{C \times C}$ denote a channel permutation matrix. QAS searches for

$$\Pi_q^* = \arg\min_{\Pi_q} \mathcal{L}_w(\Pi_q) \tag{28}$$

$$\mathcal{L}_w(\Pi_q) = \frac{1}{|\widetilde{W}|} \left\| Q_G\left(\Pi_q^{\top} \widetilde{W}\right) - \Pi_q^{\top} \widetilde{W} \right\|_F^2 \tag{29}$$

Here, $Q_G(\cdot)$ denotes the group-wise quantizer.

Pairwise Quantization-Aware Regrouping

Starting from the original grouping, QAS iteratively refines the channel assignment via pairwise group search. In each iteration, groups are first paired by alternating between two orders: a random global order for exploration, and an energy-based order that sorts groups by average channel energy and pairs neighbors with similar magnitudes, thereby covering both global exchanges and local regrouping.

Given groups A and B, QAS generates candidate exchanges based on the current quantization errors: it enumerates cross-group exchanges of the highest-error channels from both groups, randomly samples the remaining candidates for diversity, and constructs a balanced repartition over $A \cup B$. For each channel j, its preference between the two groups is measured by

$$d_j = C(j \to A) - C(j \to B), \tag{30}$$

$$C(j \to A) = \sum_r \ell(Q_{r,A}(W_{r,j}), W_{r,j}), \tag{31}$$

where $Q_{r,A}$ denotes the quantizer determined by the current weights of group A at output row r, and $\ell(\cdot)$ is the weight quantization error. Channels are sorted by d_j, with the first G assigned to A and the remaining G to B. A candidate regrouping is accepted only if it reduces the actual loss:

$$L_w(\Pi_q^{new}) < L_w(\Pi_q^{old}) \tag{32}$$

The accepted grouping is then used in the next iteration, and the process continues until no further improvement is obtained.

Permutation Fusion

The QAS permutation does not introduce an additional standalone channel-reordering operation. The preceding hierarchical rotation already applies an overlap-aware permutation Π_a and restores the channel order using $\Pi_a^\top$ after the block-wise Hadamard transform. Therefore, the QAS permutation Π_q^* can be directly fused with this inverse permutation:

$$\Pi_f = \Pi_a^\top \Pi_q^* \tag{33}$$

The complete transformation is thus

$$F = \Pi_a \overline{Q'} H_{\text{blk}} \Pi_f = \Pi_a \overline{Q'} H_{\text{blk}} \Pi_a^\top \Pi_q^* \tag{34}$$

$$F^{-1} = F^T \tag{35}$$

Therefore, QAS reduces the group-wise quantization loss of the model without incurring any additional inference-time computation.

4 Experiments

We evaluate COTQuant on Qwen2-VL-7B(Wang et al., 2024b), Qwen3-VL-8B(Bai et al., 2025), and Qwen3.5-122B(Qwen Team, 2026) across five vision-language benchmarks under low-bit weight-activation quantization. We compare against representative scaling- and rotation-based PTQ methods. Detailed experimental settings are provided in Appendix A.

4.1 Main Results

Table 1 compares COTQuant with representative PTQ methods on vision-language benchmarks. COTQuant consistently preserves near-full-precision performance under W4A4 quantization. Compared with the

Table 1: Performance comparison on Qwen2VL-7B, Qwen3VL-8B, and Qwen3.5-122B across five vision-language benchmarks. G128 refers to group-wise quantization with group size 128. The best result is marked in **bold**.

