JW-SVD: Bridging the Cross-Modal Mismatch in Post-Training MLLM Compression

Runchao Li **Yao Fu** **Mu Sheng**
Xianxuan Long **Haotian Yu** **Kenneth Loparo**
Case Western Reserve University
{rxl685, yxf484, mxs2090, xxl1514, hxy692, kal4}@case.edu

Abstract

Post-training compression of Multimodal LLMs faces a fundamental geometric conflict: parameter subspaces optimized for text often suppress orthogonal visual features. We demonstrate that standard SVD fails to resolve this *cross-modal mismatch*, causing catastrophic visual degradation. To bridge this gap, we introduce **Joint-Whitening SVD (JW-SVD)**, a dual-objective framework that aligns vision and language manifolds via a *Joint Covariance* basis, preserving features critical to both. Additionally, we propose *Global Spectrum-Aware Truncation* to dynamically transfer parameter budget from the redundant Vision Tower to the sensitive Backbone. Experiments on Qwen2.5-VL and Llama-3-Next confirm that JW-SVD demonstrates superior retention of both text and image capabilities. In addition, it resolves the modality trade-off: it recovers over 30% of perceptual performance lost by baselines while maintaining parity in textual reasoning, enabling robust multimodal performance even at extreme compression rates.

1 Introduction

Multimodal Large Language Models (MLLMs) (Liu et al., 2025; Alayrac et al., 2022; Bai et al., 2023; Huang et al., 2020; Li et al., 2025c) have advanced artificial intelligence by unifying visual perception with linguistic reasoning. However, as these models scale, their massive parameter counts impede deployment on resource-constrained edge devices. While established pruning (Ma et al., 2023; Ashkboos et al., 2024; Sreenivas et al., 2024; Long et al., 2025; Fu et al., 2025b) and quantization (Lin et al., 2024; Huang et al., 2024; van Baalen et al., 2025; Fu et al., 2025c) techniques facilitate compression, they often lack the multimodal adaptation necessary to preserve cross-modal synergy.

Singular Value Decomposition (SVD) has emerged as a promising post-training compression tool (Wang et al., 2025c; Yuan et al., 2025; Wang et al., 2025a); however, standard implementations suffer from a fundamental *cross-modal mismatch*. Through theoretical and empirical analyses, we observe that optimizing weight matrices solely for text activations leads to staggering reconstruction errors on visual data. Our layer-wise analysis reveals that this misalignment is particularly catastrophic in early layers, where image information loss can exceed text loss by nearly 700×. This phenomenon indicates that low-rank directions deemed redundant for text are often critical for visual grounding, leading to a "visual collapse" where the model's perceptual capability is severely impaired despite maintaining its linguistic proficiency.

Figure 1: Our cross-modal JW-SVD mitigated the inherent loss of text-only SVD performed on MLLM.

To address these challenges, we propose **Joint-Whitening SVD (JW-SVD)**, a post-training compression framework tailored for MLLMs. As shown in Figure 1, we reformulate the compression objective as a dual-objective minimization problem, integrating text and image statistics into a unified *Joint Covariance* matrix. This approach ensures the compressed basis preserves features essential to both modalities, effectively preventing visual collapse by maintaining alignment across disparate

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 42691–42702

manifolds. Furthermore, we exploit the inherent structural redundancy of the vision tower—which typically exhibits faster spectral decay than the backbone—to dynamically reallocate the parameter budget. By treating the vision tower as a "parameter donor," we protext the sensitive multimodal backbone from over-compression, preserving its capacity for complex cross-modal reasoning.

Our primary contributions are summarized as follows:

- We propose a rigorous framework to quantify the *cross-modal misalignment* in SVD compression, theoretically proving that text-optimized bases inherently suppress visual information.
- We introduce **Joint-Whitening SVD (JW-SVD)**, a dual-objective compression method coupled with *Global Spectrum-Aware Truncation*, which dynamically reallocates redundancy from the vision tower to the multimodal backbone.
- Extensive experiments show that JW-SVD consistently outperforms state-of-the-art baselines, recovering over 30% of perceptual capability lost by text-only methods at a 40% budget. Furthermore, we achieve effective compression of 1.8 bits per parameter when combined with quantization, establishing a new Pareto frontier for efficient MLLMs.

