

# MCTQ: Modality-Coordinated Transformation Quantization for Multimodal Large Language Models

Yingen Liu<sup>†</sup>, Lanfeng Wu<sup>†</sup>, Fan Wu<sup>†,\*</sup>, Liyan Wu<sup>†</sup>, Ruihui Li<sup>†</sup>,  
Zhuo Tang<sup>†</sup>, Eng Siong Chng<sup>‡</sup>, Kenli Li<sup>†</sup>

<sup>†</sup>Hunan University, Changsha, China

<sup>‡</sup>Nanyang Technological University, Singapore

\*Corresponding author: wufan@hnu.edu.cn

## Abstract

Post-training quantization (PTQ) for multimodal large language models (MLLMs) faces a distinct optimization challenge: text and vision tokens can induce conflicting gradients when learning a shared channel-wise transformation. We identify this issue as modality gradient conflict, characterized by opposing update directions and severe magnitude imbalance between modality-specific gradients, and propose Modality-Coordinated Transformation Quantization (MCTQ) to address it. MCTQ combines whitening-based preconditioning, modality-decomposed reconstruction, and modality-coordinated optimization to learn a single shared transformation that serves both modalities without introducing modality-specific parameters or inference overhead. Across multiple MLLMs, benchmarks, and precision settings, MCTQ consistently outperforms existing channel-wise PTQ methods, with the largest gains under aggressive low-bit quantization. These results show that coordinating modality-specific gradients is critical for learning effective shared transformations in multimodal PTQ. Code and reproduction materials are available at <https://github.com/jtsdw/UniMQ>.

## 1 Introduction

Multimodal large language models (MLLMs) (An et al., 2025; Bai et al., 2025; Gemini Team, 2023; OpenAI, 2023) have demonstrated remarkable capabilities in tasks such as visual question answering, image-text reasoning, and multimodal instruction following. However, deploying these models is challenging: beyond the massive parameters of the language model backbone, the substantial number of visual tokens incurs significant memory and computational overhead. To improve deployment efficiency, various compression techniques have been proposed, including pruning (Frantar and Alistarh, 2023), low-rank decomposition (Zhu et al.,

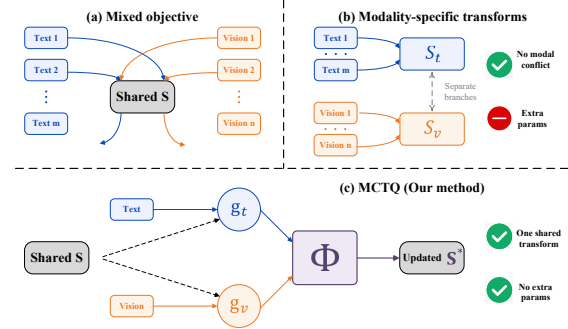


Figure 1: Illustration of shared transformation learning in multimodal PTQ. Mixed-objective optimization can induce conflicting or imbalanced modality updates, whereas modality-specific transforms avoid conflict with extra parameters. MCTQ coordinates modality gradients to learn one shared transformation without additional inference overhead.

2024), and quantization (Xiao et al., 2023; Lin et al., 2024). Among these, post-training quantization (PTQ) (Frantar et al., 2023) is a particularly promising approach—it compresses both weights and activations to low bit-widths, thereby reducing storage, memory bandwidth, and inference cost.

Recent PTQ methods reduce low-bit quantization error by applying channel-wise equivalent transformations before quantization, ranging from statistic-driven scaling (Xiao et al., 2023) to learnable transformations (Shao et al., 2024; Sun et al., 2025). Instead of relying on fixed scaling heuristics derived from activation statistics, these methods directly optimize the transformation via a layer-wise reconstruction objective, so the transformation can adapt to the varying perturbation effects across channels. Learnable equivalent transformations (LET) have therefore become a strong paradigm for low-bit PTQ. However, their suitability for MLLMs remains underexplored: in these models, text and visual tokens are jointly processed by the same quantized backbone, yet they may impose distinct requirements on the transformation that precedes quantization.

Existing MLLM quantization methods have recognized the differences between visual and textual tokens. These methods primarily address differences in activation distributions, numerical ranges, or per-modality sensitivities (Li et al., 2025; Hu et al., 2026; Yu et al., 2025). However, for LET, optimizing a shared transformation over multimodal tokens exposes a distinct challenge: **cross-modal conflict**. Specifically, text and visual tokens may favor different channel-wise scaling factors under a single shared transformation. At the gradient level, this mismatch manifests as modality-specific reconstruction objectives inducing opposing update directions or severely imbalanced magnitudes with respect to the shared transformation parameters. We term this phenomenon modality gradient conflict. This raises a central question: *How can we optimize a single shared transformation in multimodal settings?*

In this paper, we propose Modality-Coordinated Transformation Quantization (MCTQ), a PTQ framework for learning a robust shared transformation in MLLMs. MCTQ consists of three components: (i) diagonal whitening-based preconditioning, which provides a better-conditioned initialization for transformation learning; (ii) modality-decomposed reconstruction, which preserves modality-specific reconstruction gradients; and (iii) modality-coordinated optimization, which reconciles conflicting modality gradients into a unified update direction for the shared transformation. Crucially, MCTQ maintains one shared transformation throughout and fuses it into the quantized model, introducing no modality-specific parameters or inference overhead.

We evaluate MCTQ on multiple MLLMs, including Qwen2.5-VL-7B/32B and LLaVA-OneVision-1.5-8B, under W4A8 and W4A6, and additionally evaluate Qwen2.5-VL-7B under W8A8. Across model scales, architectures, and benchmarks, MCTQ consistently outperforms existing channel-wise PTQ methods, with the largest gains observed in the low-bit regime where gradient conflict is most pronounced.

In summary, our main contributions are:

- We identify modality gradient conflict as a key optimization bottleneck in shared transformation learning for multimodal PTQ, showing that it manifests as both directional interference and modality-dependent gradient dominance.

- We propose MCTQ, which reformulates shared transformation learning as a modality-coordinated optimization problem, enabling text and vision gradients to be reconciled before updating a single deployment-friendly transformation.
- We validate MCTQ on multiple MLLMs and three precision settings, demonstrating consistent improvements over existing channel-wise PTQ methods.

## 2 Related Works

**LLM Quantization.** Quantization reduces the numerical precision of model weights and activations for efficient deployment. Compared with quantization-aware training, post-training quantization (PTQ) is more practical for large language models because it only requires a small calibration set. Existing PTQ methods include weight-only and weight-activation quantization, with the latter being more challenging due to activation outliers and error accumulation. To mitigate these issues, prior work has explored activation smoothing (Xiao et al., 2023), rotation-based transformation (Ashkboos et al., 2024; Liu et al., 2025), and learnable equivalent transformation (Shao et al., 2024; Sun et al., 2025). Our work builds on learnable transformation-based PTQ and investigates *how to adapt it to multimodal settings while retaining a single shared transformation for both text and vision tokens*.

**MLLM Quantization.** Quantizing MLLMs is challenging because text and vision tokens differ in activation distributions, sensitivity patterns, and functional roles. Existing methods address this modality discrepancy from different angles. MBQ (Li et al., 2025) weights reconstruction losses by modality-specific sensitivity to balance text and vision contributions during calibration. MASQuant (Hu et al., 2026) employs modality-specific smoothing factors and uses SVD-based whitening to compensate for cross-modal activation differences in low-rank form. MQuant (Yu et al., 2025) assigns modality-specific static scales and suppresses rotation-induced outliers in Hadamard-based quantization. *Overall, these methods primarily model how modalities differ, rather than how they may conflict when optimizing learnable equivalent transformations.* Our work instead targets the constrained setting of learning a single shared transformation, resolving modality gradient con-

flict at the optimization level while preserving the deployment simplicity of PTQ.

### 3 Preliminaries

#### 3.1 Learned Shared Transformation in PTQ

In reconstruction-based PTQ, a linear layer with input activations  $\mathbf{X} \in \mathbb{R}^{n \times d}$  and weight matrix  $\mathbf{W} \in \mathbb{R}^{o \times d}$  computes the full-precision output  $\mathbf{Y} = \mathbf{X}\mathbf{W}^\top$ . Given an invertible channel-wise transformation matrix  $\mathbf{S} \in \mathbb{R}^{d \times d}$ , the same computation can be rewritten as

$$\mathbf{X}\mathbf{W}^\top = (\mathbf{X}\mathbf{S}^{-1})(\mathbf{W}\mathbf{S})^\top, \quad (1)$$

which redistributes activation and weight ranges across channels while preserving the full-precision function. Let

$$\hat{\mathbf{Y}}_{\mathbf{S}} = Q(\mathbf{X}\mathbf{S}^{-1})Q(\mathbf{W}\mathbf{S})^\top$$

denote the quantized output under the transformation  $\mathbf{S}$ . The transformation is learned by minimizing the layer-wise reconstruction error over a calibration set  $\mathcal{D}_{\text{cal}}$ :

$$\mathcal{L}_{\text{ET}}(\mathbf{S}) = \mathbb{E}_{\mathbf{X} \sim \mathcal{D}_{\text{cal}}} \left[ \left\| \mathbf{X}\mathbf{W}^\top - \hat{\mathbf{Y}}_{\mathbf{S}} \right\|_F^2 \right]. \quad (2)$$

By optimizing  $\mathcal{L}_{\text{ET}}$  directly, these methods learn a transformation  $\mathbf{S}^*$  adapted to the channel-wise effects of quantization perturbations, rather than relying on fixed heuristics derived from activation statistics.

## 4 Method

### 4.1 Overview

To optimize a shared transformation in multimodal settings, we proceed in three steps. First, diagonal whitening stabilizes early optimization by balancing channel scales before transformation learning (Section 4.2). Second, modality-decomposed reconstruction preserves separate gradient signals for text and vision tokens, preventing them from being obscured in a mixed loss (Section 4.3). Third, modality-coordinated optimization reconciles the two modality-specific gradients into a single update direction, resolving conflicts while maintaining parameter sharing (Section 4.4). Figure 2 illustrates the overall MCTQ pipeline.