Method	bits	OCRBench	TextVQA	VizWiz	MME	ScienceQA	Average$^{\uparrow}$
Qwen2-VL-7B	BF16	79.4	81.41	68.76	87.74	84.98	80.46
SmoothQuant	W4A8	64.5	74.02	57.97	85.30	80.47	72.45↓ 8.01
OmniQuant	W4A4	39.7	48.22	34.48	70.60	55.33	49.67↓ 30.79
DuQuant	W4A4	77.0	79.42	66.21	86.52	**83.54**	78.54↓ 1.92
QuaRot	W4A4	77.2	79.95	67.10	85.68	83.24	78.63↓ 1.83
COTQuant	W4A4	76.6	80.51	67.71	86.90	82.70	78.88↓ 1.58
COTQuant(G128)	W4A4	**78.6**	**80.77**	**67.91**	**86.94**	82.99	**79.44**↓ 1.02
Qwen3-VL-8B	BF16	86.5	81.58	69.55	88.80	94.35	84.16
SmoothQuant	W4A8	82.1	74.92	67.65	79.06	82.00	77.14↓ 7.06
OmniQuant	W4A8	82.4	75.72	64.72	81.17	86.91	78.18↓ 5.98
DuQuant	W4A4	84.4	80.35	69.55	86.06	**92.07**	82.49↓ 1.67
QuaRot	W4A4	79.6	73.60	62.01	79.86	82.45	75.50↓ 8.66
COTQuant	W4A4	85.5	**80.49**	**69.85**	86.06	91.99	82.78↓ 1.38
COTQuant(G128)	W4A4	**85.7**	80.45	69.62	**87.03**	91.99	**82.96**↓ 1.20
Qwen3.5-122B	BF16	87.8	83.38	70.85	92.46	97.15	86.33
SmoothQuant	W4A4	64.8	55.58	51.40	63.56	60.03	59.07↓ 27.26
OmniQuant	W4A4	67.2	59.30	54.55	69.25	65.99	63.26↓ 23.07
DuQuant	W4A4	83.6	77.02	60.84	88.12	90.52	80.02↓ 6.31
COTQuant	W4A4	85.1	**82.01**	68.23	**91.15**	95.26	84.35↓ 1.98
COTQuant(G128)	W4A4	**85.8**	81.93	**68.70**	90.94	**95.79**	**84.63**↓ 1.70

strongest W4A4 baselines, COTQuant improves the average score by 0.47–4.61 points, while retaining 98.03%–98.73% of the corresponding BF16 performance. Notably, the advantage becomes more pronounced as the model scales up. On Qwen3.5-122B, COTQuant achieves an average score of 84.63, only 1.70 points below BF16, whereas DuQuant suffers a 6.31-point degradation. This result demonstrates that the proposed cross-modal transformation remains effective when scaling to substantially larger VLMs. Moreover, QAS consistently improves CMT on all three models, confirming that quantization-aware weight regrouping provides complementary gains beyond activation-side transformation.

4.2 Ablation Study

Ablation of CMT. We first study the contribution of each component in CMT. As shown in Table 2, using block-wise Hadamard transformation alone achieves an average score of 77.96. Introducing modality-aware channel permutation improves the average score to 78.15, indicating that explicitly reorganizing cross-modal outlier channels facilitates more effective local energy redistribution. Adding a random orthogonal transformation further improves the score to 78.37, while replacing it with the learned shared orthogonal transform

Table 2: Ablation study of the proposed Cross-Modal Transform (CMT) on Qwen2-VL-7B under W4A4 quantization. The performance degradation relative to BF16 is reported after ↓.

	Component	OCRBench	TextVQA	VizWiz	MME	ScienceQA	Avg.↑
	BF16	79.4	81.41	68.76	87.74	84.98	80.46
A	Block Hadamard	75.7	79.80	66.28	86.01	82.00	77.96↓ 2.50
B	A + Modality-aware Permutation	75.6	79.81	67.68	86.06	81.60	78.15↓ 2.31
C	B + Random Orthogonal Matrix	75.9	79.81	**67.81**	85.93	82.40	78.37↓ 2.09
D	B + Learned Shared Transform	76.2	80.39	**67.81**	86.10	82.40	78.58↓ 1.88
E	Full CMT	**76.6**	**80.51**	67.71	**86.90**	**82.70**	**78.88**↓ 1.58

Table 3: Ablation study of Quantization-Aware Group Search (QAS) on Qwen2-VL-7B. G128 denotes conventional group-wise weight quantization with a group size of 128, while QAS128 applies the proposed quantization-aware channel grouping with the same group size. The performance degradation relative to BF16 is reported after ↓.