2 Related Work

Efficient Multimodal LLMs. The convergence of Large Language Models with visual encoders has yielded powerful MLLMs capable of complex reasoning (Liu et al., 2025; Alayrac et al., 2022; Bai et al., 2023; Huang et al., 2020; Zhang et al., 2025; Jiang et al., 2025; Jiang and Ferraro, 2026). However, the immense computational cost of these models limits their deployment on edge devices. While general model compression techniques such as quantization (Lin et al., 2024; Huang et al., 2024; van Baalen et al., 2025; Li et al., 2025a) and structured pruning (Ma et al., 2023; Ashkboos et al., 2024; Sreenivas et al., 2024; Xu et al., 2025) have been widely adopted, they often require extensive calibration or fine-tuning to recover performance. Low-rank decomposition offers a compelling alternative by mathematically approximating weight matrices without retraining, yet its application to the multimodal domain remains underexplored.

Post-Training Decomposition and Spectral Analysis. Singular Value Decomposition (SVD) has emerged as a robust method for compressing LLMs. Recent state-of-the-art approaches, such as SVD-LLM (Wang et al., 2025c) and ASVD (Yuan et al., 2025), improve upon standard truncation by incorporating activation statistics to preserve outliers critical for language modeling. Despite their success in NLP, these methods fundamentally rely on a unimodal assumption: they optimize the reconstruction error based solely on textual activation covariance. We identify this as a critical limitation; visual features, which often manifest as high-frequency components orthogonal to text manifolds, are inadvertently treated as noise and pruned. Unlike prior works that isolate modalities, our framework explicitly addresses this *cross-modal spectral mismatch* by enforcing a joint geometric alignment.

3 Modality Mismatch analysis

3.1 Preliminary: Text-based SVD compression on LLM

Post-training compression aims to reduce a pre-trained weight matrix $\mathbf{W} \in \mathbb{R}^{d_{out} \times d_{in}}$ to a low-rank approximation $\hat{\mathbf{W}}$ while minimizing output degradation. Unlike standard matrix approximation, which minimizes the Frobenius norm of the weight difference $\|\mathbf{W} - \hat{\mathbf{W}}\|_F$, effective compression for Large Language Models (LLMs) must account for the non-uniform distribution of input activations. Consequently, the optimization objective is defined as minimizing the reconstruction error of the activations $\mathbf{X}$:

$$\min_{\text{rank}(\hat{\mathbf{W}}) \leq k} \mathcal{L} = \|(\mathbf{W} - \hat{\mathbf{W}})\mathbf{X}\|_F^2 \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{d_{in} \times N}$ represents the calibration inputs, and k represent the target compression rank. Direct Singular Value Decomposition (SVD) on $\mathbf{W}$ fails to optimize Eq. 1 because it treats all input dimensions as equally important, ignoring the fact that activation channels in LLMs often exhibit extreme outliers and varying variances.

To align spectral truncation with output loss, prior works (Wang et al., 2025c; Chen et al., 2021) utilize *Truncation-Aware Data Whitening*. Let $\mathbf{H} = \mathbf{X}\mathbf{X}^T = \mathbf{S}\mathbf{S}^T$ denote the input covariance decomposed via Cholesky. The objective simplifies to:

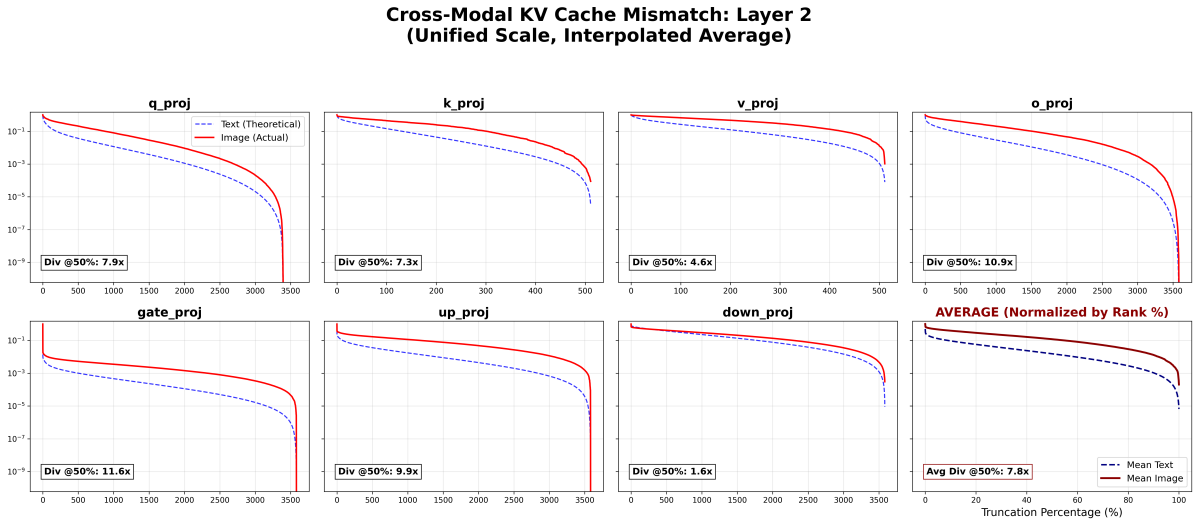

Figure 2: **Local Divergence (Layer 2).** Reconstruction error curves for linear modules in Layer 2. The blue dashed line represents theoretical text loss, while the red solid line tracks actual image reconstruction error. Note the "heavy tail" in the image loss, particularly in the MLP layers (e.g., `gate_proj`), where the divergence at 50% sparsity reaches $11.6\times$ compared to the text baseline.

$$\begin{aligned}\mathcal{L} &= \|\Delta \mathbf{S}^{-1}\mathbf{X}\|_F^2 = \text{Trace}(\Delta \underbrace{\mathbf{S}^{-1}\mathbf{X}\mathbf{X}^T\mathbf{S}^{-T}}_{\mathbf{I}}\Delta^T) \\ &= \|\Delta\|_F^2 = \sum_{i\in\text{trunc}} \sigma_i^2\end{aligned} \tag{2}$$

where σ_i are the singular values of $\mathbf{WS}$ and Δ is the truncation error. Since the whitened input space is orthonormal , the Eckart-Young-Mirsky theorem guarantees that truncating the smallest σ_i minimizes the reconstruction error on $\mathbf{X}$.

3.2 The Cross-Modal Alignment Gap

While SVD-LLM ensures optimality for the calibration modality, we demonstrate that this guarantee fails for multimodal data due to geometric misalignment.

Derivation of the Mismatch. Let $\mathbf{S}_T$ and $\mathbf{S}_I$ denote the whitening matrices for text and image activations, respectively (where $\mathbf{S}\mathbf{S}^T = \mathbf{X}\mathbf{X}^T$). When a text-optimized basis with truncation error Δ is applied to image data $\mathbf{X}_I$, the reconstruction loss $\mathcal{L}_I$ becomes:

$$\mathcal{L}_I^2 = \|(\mathbf{W} - \hat{\mathbf{W}})\mathbf{X}_I\|_F^2 = \|\Delta \mathbf{S}_T^{-1}\mathbf{X}_I\|_F^2 \tag{3}$$

Expanding the Frobenius norm via the trace property and substituting $\mathbf{X}_I\mathbf{X}_I^T = \mathbf{S}_I\mathbf{S}_I^T$ reveals the *Mismatch Matrix* $\mathbf{M} \triangleq \mathbf{S}_T^{-1}\mathbf{S}_I$:

$$\begin{aligned}\mathcal{L}_I^2 &= \text{Trace}\left(\Delta \mathbf{S}_T^{-1}(\mathbf{S}_I\mathbf{S}_I^T)\mathbf{S}_T^{-T}\Delta^T\right) \\ &= \text{Trace}\left(\Delta\mathbf{M}\mathbf{M}^T\Delta^T\right)\end{aligned} \tag{4}$$

Here, $\mathbf{M}$ quantifies the geometric deviation between the text-whitened and image-whitened subspaces. Unlike the ideal alignment scenario where $\mathbf{M} = \mathbf{I}$, our analysis shows $\mathbf{M}$ acts as a significant scaling factor. Substituting the SVD components of Δ yields the final cross-modal loss expression:

$$\mathcal{L}_I^2 = \sum_{i\in\mathcal{K}_{trunc}} \sigma_i^2 \cdot \underbrace{\|\mathbf{v}_i^T\mathbf{M}\|_2^2}_{\lambda_i} \tag{5}$$

Interpretation. Eq. 5 exposes the mechanism of cross-modal failure: the total error is the text importance (σ_i^2) weighted by the *Mismatch Factor* λ_i.

- **Scenario A (Text-on-Text):** Evaluated on text, $\mathbf{S}_I = \mathbf{S}_T \implies \mathbf{M} = \mathbf{I}$ and $\lambda_i = 1$. The loss simplifies to $\sum \sigma_i^2$, recovering the SVD-LLM optimality guarantee.

- **Scenario B (Text-on-Image):** If a direction $\mathbf{v}_i$ is textual noise (small σ_i) but aligns with the image subspace, λ_i becomes large. This amplification sustains high $\mathcal{L}_I$ even as $\mathcal{L}_T \to 0$, driving the visual collapse observed in experiments.