### 4.2 Whitening-Based Preconditioning

Low-bit quantization is sensitive to channel-wise magnitude variation in activations. In reconstruction-based PTQ, large-magnitude channels dominate the squared reconstruction loss and bias the early optimization of the shared transformation  $\mathbf{S}$ . This issue is more severe in MLLMs, where text and visual tokens may have different dominant channels and inconsistent scale preferences. We therefore apply diagonal whitening before transformation learning to rebalance channel scales and provide a better-conditioned initialization for  $\mathbf{S}$ .

Given the calibration activations  $\mathbf{X} \in \mathbb{R}^{n \times d}$  of a linear layer, we compute the channel-wise standard deviation as

$$\sigma_j = \sqrt{\text{Var}(\mathbf{X}_{:,j}) + \epsilon}, \quad (3)$$

where  $\epsilon$  ensures numerical stability. We construct the diagonal preconditioner

$$\mathbf{P} = \text{diag}(\sigma_1^\alpha, \dots, \sigma_d^\alpha), \quad (4)$$

and whiten the activations by  $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{P}^{-1}$ . Here,  $\alpha \in [0, 1]$  controls the whitening strength:  $\alpha = 0$  yields no whitening,  $\alpha = 1$  yields full diagonal whitening, and intermediate values yield partial whitening. The preconditioner is applied through an equivalent reparameterization:

$$\mathbf{X}\mathbf{W}^\top = (\mathbf{X}\mathbf{P}^{-1})(\mathbf{W}\mathbf{P})^\top. \quad (5)$$

Since  $\mathbf{P}$  is diagonal and invertible, this transformation preserves the full-precision function of the linear layer while redistributing channel-wise scales between activations and weights. Equivalently, for the  $j$ -th input channel, we have

$$\mathbf{x}'_j = \frac{\mathbf{x}_j}{\sigma_j^\alpha}, \quad \mathbf{w}'_j = \sigma_j^\alpha \mathbf{w}_j, \quad (6)$$

where  $\mathbf{w}_j$  denotes the corresponding input-channel column of  $\mathbf{W}$ . In practice, we fold  $\mathbf{P}^{-1}$  into the preceding normalization affine parameters when applicable, and fold  $\mathbf{P}$  into the linear weights, introducing no additional inference operation. This provides better-conditioned channel scales for subsequent transformation learning.

### 4.3 Modality-Decomposed Reconstruction

Recall that  $\mathcal{D}_t$  and  $\mathcal{D}_v$  denote the calibration activations collected from text and vision tokens, respectively. For a shared transformation  $\mathbf{S}$ , we define

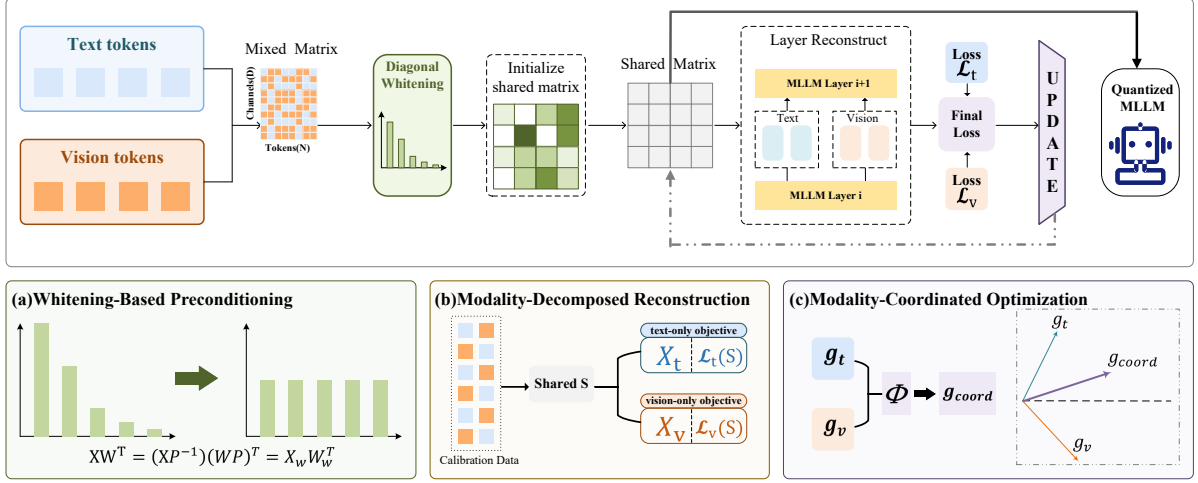


Figure 2: Overview of MCTQ, a shared-transformation PTQ framework that combines whitening-based preconditioning, modality-decomposed reconstruction, and gradient-level coordination.

modality-specific reconstruction losses as

$$\begin{aligned}\mathcal{L}_t(\mathbf{S}) &= \mathbb{E}_{\mathbf{X}_t \sim \mathcal{D}_t} \left[ \left\| \mathbf{X}_t \mathbf{W}^\top - \hat{\mathbf{Y}}_S(\mathbf{X}_t) \right\|_F^2 \right], \\ \mathcal{L}_v(\mathbf{S}) &= \mathbb{E}_{\mathbf{X}_v \sim \mathcal{D}_v} \left[ \left\| \mathbf{X}_v \mathbf{W}^\top - \hat{\mathbf{Y}}_S(\mathbf{X}_v) \right\|_F^2 \right],\end{aligned}\quad (7)$$

where, for  $m \in \{t, v\}$ ,

$$\hat{\mathbf{Y}}_S(\mathbf{X}_m) = Q(\mathbf{X}_m \mathbf{S}^{-1})Q(\mathbf{W}\mathbf{S})^\top. \quad (8)$$

Here, the same transformation  $\mathbf{S}$  is shared by both modalities.

Standard reconstruction-based PTQ merges text and vision calibration tokens and minimizes a single mixed loss  $\mathcal{L}_{\text{mix}}$ . This is equivalent to scalarizing the two modality objectives as  $\rho_t \mathcal{L}_t + \rho_v \mathcal{L}_v$ , where  $\rho_t$  and  $\rho_v$  are determined by the token composition. Although such a mixed objective may decrease during optimization, it can obscure modality-specific degradation, since the two gradients are summed before their compatibility is examined.

To preserve modality-specific optimization signals, we keep the two objectives decomposed and compute their gradients with respect to the same shared transformation:

$$\mathbf{g}_t = \nabla_{\mathbf{S}} \mathcal{L}_t(\mathbf{S}), \quad \mathbf{g}_v = \nabla_{\mathbf{S}} \mathcal{L}_v(\mathbf{S}). \quad (9)$$

Importantly, this decomposition does not introduce modality-specific transformations; both gradients are computed for the same  $\mathbf{S}$ . It only separates the reconstruction objectives during calibration, exposing the independent update signals from text and vision tokens before the shared transformation is updated.

#### 4.4 Modality-Coordinated Optimization

Given the modality-specific gradients  $\mathbf{g}_t$  and  $\mathbf{g}_v$  from Eq. (9), our goal is to update the shared transformation  $\mathbf{S}$  using a single direction that accounts for both modalities. A naive mixed update,

$$\mathbf{g}_{\text{mix}} = \rho_t \mathbf{g}_t + \rho_v \mathbf{g}_v, \quad (10)$$

uses fixed scalar weights determined by token composition or loss weighting. However, such scalarization ignores the geometry between modality gradients. When  $\langle \mathbf{g}_t, \mathbf{g}_v \rangle < 0$ , the two modalities induce conflicting update directions, so descending along one gradient may increase the reconstruction error of the other. When  $\|\mathbf{g}_t\|_2$  and  $\|\mathbf{g}_v\|_2$  differ substantially, the mixed update is dominated by the larger gradient, suppressing the optimization signal from the weaker modality.

To coordinate the two modalities, we define a gradient-level operator  $\Phi$  that maps  $\mathbf{g}_t$  and  $\mathbf{g}_v$  to a unified update direction:

$$\mathbf{g}_{\text{coord}} = \Phi(\mathbf{g}_t, \mathbf{g}_v). \quad (11)$$

Instead of assigning fixed loss weights before back-propagation,  $\Phi$  adaptively combines the two gradients according to their direction and magnitude. Specifically, following the MGDA minimum-norm formulation (Sener and Koltun, 2018), we adopt a minimum-norm convex combination:

$$\min_{w \in [0,1]} \|w \mathbf{g}_t + (1-w) \mathbf{g}_v\|_2^2, \quad (12)$$

and set

$$\mathbf{g}_{\text{coord}} = w^* \mathbf{g}_t + (1-w^*) \mathbf{g}_v. \quad (13)$$

---

**Algorithm 1** MCTQ pipeline for one linear layer.

---

**Require:** Weight  $\mathbf{W}$ , calibration activations  $\mathbf{X}_{\text{mix}}$ , modality masks  $\mathcal{M}_t, \mathcal{M}_v$ , whitening strength  $\alpha$ , learning rate  $\eta_s$ , epochs  $E$

**Ensure:** Quantized linear layer with fused transformation

- 1: Split  $\mathbf{X}_{\text{mix}}$  into  $\mathbf{X}_t$  and  $\mathbf{X}_v$  using  $\mathcal{M}_t, \mathcal{M}_v$
  - 2: Compute diagonal whitening preconditioner  $\mathbf{P}$  using Eq. (4)
  - 3: Fold  $\mathbf{P}^{-1}$  into the preceding affine operation and set  $\mathbf{W} \leftarrow \mathbf{W}\mathbf{P}$
  - 4: Recompute  $\mathbf{X}_t, \mathbf{X}_v$  with the whitened model
  - 5: Initialize shared transformation  $\mathbf{S} \leftarrow \mathbf{I}$
  - 6: **for**  $e = 1$  to  $E$  **do**
  - 7:   Compute modality losses  $\mathcal{L}_t(\mathbf{S})$  and  $\mathcal{L}_v(\mathbf{S})$  using Eq. (7)
  - 8:   Compute gradients  $\mathbf{g}_t \leftarrow \nabla_{\mathbf{S}} \mathcal{L}_t$ ,  $\mathbf{g}_v \leftarrow \nabla_{\mathbf{S}} \mathcal{L}_v$
  - 9:   Obtain coordinated gradient  $\mathbf{g}_{\text{coord}} \leftarrow \Phi(\mathbf{g}_t, \mathbf{g}_v)$  using Eq. (14)
  - 10:   Update  $\mathbf{S} \leftarrow \mathbf{S} - \eta_s \mathbf{g}_{\text{coord}}$
  - 11: **end for**
  - 12: Fold  $\mathbf{S}^{\star-1}$  into the preceding affine operation and quantize  $\mathbf{W}\mathbf{S}^{\star}$
- 

We provide a theoretical justification for this choice in Appendix D.