Method	OCRBench	TextVQA	VizWiz	MME	ScienceQA	Avg.↑
BF16	79.4	81.41	68.76	87.74	84.98	80.46
CMT	76.6	80.51	67.71	86.90	82.70	78.88↓ 1.58
CMT + G128	76.8	80.57	67.77	86.86	82.85	78.97↓ 1.49
CMT + QAS128	78.6	**80.77**	**67.91**	**86.94**	82.99	**79.44**↓ 1.02
DuQuant	77.0	79.42	66.21	86.52	83.54	78.54↓ 1.92
DuQuant + G128	77.5	79.48	66.47	86.65	83.49	78.72↓ 1.74
DuQuant + QAS128	**79.1**	79.73	67.10	86.77	**83.64**	79.23↓ 1.23

increases the average performance to 78.58, demonstrating the benefit of optimizing the cross-modal outlier subspace rather than relying on a random rotation. Finally, the full CMT achieves an average score of 78.88, reducing the gap to the BF16 model to 1.58 points. Overall, the above experiments validate the contribution of each component.

Effectiveness of QAS. We evaluate QAS on top of both CMT and DuQuant to verify its effectiveness over conventional group-wise quantization. As shown in Table 3, replacing G128 with QAS128 improves the average score from 78.97 to 79.44 on CMT, yielding a 0.47-point gain. A consistent improvement is observed on DuQuant, where QAS128 increases the average score from 78.72 to 79.23, outperforming conventional G128 by 0.51 points. These results demonstrate that QAS provides additional optimization beyond standard group-wise quantization by explicitly searching for more quantization-friendly channel assignments.

5 Conclusion

In this work, we propose COTQuant, a post-training quantization framework for large vision-language models. COTQuant combines Cross-Modal Transform (CMT) to alleviate heterogeneous multimodal activation outliers with Quantization-Aware Group Search (QAS) to reduce group-wise weight quantization error. Extensive experiments across multiple VLMs and five multimodal benchmarks demonstrate that COTQuant consistently preserves strong full-precision performance under W4A4 quantization. In particular, it scales effectively to Qwen3.5-122B, achieving an average score of 84.63 with only a 1.70-point degradation from BF16, highlighting the effectiveness and scalability of our approach.

AI USE STATEMENT

In this work, we used DeepSeek for language translation and ChatGPT for literature retrieval and polishing. All AI-assisted content was manually reviewed by the authors. We take full responsibility for the final content of this work, including all text, claims, experiments, and artifacts produced with the aid of generative AI tools.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736. Curran Associates, Inc., 2022. doi: 10.52202/068431-1723. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L Croci, Bo Li, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. *arXiv preprint arXiv:2404.00456*, 2024.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Kaichen Zhang* Fanyi Pu* Xinrun Du Yuhao Dong Haotian Liu Yuanhan Zhang Ge Zhang Chunyuan Li Bo Li*, Peiyuan Zhang* and Ziwei Liu. Lmms-eval: Accelerating the development of large multimoal models, March 2024. URL https://github.com/EvolvingLMMs-Lab/lmms-eval.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=sbusw6LD41.

Peijie Dong, Lujun Li, Zhenheng Tang, Xiang Liu, Xinglin Pan, Qiang Wang, and Xiaowen Chu. Pruner-zero: Evolving symbolic pruning metric from scratch for large language models. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024. URL https://arxiv.org/abs/2406.02924. [arXiv: 2406.02924].

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training compression for generative pretrained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *ICLR*, 2016.

Jung Hwan Heo, Jeonghoon Kim, Beomseok Kwon, Byeongwook Kim, Se Jung Kwon, and Dongsoo Lee. Rethinking channel dimensions to isolate outliers for low-bit weight quantization of large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 12744–12762, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/374050dc3f211267bd6bf0ea24eae184-Paper-Conference.pdf.

Shiyao Li, Yingchun Hu, Xuefei Ning, Xihui Liu, Ke Hong, Xiaotao Jia, Xiuhong Li, Yaqi Yan, Pei Ran, Guohao Dai, Shengen Yan, Huazhong Yang, and Yu Wang. Mbq: Modality-balanced quantization for large vision-language models, 2024. URL https://arxiv.org/abs/2412.19509.

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. *arXiv preprint arXiv:2406.01721*, 2024a.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for llm compression and acceleration. In *MLSys*, 2024b.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Jing Liu, Ruihao Gong, Xiuying Wei, Zhiwei Dong, Jianfei Cai, and Bohan Zhuang. QLLM: Accurate and efficient low-bitwidth quantization for large language models. In *International Conference on Learning Representations (ICLR)*, 2024a.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024b.

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

Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? Adaptive rounding for post-training quantization. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7197–7206. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/nagel20a.html.