For two modalities, Eq. (12) admits a closed-form solution:

$$w^{\star} = \Pi_{[0,1]} \left( \frac{\|\mathbf{g}_v\|_2^2 - \langle \mathbf{g}_t, \mathbf{g}_v \rangle}{\|\mathbf{g}_t - \mathbf{g}_v\|_2^2} \right), \quad (14)$$

where  $\Pi_{[0,1]}(\cdot)$  denotes clipping to the interval  $[0, 1]$ . The shared transformation is then updated by

$$\mathbf{S} \leftarrow \mathbf{S} - \eta \mathbf{g}_{\text{coord}}. \quad (15)$$

This coordination has two desirable effects. First, when the modality gradients point in incompatible directions, the solution searches for a direction in their convex hull that minimizes mutual interference, rather than blindly following their scalar-weighted sum. Second, when one modality produces a much larger gradient norm, the optimal weight automatically down-weights the dominant gradient, preventing it from overwhelming the shared update. Therefore, MCTQ resolves modality conflict after the modality-specific gradients are exposed, while still updating a single shared transformation  $\mathbf{S}$ .

## 5 Experiments

### 5.1 Experimental Setup

**Model and Quantization Setting.** We conduct experiments on Qwen2.5-VL-7B-Instruct, Qwen2.5-VL-32B-Instruct (Bai et al., 2025), and LLaVA-OneVision-1.5-8B (An et al., 2025). Our primary setting is W4A6. We additionally report W4A8 for all models and W8A8 for Qwen2.5-VL-7B. We use per-token activation quantization and per-channel weight quantization, with min-max asymmetric quantization for activations and min-max symmetric quantization for weights.

**Calibration Data.** We use 128 image-caption pairs from the COCO val2017 caption annotations (Lin et al., 2014) as multimodal calibration data. Each sample is converted into the chat-format input of Qwen2.5-VL (Bai et al., 2025) and is used for both activation-statistics collection and reconstruction-based optimization. For whitening-based methods, activation statistics are recollected after whitening.

**Baselines.** We compare MCTQ against both LLM-oriented and MLLM-oriented channel-wise PTQ methods: SmoothQuant (Xiao et al., 2023) and OmniQuant (Shao et al., 2024) are representative channel-wise transformation methods originally designed for LLMs (statistic-based and learnable, respectively), while MASQuant (Hu et al., 2026) and MBQ (Li et al., 2025) are recent methods specifically tailored for MLLMs.

**Implementation Details.** Unless otherwise specified, we apply partial diagonal whitening with  $\alpha_{\text{attn}} = 0.3$  and  $\alpha_{\text{mlp}} = 0.3$ , followed by 30 reconstruction epochs for transformation learning. We set the learning rates to  $5 \times 10^{-4}$  for the shared transformation and  $1 \times 10^{-3}$  for the clipping parameters.

**Evaluation Benchmarks.** We evaluate on six vision-language benchmarks: OCRBench (Liu et al., 2024), SEEDBench (Li et al., 2024), MMMU (Yue et al., 2024), VizWiz (Gurari et al., 2018), ScienceQA (Lu et al., 2022), and TextVQA (Singh et al., 2019). We report task accuracy and the arithmetic mean across all tasks.

### 5.2 Main Results

Table 1 reports W8A8 results on Qwen2.5-VL-7B and W4A8/W4A6 results on both Qwen2.5-VL-7B and Qwen2.5-VL-32B (Bai et al., 2025).

We further evaluate MCTQ on LLaVA-OneVision-1.5-8B (An et al., 2025), which adopts a different vision-language architecture, and provide the full results in Appendix A.

On Qwen2.5-VL-7B at W8A8, all methods remain close to the FP16 baseline, indicating that this setting is relatively mild for existing PTQ methods. MCTQ achieves an average accuracy of 68.71, maintaining performance comparable to the full-precision model. As the precision decreases, the advantage of MCTQ becomes more evident. Under the primary W4A6 setting, MCTQ achieves 66.26 average accuracy on Qwen2.5-VL-7B, outperforming OmniQuant by 1.30 points and SmoothQuant by 6.36 points. On Qwen2.5-VL-32B, MCTQ reaches 66.51 under W4A6, surpassing OmniQuant and MASQuant by more than two points. A similar trend is observed under W4A8, where MCTQ consistently obtains the best average performance across both Qwen2.5-VL models.

### 5.3 Analysis of Modality Gradient Conflict

To verify whether multimodal PTQ suffers from optimization conflict on the shared transformation, we analyze modality-specific gradients before applying coordinated optimization. We decompose the reconstruction objective into text- and vision-conditioned losses, compute their gradients with respect to the same shared transformation parameters, and record the cosine similarity and norm ratio between the two gradients for each sampled update. The cosine similarity measures directional compatibility, while the norm ratio reflects modality-dependent dominance.

Figure 3 shows that many samples either fall into the negative-cosine region or exhibit large gradient-norm ratios. This indicates that modality conflict is not limited to opposing gradient directions; it also appears as severe imbalance where one modality dominates the shared update. Representative layer-wise histograms in Figure 4 further confirm that such patterns occur across different layers and transformation types. These results provide direct evidence that a unified mixed objective can obscure incompatible or biased modality updates, motivating MCTQ to preserve and coordinate modality-specific gradients before updating the shared transformation. Additional visualizations are provided in Appendix C.

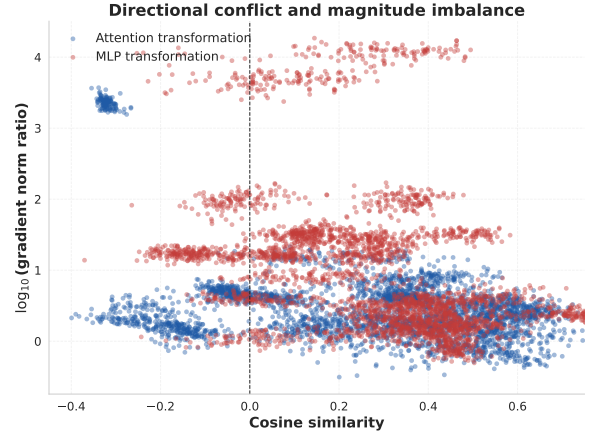


Figure 3: Global visualization of modality-gradient conflict on shared transformation parameters. The horizontal axis shows the cosine similarity between text and vision gradients, while the vertical axis shows the logarithm of their norm ratio.

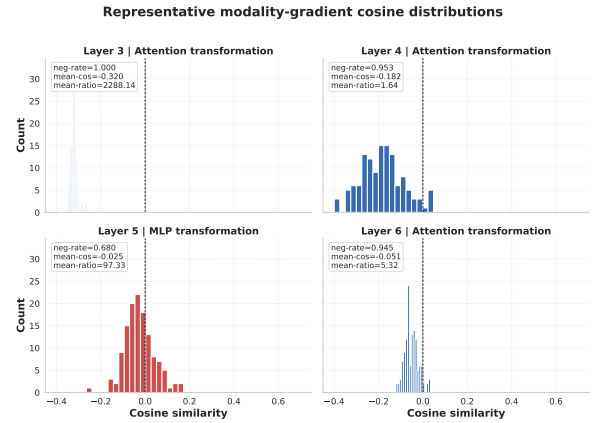


Figure 4: Representative examples of modality-gradient cosine distributions on shared transformation parameters. Each panel corresponds to one illustrative layer group and reports its negative-alignment rate, mean cosine similarity, and mean gradient norm ratio.

### 5.4 Gradient-Scale Intervention

To test whether gradient-magnitude imbalance is an actionable optimization factor rather than merely a correlational observation, we directly rescale only the text-side transformation gradient as  $\tilde{\mathbf{g}}_t = \alpha_t \mathbf{g}_t$  during shared-transformation learning on Qwen2.5-VL-7B under W4A6, while keeping the vision gradient and all other settings fixed. Table 2 reports the resulting median text-to-vision gradient-norm ratio and downstream performance. The most balanced setting,  $\alpha_t = 0.5$ , yields the best average accuracy, whereas shifting the ratio toward either modality reduces performance. This controlled intervention supports gradient imbalance as an important and actionable bottleneck, while not implying that it is

| Model          | Bitwidth | Method      | SEED         | VizWiz       | ScienceQA    | TextVQA      | MMMU         | OCRBench     | Avg. ( $\uparrow$ ) |
|----------------|----------|-------------|--------------|--------------|--------------|--------------|--------------|--------------|---------------------|
| Qwen2.5-VL-7B  | FP16     | –           | 74.36        | 42.65        | 88.59        | 82.58        | 50.89        | 68.90        | 68.00               |
|                | W8A8     | SmoothQuant | 74.30        | 42.87        | 88.45        | 82.44        | 50.44        | 69.00        | 67.92               |
|                |          | OmniQuant   | 74.33        | <b>44.12</b> | <b>88.80</b> | 85.44        | 50.67        | 69.00        | <b>68.72</b>        |
|                |          | MASQuant    | 74.29        | 42.86        | 88.33        | <b>85.49</b> | 50.00        | 68.70        | 68.28               |
|                |          | MBQ         | 74.44        | 42.83        | 88.54        | 82.50        | <b>50.89</b> | <b>69.20</b> | 68.07               |
|                |          | MCTQ        | <b>74.47</b> | 43.84        | 88.59        | 85.38        | <b>50.89</b> | 69.10        | 68.71               |
|                | W4A8     | SmoothQuant | 72.08        | 39.05        | 81.33        | 76.24        | 41.78        | 62.30        | 62.13               |
|                |          | OmniQuant   | 73.74        | 40.91        | 86.11        | 83.57        | 46.78        | 67.10        | 66.37               |
|                |          | MASQuant    | 73.47        | 39.84        | 86.35        | 83.65        | 46.11        | 68.00        | 66.24               |
|                |          | MBQ         | 72.83        | 35.72        | 84.48        | 69.45        | 43.89        | 57.20        | 60.60               |
|                |          | MCTQ        | <b>74.01</b> | <b>41.38</b> | <b>86.72</b> | <b>83.91</b> | <b>49.44</b> | <b>68.60</b> | <b>67.34</b>        |
|                | W4A6     | SmoothQuant | 69.91        | 38.04        | 78.99        | 74.16        | 38.67        | 59.60        | 59.90               |
|                |          | OmniQuant   | 72.87        | 39.19        | 84.67        | 81.78        | 44.56        | 66.70        | 64.96               |
|                |          | MASQuant    | 67.65        | 25.95        | 67.11        | 67.47        | 40.00        | 55.40        | 53.93               |
|                |          | MBQ         | 69.57        | 33.22        | 76.56        | 66.03        | 41.89        | 54.30        | 56.93               |
|                |          | MCTQ        | <b>73.69</b> | <b>41.36</b> | <b>85.31</b> | <b>82.83</b> | <b>46.78</b> | <b>67.60</b> | <b>66.26</b>        |
| Qwen2.5-VL-32B | FP16     | –           | 74.21        | 41.60        | 92.24        | 77.14        | 58.00        | 66.70        | 68.32               |
|                | W4A8     | SmoothQuant | 72.92        | 41.35        | 69.30        | 75.87        | 53.56        | 64.30        | 62.88               |
|                |          | OmniQuant   | 73.29        | 41.22        | 88.29        | 78.47        | 55.67        | 64.90        | 66.97               |
|                |          | MASQuant    | 73.49        | 39.87        | 87.88        | 77.40        | <b>56.22</b> | 64.10        | 66.49               |
|                |          | MBQ         | 73.20        | 37.93        | 84.44        | 75.79        | 53.67        | 63.10        | 64.69               |
|                |          | MCTQ        | <b>73.62</b> | <b>42.30</b> | <b>90.31</b> | <b>78.68</b> | 54.00        | <b>65.10</b> | <b>67.34</b>        |
|                | W4A6     | SmoothQuant | 71.31        | 37.58        | 80.15        | 73.10        | 49.89        | 62.90        | 62.49               |
|                |          | OmniQuant   | 72.86        | 40.29        | 75.17        | 76.89        | <b>54.67</b> | <b>65.10</b> | 64.16               |
|                |          | MASQuant    | 72.60        | 35.56        | 85.59        | 74.64        | 53.33        | 63.30        | 64.17               |
|                |          | MBQ         | 68.44        | 32.91        | 66.12        | 72.57        | 49.22        | 60.80        | 58.34               |
|                |          | MCTQ        | <b>73.02</b> | <b>42.71</b> | <b>87.79</b> | <b>78.15</b> | 52.89        | 64.50        | <b>66.51</b>        |

Table 1: Main results on Qwen2.5-VL-7B under W8A8, W4A8, and W4A6, and on Qwen2.5-VL-32B under W4A8 and W4A6. The best result under each bit-width is highlighted in bold.

| $\alpha_t$ | Ratio       | OCR          | MMMU         | SQA          | TVQA         | Avg.         |
|------------|-------------|--------------|--------------|--------------|--------------|--------------|
| 0.25       | 0.63        | 65.40        | 45.33        | <b>85.69</b> | 82.33        | 69.69        |
| 0.50       | <b>0.85</b> | <b>67.50</b> | 47.67        | <b>85.69</b> | 82.19        | <b>70.76</b> |
| 1.00       | 1.78        | 66.10        | <b>48.44</b> | 85.10        | 82.29        | 70.48        |
| 2.00       | 3.46        | 66.00        | 45.67        | 84.72        | <b>82.75</b> | 69.78        |

Table 2: Intervention on the text-side gradient scale under W4A6. Ratio denotes the median text-to-vision gradient-norm ratio; values closer to one indicate better magnitude balance.

the sole cause of quantization degradation.

## 5.5 Ablation Study

We conduct three ablation studies under the W4A6 setting to isolate the contribution of each MCTQ component and to compare coordination strategies. All experiments in this subsection use four representative tasks (MMMU, OCRBench, ScienceQA, and TextVQA) for efficient evaluation. Additional ablations on reconstruction epochs and calibration set size are provided in Appendix B.

**Component ablation.** Table 3 ablates the main designs of MCTQ under the W4A6 setting. Starting from the vanilla LET baseline, adding whitening-based preconditioning improves the average accuracy from 67.50 to 68.93, indicating that channel-scale rebalancing provides a better-conditioned initialization for shared transformation learning. Applying modality-decomposed reconstruction and coordinated optimization without whitening further improves the average accuracy to 69.39, outperforming LET by +1.89. This shows that explicitly preserving and coordinating modality-specific gradients is beneficial even without preconditioning. The full MCTQ pipeline achieves the best average accuracy of 70.63, improving over LET by +3.13, over LET + W by +1.70, and over LET + D + C by +1.24. These results suggest that whitening-based preconditioning and modality-coordinated optimization are complementary: the former improves the channel-scale geometry, while the latter reconciles text and vision gradients into a more effective shared update.

| Variant         | MMMU         | OCR          | SQA          | TVQA         | Avg.         |
|-----------------|--------------|--------------|--------------|--------------|--------------|
| LET             | 43.56        | 62.50        | 83.83        | 80.11        | 67.50        |
| LET + W         | 45.00        | 64.90        | 85.05        | 80.77        | 68.93        |
| LET + D + C     | 45.33        | 66.20        | 84.24        | 81.79        | 69.39        |
| LET + W + D + C | <b>46.78</b> | <b>67.60</b> | <b>85.31</b> | <b>82.83</b> | <b>70.63</b> |

Table 3: Component ablation under the W4A6 setting. W, D, and C denote whitening-based preconditioning, modality-decomposed reconstruction, and modality-coordinated optimization, respectively. LET + W + D + C corresponds to the full MCTQ pipeline.

| Coordinator         | MMMU         | OCR          | SQA          | TVQA         | Avg.         |
|---------------------|--------------|--------------|--------------|--------------|--------------|
| Vanilla             | 45.33        | <b>67.90</b> | 85.12        | 82.57        | 70.23        |
| Projection (PCGrad) | 46.33        | 65.80        | 84.51        | 82.41        | 69.76        |
| Ours                | <b>47.67</b> | 67.40        | <b>85.83</b> | <b>82.84</b> | <b>70.94</b> |

Table 4: Comparison of coordination strategies for  $\Phi$  under the W4A6 setting. Ours performs minimum-norm coordination by considering both gradient direction and magnitude.

**Coordination strategy.** Table 4 evaluates alternative designs for the coordination operator  $\Phi$ . Vanilla summation directly aggregates modality-specific gradients without explicit conflict resolution. Projection-based reconciliation, implemented with PCGrad (Yu et al., 2020), mainly addresses directional conflict by removing opposing gradient components, but does not explicitly correct magnitude imbalance. Its lower average performance than vanilla summation (69.76 vs. 70.23) suggests that handling directional conflict alone may be insufficient for multimodal shared transformation learning. In contrast, our coordinator accounts for both gradient direction and norm, achieving the best average accuracy of 70.94 and leading on three of four tasks.

**Whitening strength.** We sweep  $\alpha_{\text{attn}}$  and  $\alpha_{\text{mlp}}$  over  $\{0, 0.3, 0.6, 1.0\}$  under W4A6 and report five representative configurations in Table 5. Without whitening, the average accuracy drops to 68.85, while full whitening remains suboptimal at 69.39. Partial whitening performs best, with  $\alpha_{\text{attn}} = \alpha_{\text{mlp}} = 0.3$  achieving the highest score on all reported metrics and an average accuracy of 70.73. Other moderate settings also remain competitive, achieving 70.12 and 70.05. These results suggest that partial whitening provides effective preconditioning, whereas overly weak or overly aggressive whitening is less suitable.

| $\alpha_{\text{attn}}$ | $\alpha_{\text{mlp}}$ | MMMU         | OCR          | SQA          | TVQA         | Avg.         |
|------------------------|-----------------------|--------------|--------------|--------------|--------------|--------------|
| 0.0                    | 0.0                   | 46.00        | 65.40        | 81.38        | 82.63        | 68.85        |
| 0.3                    | 0.3                   | <b>47.18</b> | <b>67.60</b> | <b>85.31</b> | <b>82.83</b> | <b>70.73</b> |
| 0.6                    | 0.3                   | 46.89        | 66.80        | 84.14        | 82.63        | 70.12        |
| 0.6                    | 0.6                   | 46.44        | 66.30        | 84.96        | 82.49        | 70.05        |
| 1.0                    | 1.0                   | 46.89        | 66.10        | 82.31        | 82.27        | 69.39        |

Table 5: Effect of whitening strength  $\alpha$  under the W4A6 setting. Partial whitening with  $\alpha_{\text{attn}} = \alpha_{\text{mlp}} = 0.3$  achieves the best overall performance.