NVIDIA Corporation. Nvidia ada gpu architecture. White Paper V2.02, NVIDIA Corporation, 2023. URL https://images.nvidia.com/aem-dam/Solutions/geforce/ada/nvidia-ada-gpu-architecture.pdf. Accessed: 2025-07-29.

Jian Ouyang, Xueliang Du, Yin Ma, and Jiaqiang Liu. 3.3 kunlun: A 14nm high-performance ai processor for diversified workloads. In *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, volume 64, pp. 50–51, 2021. doi: 10.1109/ISSCC42613.2021.9366056.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Sai Sanjeet, Ian Colbert, Pablo Monteagudo-Lago, Giuseppe Franco, Yaman Umuroglu, and Nicholas J. Fraser. Mixquant: Pushing the limits of block rotations in post-training quantization. *arXiv preprint arXiv:2601.22347*, 2026.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang Zhang, Peng Gao, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. *arXiv preprint arXiv:2308.13137*, 2023.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Vijay Thakkar, Pradeep Ramani, Cris Cecka, Aniket Shivam, Honghao Lu, Ethan Yan, Jack Kosaian, Mark Hoemmen, Haicheng Wu, Andrew Kerr, Matt Nicely, Duane Merrill, Dustyn Blasig, Aditya Atluri, Fengqi Qiao, Piotr Majcher, Paul Springer, Markus Hohnerbach, Jin Wang, and Manish Gupta. CUTLASS, January 2023. URL https://github.com/NVIDIA/cutlass.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. QuIP: Even better LLM quantization with hadamard incoherence and lattice codebooks. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48630–48656. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/tseng24a.html.

Zhongwei Wan, Ziang Wu, Che Liu, Jinfa Huang, Zhihong Zhu, Peng Jin, Longyue Wang, and Li Yuan. Look-m: Look-once optimization in kv cache for efficient multimodal long-context inference. *arXiv preprint arXiv:2406.18139*, 2024.

Changyuan Wang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. Q-VLM: Post-training quantization for large vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=gxMfNArldP.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Xiuying Wei, Yunchen Zhang, Yuhang Li, Xiangguo Zhang, Ruihao Gong, Jinyang Guo, and Xianglong Liu. Outlier suppression+: Accurate quantization of large language models by equivalent and optimal shifting and scaling. *arXiv preprint arXiv:2304.09145*, 2023.

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*, 2023.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27168–27183. Curran Associates, Inc., 2022. doi: 10.52202/068431-1970. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/adf7fa39d65e2983d724ff7da57f00ac-Paper-Conference.pdf.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024. URL https://arxiv.org/abs/2407.12772.

Yilong Zhao, Chien-Yu Lin, Kan Zhu, Zihao Ye, Lequn Chen, Size Zheng, Luis Ceze, Arvind Krishnamurthy, Tianqi Chen, and Baris Kasikci. Atom: Low-bit quantization for efficient and accurate llm serving. In P. Gibbons, G. Pekhimenko, and C. De Sa (eds.), *Proceedings of Machine Learning and Systems*, volume 6, pp. 196–209, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/file/5edb57c05c81d04beb716ef1d542fe9e-Paper-Conference.pdf.

A Experimental Setup

Base Models. We evaluate COTQuant on three representative vision-language models, including Qwen2-VL-7B, Qwen3-VL-8B, and Qwen3.5-122B. All models use their official pretrained weights without additional fine-tuning.

Benchmark and Datasets. We evaluate the quantized models on five widely used vision-language benchmarks, including OCRBench(Liu et al., 2024b), TextVQA(Singh et al., 2019), VizWiz(Gurari et al., 2018), MME(Fu et al., 2023), and ScienceQA(Lu et al., 2022). These benchmarks cover diverse multimodal capabilities, including optical character recognition, visual question answering, visual perception, and multimodal reasoning. The average score over the five benchmarks is reported as the overall performance metric. We adopt COCO(Lin et al., 2014) as the calibration dataset, from which we randomly sample a subset of 128 samples for calibration. Experiments are conducted on lmms-eval(Zhang et al., 2024; Bo Li* & Liu, 2024), a standard evaluation platform for vision-language benchmarks.

Baselines. We compare COTQuant with representative post-training quantization methods, including SmoothQuant and OmniQuant, which suppress outliers via per-channel scaling, and DuQuant and QuaRot, which adopt orthogonal rotations. We also report the corresponding 16-bit models as full-precision references for measuring the performance degradation introduced by quantization.