## 5.6 Efficiency Analysis

We evaluate the practical deployment efficiency of the MCTQ-quantized model by measuring prefill latency and peak GPU memory usage on a single NVIDIA A100-80G GPU under W4A8 quantization. We report efficiency under W4A8 because 8-bit activation inference is more mature in current deployment stacks, while W4A6 is used as an aggressive accuracy-oriented setting for evaluating quantization robustness.<sup>1</sup> Table 6 reports results across two prefill lengths, 1024 and 2048 tokens, and two batch sizes, 1 and 4. Compared with the FP16 baseline, W4A8 quantization achieves  $1.31\times$ – $1.38\times$  prefill speedup and  $1.10\times$ – $1.60\times$  peak memory saving. The memory reduction is more pronounced at batch size 1, where model weights account for a larger fraction of peak memory. At larger batch sizes, the relative gain becomes smaller because the KV cache and runtime activations occupy a larger share of memory and are not quantized in our setting. Importantly, MCTQ fuses the learned shared transformation into the quantized weights, introducing no modality-specific parameters, runtime branches, or additional transformation operations. Thus, the accuracy gains of MCTQ are obtained while preserving the deployment efficiency of standard quantized linear inference.

## 6 Conclusion

In this paper, we identified modality gradient conflict as a key optimization bottleneck in shared transformation learning for multimodal PTQ, manifested as both directional interference and modality-dependent gradient-magnitude dominance. To address this issue, we proposed MCTQ, a modality-coordinated PTQ frame-

<sup>1</sup>Although recent GPU architectures have started to support emerging low-precision floating-point formats such as FP6/FP4, widely available INT6 activation kernels remain less standardized than 4-bit, 8-bit, and FP8 inference kernels.

| Len. | BS | Method | Time (ms)     | Spd.         | Mem. (GB)    | Save.        |
|------|----|--------|---------------|--------------|--------------|--------------|
| 1024 | 1  | FP16   | 148.84        | 1.00×        | 17.12        | 1.00×        |
|      |    | W4A8   | <b>113.55</b> | <b>1.31×</b> | <b>10.69</b> | <b>1.60×</b> |
| 1024 | 4  | FP16   | 603.54        | 1.00×        | 18.45        | 1.00×        |
|      |    | W4A8   | <b>438.82</b> | <b>1.38×</b> | <b>14.03</b> | <b>1.32×</b> |
| 2048 | 1  | FP16   | 299.85        | 1.00×        | 17.57        | 1.00×        |
|      |    | W4A8   | <b>220.90</b> | <b>1.36×</b> | <b>11.80</b> | <b>1.49×</b> |
| 2048 | 4  | FP16   | 1265.98       | 1.00×        | 20.33        | 1.00×        |
|      |    | W4A8   | <b>948.59</b> | <b>1.33×</b> | <b>18.53</b> | <b>1.10×</b> |

Table 6: Efficiency comparison between FP16 and W4A8 quantization.

work that combines whitening-based preconditioning, modality-decomposed reconstruction, and modality-coordinated optimization to learn a single shared transformation for both text and vision tokens. Experiments across multiple MLLMs, multimodal benchmarks, and precision settings demonstrate that MCTQ consistently improves quantization quality, with the most pronounced gains under aggressive low-bit quantization. Our results highlight the importance of gradient-level coordination in multimodal model compression and suggest that optimization-aware quantization is a promising direction for building efficient MLLMs.

## Limitations

Although MCTQ demonstrates consistent improvements for multimodal PTQ, it still has several limitations.

First, our current study focuses on vision-language MLLMs and considers two modalities: text and vision. However, the proposed formulation is not inherently restricted to two modalities: modality-decomposed reconstruction and gradient-level coordination can in principle be extended to additional modalities such as speech or audio. Nevertheless, omni-modal models may introduce more complex modality interactions, heterogeneous token structures, and higher-order gradient conflicts beyond pairwise text-vision coordination. Extending MCTQ to omni-modal quantization is therefore an important direction for future work.

Second, although MCTQ shows improvements under W4A4 quantization, there remains a considerable gap between W4A4 performance and the full-precision baseline. This suggests that extremely low-bit weight-activation quantization for

MLLMs is still far from solved. In particular, when both weights and activations are compressed to 4 bits, quantization noise, modality-specific activation outliers, and cross-modal optimization conflict become substantially more severe. Future work may explore stronger transformation parameterizations, more fine-grained modality coordination, or joint treatment of activation clipping and transformation learning to further improve robustness in this aggressive regime.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62472162, 62225205, and 62532005).

## References

- Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, et al. 2025. [LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training](#). *Preprint*, arXiv:2509.23661.
- Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefer, and James Hensman. 2024. [QuaRot: Outlier-free 4-bit inference in rotated LLMs](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 100213–100240.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, et al. 2025. [Qwen2.5-VL technical report](#). *Preprint*, arXiv:2502.13923.
- Elias Frantar and Dan Alistarh. 2023. SparseGPT: Massive language models can be accurately pruned in one-shot. In *International conference on machine learning*, pages 10323–10337. PMLR.
- Elias Frantar, Saleh Ashkboos, Torsten Hoefer, and Dan Alistarh. 2023. [GPTQ: Accurate post-training quantization for generative pre-trained transformers](#). In *The Eleventh International Conference on Learning Representations*.
- Gemini Team. 2023. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805.
- Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of*

- the IEEE conference on computer vision and pattern recognition, pages 3608–3617.
- Lulu Hu, Wenhui Xiao, Xin Chen, Xinhua Xu, Bowen Xu, Kun Li, and Yongliang Tao. 2026. MASQuant: Modality-aware smoothing quantization for multimodal large language models. *arXiv preprint arXiv:2603.04800*.
- Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. SEED-Bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.
- 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. 2025. MBQ: Modality-balanced quantization for large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4167–4177.
- Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. *Proceedings of machine learning and systems*, 6:87–100.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. 2021. Conflict-averse gradient descent for multi-task learning. *Advances in neural information processing systems*, 34:18878–18890.
- Xiaohao Liu, Xiaobo Xia, Manyi Zhang, Ji-Fu Li, Xianzhi Yu, Fei Shen, Xiu Su, See-Kiong Ng, and Tat-Seng Chua. 2026. FreeAct: Freeing activations for LLM quantization. *arXiv preprint arXiv:2603.01776*.
- Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024. OCR-Bench: On the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, 67(12):220102.
- Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. 2025. SpinQuant: LLM quantization with learned rotations. In *The Thirteenth International Conference on Learning Representations*.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521.
- OpenAI. 2023. [Gpt-4 technical report](#). Preprint, arXiv:2303.08774.
- Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31.
- Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Gao Peng, Yu Qiao, and Ping Luo. 2024. [OmniQuant: Omnidirectionally calibrated quantization for large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.
- 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. 2025. [FlatQuant: Flatness matters for LLM quantization](#). In *Proceedings of the 42nd International Conference on Machine Learning*.
- Changyuan Wang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. 2024. Q-VLM: Post-training quantization for large vision-language models. *Advances in Neural Information Processing Systems*, 37:114553–114573.
- Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *International conference on machine learning*, pages 38087–38099. PMLR.
- JiangYong Yu, Sifan Zhou, Dawei Yang, Shuoyu Li, Shuo Wang, Xing Hu, Chen Xu, Zukang Xu, Changyong Shu, and Zhihang Yuan. 2025. [MQuant: Unleashing the inference potential of multimodal large language models via static quantization](#). In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 1783–1792.
- Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, et al. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Yuan Zhang, Chunkai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. 2024. [SparseVLM: Visual token sparsification for efficient vision-language model inference](#). *ArXiv*, abs/2410.04417.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2024. [A survey on model compression for large language models](#). *Transactions of the Association for Computational Linguistics*, 12:1556–1577.

## A Additional Results

This section provides additional empirical results that complement the main paper.

Table 7 reports cross-architecture results on LLaVA-OneVision-1.5-8B, which differs from Qwen2.5-VL in model architecture and vision-language integration. MCTQ obtains the best average performance under both W4A8 and W4A6 settings, improving over the strongest baseline from 66.15 to 66.55 and from 63.14 to 64.60, respectively. Although the gains are more moderate than those observed on Qwen2.5-VL, the consistent improvements indicate that MCTQ is not limited to a specific model family and can generalize to different MLLM architectures.

Table 8 presents an extreme W4A4 stress test on Qwen2.5-VL-7B. Under this highly aggressive setting, existing PTQ baselines largely collapse, with average accuracy dropping to very low levels across multimodal benchmarks. In contrast, MCTQ substantially alleviates this degradation and retains non-trivial performance, improving the average accuracy from the best baseline score of 6.30 to 19.40. This result suggests that coordinating modality-specific gradients is especially beneficial when quantization noise becomes severe. Nevertheless, a large gap to the FP16 baseline remains, indicating that fully practical W4A4 quantization for MLLMs is still an open challenge. We therefore view W4A4 as a stress-test regime rather than a fully deployment-ready setting, and leave more robust extreme low-bit multimodal quantization to future work.

## B Training Dynamics and Calibration Robustness

Table 9 reports MCTQ performance under varying reconstruction epochs under the W4A6 setting on Qwen2.5-VL-7B. The performance improves consistently as the number of reconstruction epochs increases from 5 to 30, indicating that sufficient transformation learning is important for effective modality-coordinated optimization. The best average accuracy is achieved at 30 epochs, while further increasing the number of epochs to 50 leads to performance degradation, suggesting mild overfitting to the calibration set. Therefore, we use 30 reconstruction epochs as the default setting, which provides the best accuracy while avoiding unnecessary calibration overfitting.