Quantization Settings. We focus on W4A4 post-training quantization. Following the quantization setup used throughout this work, activations are quantized in a per-token manner, while weights are quantized with channel-wise or group-wise quantization depending on the evaluated setting. For COTQuant with QAS, we use a group size of 128 unless otherwise specified. For baselines that fail to produce valid outputs under W4A4, we resort to W4A8 quantization; otherwise, all methods are evaluated under W4A4.

Hardware. Experiments on Qwen2-VL-7B and Qwen3-VL-8B are conducted on NVIDIA GeForce RTX 4090 GPUs(NVIDIA Corporation, 2023), while experiments on the larger Qwen3.5-122B model are conducted on Kunlun Xin P800 accelerators(Ouyang et al., 2021).

B Related Work

B.1 Post-Training Quantization for Large Language Models

Post-training quantization (PTQ) has become a practical approach for reducing the memory and computational costs of large language models (LLMs) without expensive retraining. Existing methods can be broadly categorized into weight-only and weight-activation quantization. Weight-only methods, such as GPTQ Frantar et al. (2022) and AWQ Lin et al. (2024b), reduce model storage and memory traffic while maintaining activations at higher precision. In contrast, weight-activation quantization enables low-precision matrix multiplication but is more challenging due to severe activation outliers.

A major line of work therefore focuses on mitigating activation outliers. SmoothQuant Xiao et al. (2023) transfers quantization difficulty from activations to weights through equivalent channel-wise scaling, while OmniQuant Shao et al. (2023) learns equivalent transformations and clipping parameters from calibration data. More recently, orthogonal transformations have been employed to redistribute outlier energy without changing the original full-precision computation. QuaRot Ashkboos et al. (2024) applies randomized Hadamard rotations to obtain more quantization-friendly weight and activation distributions. DuQuant Lin et al. (2024a) further combines block-wise rotations with channel permutation to distribute activation outliers across blocks. These approaches demonstrate the effectiveness of equivalent transformations for low-bit weight-activation quantization, but are primarily developed for LLMs.

Group-wise quantization is widely adopted in low-bit weight quantization, where weight channels are partitioned into fixed-size groups and each group is assigned independent quantization parameters. By reducing the dynamic range covered by each quantizer, it can effectively alleviate weight quantization error. However, the resulting quantization error also depends on how channels are grouped, since different channel assignments induce different within-group ranges and quantization scales.

B.2 Quantization for Vision-Language Models

Most existing PTQ techniques are originally designed for LLMs and treat all input tokens under a shared quantization strategy when transferred to vision-language models (VLMs). Recent works have begun to explicitly study PTQ for VLMs. Q-VLM Wang et al. (2024a) improves quantization by modeling cross-layer dependencies and incorporating them into the search for quantization strategies, while MBQ Li et al. (2024) further identifies different quantization sensitivities between vision and language tokens and incorporates modality-specific sensitivity into the calibration objective. These studies demonstrate the importance of considering VLM-specific characteristics during quantization.

Our method explicitly accounts for modality differences, in particular the channel-wise outlier structures of text and vision activations, which are heterogeneous yet partially overlapping. COTQuant processes multimodal activation outliers via CMT and weight-side distortion via QAS, thereby enabling accurate W4A4 VLM quantization.

C Additional Results

To further evaluate COTQuant under different quantization precisions, we conduct additional experiments under the W6A6 setting on Qwen2-VL-7B and Qwen3-VL-8B. As shown in Table 4, all methods benefit

Table 4: Performance comparison on Qwen2-VL-7B and Qwen3-VL-8B under W6A6 quantization across five vision-language benchmarks. The best quantized result is marked in **bold**.