Table 10 evaluates MCTQ with different calibra-

| Model                  | Bitwidth | Method      | SEED         | VizWiz       | ScienceQA    | TextVQA      | MMMU         | OCRBench     | Avg. ( $\uparrow$ ) |
|------------------------|----------|-------------|--------------|--------------|--------------|--------------|--------------|--------------|---------------------|
| LLaVA-OneVision-1.5-8B | FP16     | –           | 72.47        | 39.81        | 94.81        | 79.02        | 57.00        | 65.50        | 68.10               |
|                        | W4A8     | SmoothQuant | 69.20        | 25.15        | 87.62        | 63.12        | 51.22        | 58.30        | 59.10               |
|                        |          | OmniQuant   | <b>71.25</b> | 35.53        | <b>93.40</b> | 75.60        | 55.44        | <b>65.70</b> | 66.15               |
|                        |          | MASQuant    | 70.12        | 31.18        | 85.38        | 70.68        | <b>58.78</b> | 62.50        | 63.11               |
|                        |          | MBQ         | 69.46        | 32.90        | 91.77        | 69.53        | 49.78        | 59.90        | 62.22               |
|                        |          | MCTQ        | 69.97        | <b>37.92</b> | 93.26        | <b>77.07</b> | 56.78        | 64.30        | <b>66.55</b>        |
|                        | W4A6     | SmoothQuant | 66.80        | 21.32        | 84.30        | 54.31        | 45.56        | 54.80        | 54.52               |
|                        |          | OmniQuant   | 69.21        | 33.53        | 88.32        | 73.34        | 53.31        | 61.10        | 63.14               |
|                        |          | MASQuant    | 68.42        | 27.54        | 77.39        | 67.32        | <b>55.78</b> | 60.80        | 59.54               |
|                        |          | MBQ         | 66.22        | 29.14        | 83.57        | 63.67        | 45.67        | 55.10        | 57.23               |
|                        |          | MCTQ        | <b>70.08</b> | <b>35.09</b> | <b>90.26</b> | <b>74.29</b> | 54.89        | 63.00        | <b>64.60</b>        |

Table 7: Quantization results on LLaVA-OneVision-1.5-8B (An et al., 2025) under W4A8 and W4A6 settings. The best result under each bit-width is highlighted in bold.

| Model         | Bitwidth | Method      | SEED         | VizWiz      | ScienceQA    | TextVQA      | MMMU         | OCRBench     | Average ( $\uparrow$ ) |
|---------------|----------|-------------|--------------|-------------|--------------|--------------|--------------|--------------|------------------------|
| Qwen2.5-VL-7B | FP16     | –           | 74.36        | 42.65       | 88.59        | 82.58        | 50.89        | 68.90        | 68.00                  |
|               | W4A4     | SmoothQuant | 3.61         | 0.00        | 4.83         | 0.05         | 17.44        | 0.40         | 4.39                   |
|               |          | OmniQuant   | 8.79         | 0.10        | 1.70         | 0.11         | 25.00        | 2.10         | 6.30                   |
|               |          | MASQuant    | 1.90         | 0.00        | 1.65         | 0.01         | 10.33        | 0.00         | 2.32                   |
|               |          | MBQ         | 4.68         | 0.00        | 0.57         | 0.02         | 26.56        | 0.10         | 5.32                   |
|               |          | MCTQ        | <b>25.02</b> | <b>4.88</b> | <b>21.67</b> | <b>13.03</b> | <b>33.67</b> | <b>18.10</b> | <b>19.40</b>           |

Table 8: Extreme W4A4 quantization stress test on Qwen2.5-VL-7B (Bai et al., 2025). Most existing PTQ baselines nearly collapse under W4A4, while MCTQ retains non-trivial performance and substantially improves the average accuracy.

| Epochs | OCR          | MMMU         | SQA          | TVQA         | Avg.         |
|--------|--------------|--------------|--------------|--------------|--------------|
| 5      | 66.30        | 44.56        | 84.44        | 82.05        | 69.34        |
| 10     | 66.40        | 44.89        | 84.39        | 82.51        | 69.55        |
| 20     | 67.30        | 46.78        | 85.17        | 82.87        | 70.53        |
| 30     | <b>67.30</b> | <b>47.22</b> | <b>85.38</b> | <b>83.02</b> | <b>70.73</b> |
| 50     | 64.70        | 45.78        | 85.19        | 81.92        | 69.40        |

Table 9: Effect of reconstruction epochs on MCTQ under the W4A6 setting.

tion set sizes. Performance remains stable across a wide range (16–512 samples), with the mean varying within 0.9 points. Notably, even 16 calibration samples (69.97) perform comparably to 512 samples (70.21), indicating that MCTQ does not rely on a large calibration set and that coordination effectively exploits the available multimodal signal.

### C Additional layer-level conflict visualizations.

Figure 5 gives a compact layer-aggregated view of multimodal conflict. Some layers show stronger directional incompatibility, while others are dominated by large gradient-magnitude imbalance, indicating that the conflict is structured rather than random.

| Samples | OCR   | MMMU         | SQA          | TVQA         | Avg.         |
|---------|-------|--------------|--------------|--------------|--------------|
| 16      | 66.40 | 46.11        | 84.72        | 82.65        | 69.97        |
| 32      | 65.50 | 45.67        | 83.92        | 82.17        | 69.31        |
| 64      | 66.20 | 44.89        | 84.89        | 82.35        | 69.58        |
| 128     | 66.10 | 46.67        | 84.91        | 82.43        | 70.03        |
| 256     | 66.20 | 45.00        | 85.62        | <b>82.65</b> | 69.87        |
| 512     | 65.60 | <b>47.44</b> | <b>85.64</b> | 82.17        | <b>70.21</b> |

Table 10: Effect of calibration set size on MCTQ under the W4A6 setting. All experiments use 30 reconstruction epochs.

Figure 6 further shows the top-12 layers ranked by the aggregated conflict score. These examples complement the representative plots in the main text by showing that the reported conflict patterns are part of a broader layer-level structure.

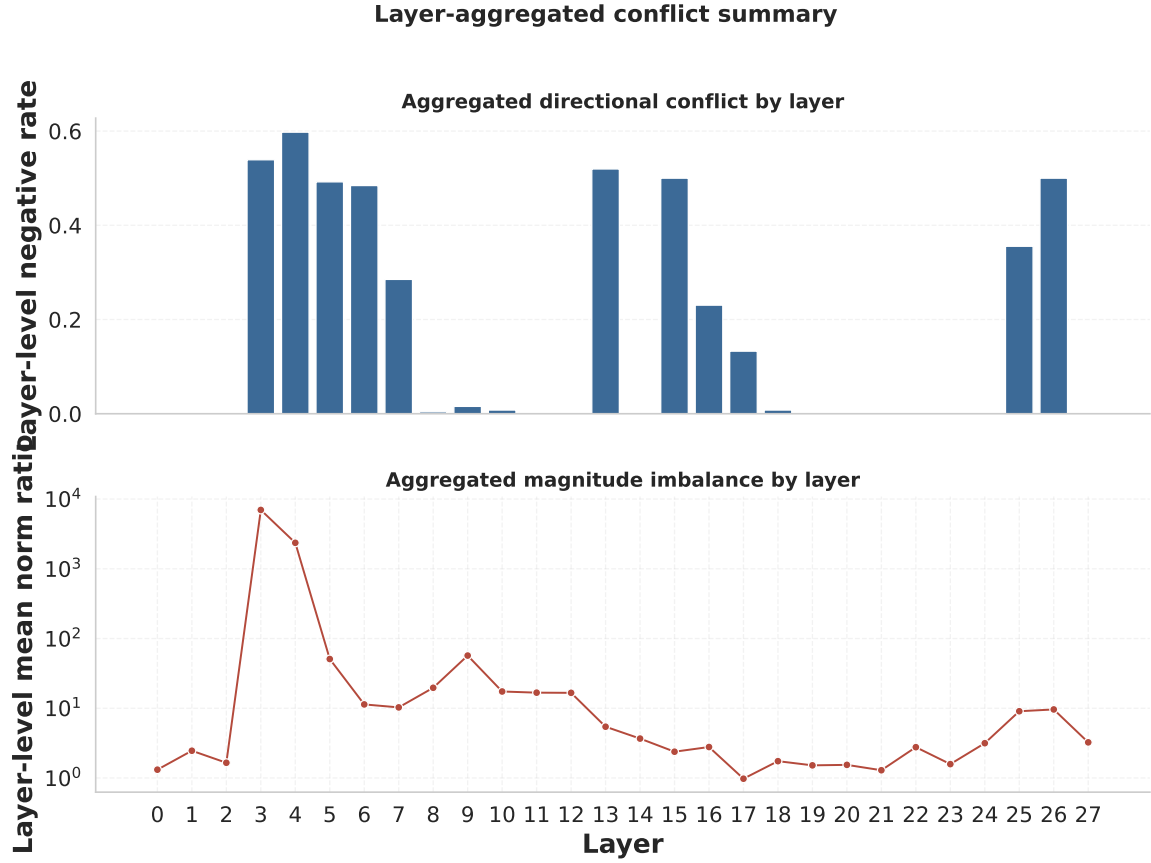


Figure 5: Layer-aggregated modality-gradient conflict statistics. The top panel shows the negative-alignment rate of text and vision gradients for each layer, and the bottom panel shows the corresponding mean gradient norm ratio.