Method	bits	OCRBench	TextVQA	VizWiz	MME	ScienceQA	Average$^{\uparrow}$
Qwen2-VL-7B	BF16	79.4	81.41	68.76	87.74	84.98	80.46
SmoothQuant	W6A6	77.6	80.05	67.70	87.49	83.99	79.37↓ 1.09
OmniQuant	W6A6	77.8	80.00	66.68	87.11	83.59	79.04↓ 1.42
DuQuant	W6A6	79.2	81.36	68.25	86.98	84.23	80.00↓ 0.46
QuaRot	W6A6	80.0	81.28	68.78	87.70	84.48	80.45↓ 0.01
COTQuant	W6A6	79.4	81.40	69.00	**88.12**	84.88	80.56↑ 0.10
COTQuant(G128)	W6A6	**80.4**	**81.42**	**69.03**	88.04	**85.13**	**80.80**↑ 0.34
Qwen3-VL-8B	BF16	86.5	81.58	69.55	88.80	94.35	84.16
SmoothQuant	W6A6	86.0	80.33	68.48	86.94	92.22	82.79↓ 1.37
OmniQuant	W6A6	85.4	80.10	68.64	86.90	92.17	82.64↓ 1.52
DuQuant	W6A6	85.8	80.91	69.37	87.62	93.90	83.52↓ 0.64
QuaRot	W6A6	85.5	81.46	69.40	88.21	94.25	83.76↓ 0.40
COTQuant	W6A6	86.1	81.64	69.53	88.75	**94.45**	84.09↓ 0.07
COTQuant(G128)	W6A6	**86.3**	**81.66**	**69.90**	**88.92**	94.40	**84.24**↑ 0.08

from the increased numerical precision compared with the more challenging W4A4 setting, while COTQuant consistently achieves the strongest overall performance.

On Qwen2-VL-7B, COTQuant achieves an average score of 80.56. Incorporating QAS further improves the average score to 80.80, yielding a 0.35-point improvement over QuaRot and slightly surpassing the reported BF16 reference. A similar trend is observed on Qwen3-VL-8B, where COTQuant obtains an average score of 84.09, and COTQuant with QAS128 further reaches 84.24, compared with 83.76 for QuaRot. These results demonstrate that the proposed cross-modal transformation and quantization-aware grouping can remain full-precision performance under W6A6 quantization.

D Inference Efficiency

We evaluate the practical inference efficiency of different PTQ methods on Qwen2-VL-7B using real low-bit inference kernels. Following the implementation of QuaRot, we employ CUDA-based 4-bit kernels and use NVIDIA CUTLASS for INT4 GEMM operations on Tensor Cores(Thakkar et al., 2023). All 4-bit methods reduce the model-weight memory from 15810.0 MiB to 6994.9 MiB, corresponding to a 2.26× compression ratio.

As shown in Table 5, real low-bit inference substantially improves both computational and memory efficiency over BF16. COTQuant achieves a prefill throughput of 8202.5, corresponding to a 1.78× speedup, while its decoding throughput reaches 150.7 with a 2.80× speedup. The peak runtime memory is reduced from 15874.0 MiB to 7170.1 MiB, orresponding to a 2.21× compression ratio. Compared with other PTQ transformations, COTQuant maintains comparable runtime efficiency while achieving the strongest average accuracy among the valid quantized methods. These results show that COTQuant preserves the practical

Table 5: Inference efficiency on Qwen2-VL-7B. Prefill throughput is measured at a sequence length of 2048. Throughput is reported in tokens/s.

Method	Average ↑	Prefill		Decode		Peak Memory	
		Throughput ↑	Speedup ↑	Throughput ↑	Speedup ↑	MiB ↓	Compression ↑
BF16	80.46	4597.6	1.00×	53.8	1.00×	15874.0	1.00×
OmniQuant	49.67	9023.8	1.96×	152.5	2.83×	7070.7	2.25×
QuaRot	78.63	10030.4	2.18×	155.5	2.89×	7123.9	2.23×
DuQuant	78.54	7200.8	1.57×	143.4	2.67×	7150.9	2.22×
COTQuant	78.88	8202.5	1.78×	150.7	2.80×	7170.1	2.21×

acceleration and memory advantages of low-bit inference while substantially improving the usability and accuracy of the quantized model.

E QUALITATIVE RESULTS



Figure 3: Qualitative comparison between COTQuant and existing PTQ methods on text-centric vision-language tasks.

Figure 3 presents representative qualitative comparisons between COTQuant and existing PTQ baselines on text-centric vision-language tasks. COTQuant preserves accurate text localization and recognition under low-bit quantization, whereas the compared methods exhibit different failure modes, including irrelevant responses, incorrect text-region localization, and hallucinated textual outputs. These examples qualitatively illustrate the improved robustness of COTQuant in preserving multimodal understanding after quantization.