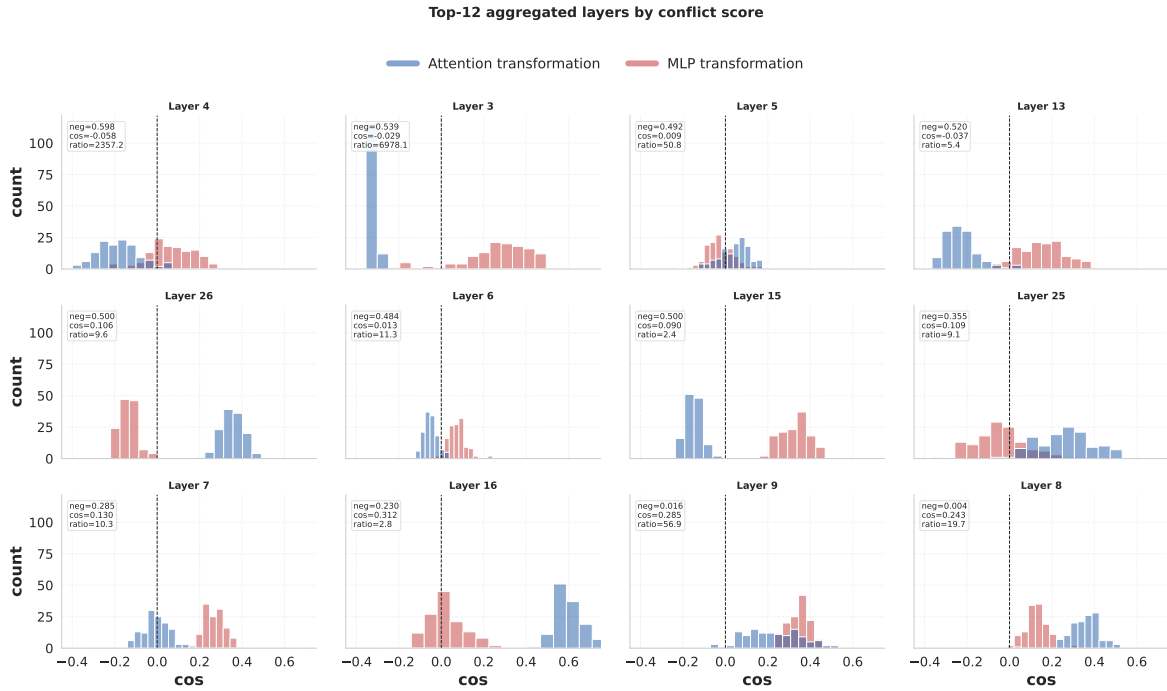


Figure 6: Top-12 aggregated layers ranked by conflict score. Each panel overlays the cosine-similarity distributions of modality gradients within one layer and reports the corresponding conflict statistics.

## D Theoretical Justification of Modality-Coordinated Optimization

**Setup.** We provide a local first-order justification for the modality-coordinated optimization used in MCTQ. Let  $\theta = \text{vec}(\mathbf{S})$  denote the vectorized shared transformation parameters. For modality  $m \in \{t, v\}$ , where  $t$  and  $v$  denote text and vision, respectively, let  $\mathcal{L}_m(\theta)$  denote the modality-specific reconstruction loss defined in Eq. (7), and let  $\mathbf{g}_m = \nabla_{\theta} \mathcal{L}_m(\theta)$  be its gradient with respect to the shared transformation. The vectorization is only for notational convenience; the Euclidean inner product between vectorized gradients is equivalent to the Frobenius inner product between the corresponding matrix gradients.

Given the two modality gradients  $\mathbf{g}_t$  and  $\mathbf{g}_v$ , we consider their convex hull:

$$\mathcal{C} = \text{conv}(\mathbf{g}_t, \mathbf{g}_v) = \{w\mathbf{g}_t + (1-w)\mathbf{g}_v \mid w \in [0, 1]\}. \quad (16)$$

MCTQ selects the coordinated gradient as the minimum-norm element in this convex hull:

$$\mathbf{g}_{\text{coord}} = \arg \min_{\mathbf{g} \in \mathcal{C}} \|\mathbf{g}\|_2^2. \quad (17)$$

Equivalently,

$$\mathbf{g}_{\text{coord}} = w^* \mathbf{g}_t + (1 - w^*) \mathbf{g}_v, \quad (18)$$

where  $w^* \in [0, 1]$  is obtained by solving the one-dimensional quadratic problem in Eq. (12). The following results show that this choice yields a principled shared update direction for multimodal transformation learning. Since quantization-aware reconstruction is optimized with straight-through gradient estimation in practice, the following analysis concerns the local first-order behavior of the surrogate reconstruction objectives with respect to the learnable transformation parameters. The results do not claim global optimality or direct downstream accuracy guarantees; rather, they justify why the coordinated direction is preferable to a fixed scalarization when modality gradients conflict.

**Lemma 1. Closed-form coordination weight.** For two modality gradients  $\mathbf{g}_t$  and  $\mathbf{g}_v$ , the solution of

$$\min_{w \in [0, 1]} \|w\mathbf{g}_t + (1-w)\mathbf{g}_v\|_2^2 \quad (19)$$

is given by

$$w^* = \Pi_{[0, 1]} \left( \frac{\|\mathbf{g}_v\|_2^2 - \langle \mathbf{g}_t, \mathbf{g}_v \rangle}{\|\mathbf{g}_t - \mathbf{g}_v\|_2^2} \right), \quad (20)$$

where  $\Pi_{[0, 1]}(\cdot)$  denotes clipping to the interval  $[0, 1]$ . Accordingly, the coordinated gradient is

$$\mathbf{g}_{\text{coord}} = w^* \mathbf{g}_t + (1 - w^*) \mathbf{g}_v. \quad (21)$$

*Proof.* Let

$$f(w) = \|w\mathbf{g}_t + (1-w)\mathbf{g}_v\|_2^2. \quad (22)$$

This objective can be rewritten as

$$f(w) = \|\mathbf{g}_v + w(\mathbf{g}_t - \mathbf{g}_v)\|_2^2. \quad (23)$$

Expanding the quadratic form gives

$$f(w) = \|\mathbf{g}_v\|_2^2 + 2w\langle \mathbf{g}_v, \mathbf{g}_t - \mathbf{g}_v \rangle + w^2\|\mathbf{g}_t - \mathbf{g}_v\|_2^2. \quad (24)$$

Taking the derivative with respect to  $w$ , we obtain

$$f'(w) = 2\langle \mathbf{g}_v, \mathbf{g}_t - \mathbf{g}_v \rangle + 2w\|\mathbf{g}_t - \mathbf{g}_v\|_2^2. \quad (25)$$

Setting  $f'(w) = 0$  yields the unconstrained minimizer

$$w_0 = \frac{-\langle \mathbf{g}_v, \mathbf{g}_t - \mathbf{g}_v \rangle}{\|\mathbf{g}_t - \mathbf{g}_v\|_2^2} = \frac{\|\mathbf{g}_v\|_2^2 - \langle \mathbf{g}_t, \mathbf{g}_v \rangle}{\|\mathbf{g}_t - \mathbf{g}_v\|_2^2}. \quad (26)$$

Since Eq. (19) constrains  $w \in [0, 1]$ , the constrained optimum is obtained by projecting the unconstrained minimizer onto  $[0, 1]$ :

$$w^* = \Pi_{[0, 1]}(w_0). \quad (27)$$

This proves Eq. (20). When  $\mathbf{g}_t = \mathbf{g}_v$ , the denominator becomes zero and the objective is constant for all  $w \in [0, 1]$ ; in this degenerate case, we set  $w^* = 1/2$  in practice.  $\square$

**Proposition 1. Common descent property.** Let

$$\mathbf{g}_{\text{coord}} = \arg \min_{\mathbf{g} \in \mathcal{C}} \|\mathbf{g}\|_2^2, \quad \mathcal{C} = \text{conv}(\mathbf{g}_t, \mathbf{g}_v). \quad (28)$$

If  $\mathbf{g}_{\text{coord}} \neq \mathbf{0}$ , then  $-\mathbf{g}_{\text{coord}}$  is a common descent direction for both modality-specific objectives. Specifically, for  $m \in \{t, v\}$ , we have

$$\langle \mathbf{g}_m, -\mathbf{g}_{\text{coord}} \rangle \leq -\|\mathbf{g}_{\text{coord}}\|_2^2 < 0. \quad (29)$$

*Proof.* Since  $\mathbf{g}_{\text{coord}}$  is the minimum-norm element in the closed convex set  $\mathcal{C}$ , it is the Euclidean projection of the origin onto  $\mathcal{C}$ . By the first-order optimality condition of Euclidean projection, for any  $\mathbf{g} \in \mathcal{C}$ ,

$$\langle \mathbf{g} - \mathbf{g}_{\text{coord}}, \mathbf{g}_{\text{coord}} - \mathbf{0} \rangle \geq 0. \quad (30)$$

Equivalently,

$$\langle \mathbf{g}, \mathbf{g}_{\text{coord}} \rangle \geq \|\mathbf{g}_{\text{coord}}\|_2^2, \quad \forall \mathbf{g} \in \mathcal{C}. \quad (31)$$

Because both modality gradients  $\mathbf{g}_t$  and  $\mathbf{g}_v$  belong to  $\mathcal{C}$ , taking  $\mathbf{g} = \mathbf{g}_m$  gives

$$\langle \mathbf{g}_m, \mathbf{g}_{\text{coord}} \rangle \geq \|\mathbf{g}_{\text{coord}}\|_2^2, \quad m \in \{t, v\}. \quad (32)$$

Multiplying both sides by  $-1$ , we obtain

$$\langle \mathbf{g}_m, -\mathbf{g}_{\text{coord}} \rangle \leq -\|\mathbf{g}_{\text{coord}}\|_2^2. \quad (33)$$

If  $\mathbf{g}_{\text{coord}} \neq \mathbf{0}$ , the right-hand side is strictly negative. Therefore,  $-\mathbf{g}_{\text{coord}}$  is a descent direction for both  $\mathcal{L}_t$  and  $\mathcal{L}_v$  in the first-order sense.  $\square$

This proposition shows that the coordinated gradient is not merely a smaller update. Instead, when it is nonzero, it defines a shared direction that is locally compatible with both modality-specific reconstruction objectives. Thus, the update avoids the first-order failure case in which optimizing the shared transformation for one modality directly increases the reconstruction loss of the other. When  $\mathbf{g}_{\text{coord}} = \mathbf{0}$ , the origin lies in the convex hull of the modality gradients, indicating a Pareto-stationary point for the two local objectives.

**Corollary 1. Loss decrease under smoothness.** Assume that each modality-specific reconstruction loss  $\mathcal{L}_m$ ,  $m \in \{t, v\}$ , is  $\beta_m$ -smooth with respect to  $\theta = \text{vec}(\mathbf{S})$ . Let

$$\theta^+ = \theta - \eta \mathbf{g}_{\text{coord}}, \quad (34)$$

where  $\mathbf{g}_{\text{coord}} \neq \mathbf{0}$  is the coordinated gradient defined in Eq. (17). If  $0 < \eta < 2/\max_m \beta_m$ , then for each modality  $m \in \{t, v\}$ ,

$$\mathcal{L}_m(\theta^+) \leq \mathcal{L}_m(\theta) - \eta \left(1 - \frac{\beta_m \eta}{2}\right) \|\mathbf{g}_{\text{coord}}\|_2^2. \quad (35)$$

Therefore, the coordinated update decreases both modality-specific reconstruction losses under a sufficiently small step size.

*Proof.* By the  $\beta_m$ -smoothness of  $\mathcal{L}_m$ , we have

$$\begin{aligned} \mathcal{L}_m(\theta - \eta \mathbf{g}_{\text{coord}}) &\leq \mathcal{L}_m(\theta) - \eta \langle \mathbf{g}_m, \mathbf{g}_{\text{coord}} \rangle \\ &\quad + \frac{\beta_m \eta^2}{2} \|\mathbf{g}_{\text{coord}}\|_2^2. \end{aligned} \quad (36)$$

From Proposition 1, the coordinated gradient satisfies

$$\langle \mathbf{g}_m, \mathbf{g}_{\text{coord}} \rangle \geq \|\mathbf{g}_{\text{coord}}\|_2^2, \quad m \in \{t, v\}. \quad (37)$$

Substituting this inequality into the smoothness bound gives

$$\begin{aligned} \mathcal{L}_m(\theta - \eta \mathbf{g}_{\text{coord}}) &\leq \mathcal{L}_m(\theta) - \eta \|\mathbf{g}_{\text{coord}}\|_2^2 \\ &\quad + \frac{\beta_m \eta^2}{2} \|\mathbf{g}_{\text{coord}}\|_2^2. \end{aligned} \quad (38)$$

Rearranging terms yields

$$\mathcal{L}_m(\theta^+) \leq \mathcal{L}_m(\theta) - \eta \left(1 - \frac{\beta_m \eta}{2}\right) \|\mathbf{g}_{\text{coord}}\|_2^2. \quad (39)$$

When  $0 < \eta < 2/\max_m \beta_m$ , the coefficient  $1 - \beta_m \eta/2$  is positive for both modalities. Thus, the update decreases both  $\mathcal{L}_t$  and  $\mathcal{L}_v$ .  $\square$

**Interpretation.** The above results clarify why the minimum-norm coordination is suitable for modality-coordinated transformation learning. First, Proposition 1 shows that when  $\mathbf{g}_{\text{coord}} \neq \mathbf{0}$ , the update direction  $-\mathbf{g}_{\text{coord}}$  is a common descent direction for both modality-specific reconstruction losses. Therefore, unlike a fixed mixed update, the coordinated update avoids the first-order failure case where reducing the reconstruction error of one modality directly increases that of the other.

Second, the closed-form solution in Lemma 1 shows that the coordination weight depends on both the gradient norms and their inner product. Thus, the update is not determined by token composition or fixed loss weights, but adapts to the geometry of the modality gradients. For example, when  $\mathbf{g}_t \perp \mathbf{g}_v$ , the optimal weight becomes

$$w^* = \frac{\|\mathbf{g}_v\|_2^2}{\|\mathbf{g}_t\|_2^2 + \|\mathbf{g}_v\|_2^2}, \quad (40)$$

so the modality with a larger gradient norm receives a smaller coefficient. This behavior helps mitigate modality-dependent gradient dominance, although it should be interpreted as an adaptive balancing effect rather than a complete theoretical resolution of all magnitude-imbalance phenomena.

Finally, we emphasize that this guarantee is local and first-order. It does not imply global optimality, nor does it by itself guarantee downstream task accuracy improvement. Rather, it justifies MCTQ's optimization principle: once modality-specific gradients are exposed, coordinating them in gradient space provides a principled shared update that simultaneously addresses directional conflict and adaptively mitigates gradient dominance.

### Example 1. Workstation Scene (COCO 1503)

**Question:** Please provide a detailed description of this image.



**FP16 Response:** The image depicts a well-organized computer workstation setup on a white desk. The centerpiece is an iMac G5 desktop computer with its distinctive rounded design and silver base, displaying a webpage on its screen. To the left of the iMac sits a silver MacBook laptop, partially open, suggesting recent use. A small external hard drive is positioned between the two devices, indicating data transfer or backup activities. In front of the iMac, a white Apple keyboard and a matching Apple Magic Mouse rest on the desk, ready for use. Two small, round, silver speakers flank the keyboard, enhancing the audio experience. The desk is situated near a window with partially closed blinds, allowing natural light to filter into the room. A floor lamp stands in the background, providing additional lighting. The overall setting suggests a clean, modern workspace, likely used for work or personal computing tasks.

**MCTQ (W4A6) Response:** The image showcases a well-organized computer workstation on a white desk. The central focus is an iMac G4 desktop computer with a distinctive round base and a screen displaying what appears to be a web page or software interface. To the left of the iMac, there is a silver Apple iBook G4 laptop, suggesting a dual-computer setup for work or multitasking. Below the iBook, a small external hard drive is placed, likely used for data storage or backup purposes. To the right of the iMac, a printer or scanner is visible, indicating a home office or personal workspace setup. The desk also features two small, circular speakers, one on each side of the iMac, which suggest that the user might have enjoyed listening to music or other audio while working. The background includes a window with blinds partially open, allowing natural light to enter the room. A floor lamp stands in the corner, providing additional lighting when needed.

**MBQ (W4A6) Response:** The image depicts a well-organized workspace featuring multiple Apple-branded devices arranged neatly on a desk beneath a window with horizontal blinds. Prominently positioned is an iMac G5 desktop computer displaying its interface screen atop sleek silver speakers symmetrically placed flanking it. Adjacent to this setup sits an open silver MacBook Pro laptop oriented horizontally across its docking station. Beneath these electronics rests a compact external hard drive, suggesting functionality for expanded storage capabilities. Completing the peripherals arrayed before us are two circular white Apple products, possibly small accessories, resting on the desk surface. A centrally placed Apple keyboard and mouse further reinforce the tidy workstation layout.

## Example 2. Outdoor Dog Scene (COCO 366611)

**Question:** Please provide a detailed description of this image.



---

**FP16 Response:** The image shows a brindle and white dog lying on a patch of grass, actively chewing on a red and white soccer ball. The dog appears to be engaged in play, with its mouth gripping the ball firmly. The grassy area is uneven, with some patches of bare soil visible, suggesting recent digging or play activity. To the right of the dog, there is a gravel path that runs parallel to the grassy area, indicating a boundary between the lawn and another section of the yard. The lighting suggests it is daytime, with sunlight casting shadows across the scene, enhancing the textures of the grass and gravel. The overall setting seems to be a backyard or a similar outdoor space where the dog is enjoying some free time.

---

**MCTQ (W4A6) Response:** The image shows a brindle and white dog lying on a patch of grass, playfully interacting with a red and white soccer ball. The dog appears to be biting or nuzzling the ball, suggesting it is engaged in play. The grassy area where the dog and ball are situated is bordered by a gravel path on one side, which seems to lead into a more open, sunlit area. The lighting suggests it is daytime, likely late afternoon given the long shadows cast. The overall setting appears to be a backyard or a similar outdoor space, possibly used for pets to play and explore. The dog's posture and the relaxed environment suggest a casual, leisurely moment captured during a warm day.

---

**MBQ (W4A6) Response:** The provided image captures a tranquil outdoor scene featuring a brindle-coated dog lying on its side on a patch of unevenly trimmed grass. Adjacent to it is a reddish-brown soccer ball positioned centrally between the dog's front paws. The spatial arrangement places the majority of the frame within a naturalized yard bordered by patches of bare soil and gravel toward the upper-right quadrant. Sunlight filters down from above, casting distinct shadows across the grassy expanse beneath the dog. The ground is punctuated by small yellow-green patches, suggesting a slightly worn outdoor play area.

### Example 3. Multi-Person Indoor Scene (COCO 9400)

**Question:** Please provide a detailed description of this image.



**FP16 Response:** The image depicts a group of people gathered around a table in what appears to be a casual, possibly informal setting, such as a cafe or a community center. The individuals are focused on their laptops, suggesting they might be working, studying, or participating in an online activity together. The table is cluttered with various items including laptops, cups, and condiments like ketchup, indicating a relaxed and collaborative atmosphere. The background features a checkered wall and some hanging lights, adding to the cozy ambiance. The participants appear engaged and concentrated, with one person leaning forward intently, while others look at their screens or converse softly. The overall scene conveys a sense of shared purpose and interaction among the group.

**MCTQ (W4A6) Response:** The image depicts a group of people gathered around a table in what appears to be a casual, possibly restaurant-like setting with a checkered wall and a bar area in the background. The focus is on several individuals who are intently working on their laptops, suggesting a collaborative or competitive environment. The laptops vary in brand and model, with one prominently displaying an Apple logo. The table is cluttered with various items including cups, bottles (likely containing beverages), and other personal belongings, indicating a relaxed yet focused atmosphere. The lighting is dim, typical of an evening setting, which adds to the ambiance of a late-night work session or gathering. The individuals seem engaged, with some leaning forward towards their screens, while others appear to be in conversation or contemplation. The overall scene suggests a blend of work and social interaction in a relaxed environment.

**MBQ (W4A6) Response:** The setting appears to be a dimly lit cafe or restaurant-like environment where individuals are deeply engrossed in their laptops. Multiple Apple iMacs dominate the foreground, suggesting that this could be either a tech meet-up or informal coding session among enthusiasts.