3.3 Empirical Verification: The Modal Gap in Practice

To quantify the theoretical mismatch derived in Section 3.2, we perform a layer-wise analysis on Qwen2.5-VL-7B. We measure the *Cumulative Reconstruction Error (CRE)* under text-based truncation, contrasting the theoretical text loss ($\mathcal{L}_T = \sum \sigma_i^2$) with the empirical image loss ($\mathcal{L}_I = \sum \sigma_i^2 \lambda_i$).

Local Divergence: The "Heavy Tail" Phenomenon. We examine modular divergence in Layer 2 (Figure 2), where two critical behaviors validate our model:

- **The Heavy Tail:** Consistent with Eq. 5, the image reconstruction error (red line) exhibits a significant "heavy tail." loss for image While $\mathcal{L}_T$ decays exponentially—indicating σ_i effectively captures text variance—$\mathcal{L}_I$ remains orders of magnitude higher in the high-rank region. This confirms that for indices $i > k$, the mismatch factor λ_i is large: directions discarded as textual noise act as critical signals for visual grounding.
- **Module Sensitivity:** Misalignment is highly module-dependent. Notably, MLP expansion layers (e.g., `gate_proj`, `up_proj`) exhibit severe divergence compared to projection layers. For instance, `gate_proj` suffers an $11.6\times$ divergence at 50% sparsity, whereas the output mapping `down_proj` shows only $1.6\times$. This suggests that modality-specific features are sequestered in the high-dimensional intermediate subspaces of feed-forward networks, orthogonal to the text manifold.

Global Analysis: Systemic Misalignment. To map the distribution of this error, we compute the Cumulative Error Ratio $\rho = \sum \mathcal{L}_I / \mathcal{L}_T$ across the full depth (Figure 3):

- **Early-Layer Disjointness:** Misalignment is catastrophic in Layers 0–2, where image loss exceeds text estimates by up to $697.5\times$. This confirms that early layers process raw, unintegrated modality features in nearly orthogonal manifolds, rendering them hypersensitive to text-based truncation.
- **The MLP Gap:** While the gap narrows as modalities integrate in deeper layers, expansion layers in MLP modules (orange/red lines) consistently sustain higher error ratios ($4 \times -8\times$) than other blocks. This indicates that feed-forward networks serve as the primary provider for modality-specific information throughout the depth of the model.

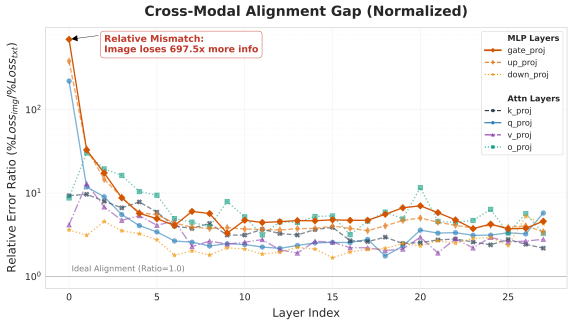

Figure 3: **Global Alignment Gap.** The Cumulative Error Ratio $\rho = \mathcal{L}_I / \mathcal{L}_T$ across all model layers. Early layers (0–2) exhibit catastrophic misalignment, with image information loss exceeding text loss by up to $697.5\times$. While the gap narrows in deeper layers, most modules consistently maintain higher mismatch ratios ($4 \times -8\times$) than attention heads.

Implication for Compression. These findings demonstrate that uniform truncation is fundamentally flawed for MLLMs. The image mismatch in early layers and MLPs necessitate a *Dynamic Truncation* strategy that specifically preserves these high-mismatch subspaces.

4 Methodology: Joint-Whitening SVD

To bridge the modal gap identified in Section 3.2, we propose **Joint-Whitening SVD (JW-SVD)**, a post-training compression framework designed to minimize cross-modal reconstruction error. Our method comprises two core components: *Joint Covariance Whitening*, which integrates multimodal statistics into the optimization objective, and *Global Spectrum-Aware Truncation*, which dynamically allocates the compression budget across the vision tower and language backbone to prevent feature collapse.

4.1 Joint Covariance Whitening

4.1.1 Problem Formulation: Dual-Objective Minimization

Consider a linear layer $\mathbf{W} \in \mathbb{R}^{d_{out} \times d_{in}}$ with calibration activations $\mathbf{X}_T$ (text) and $\mathbf{X}_I$ (image). Standard SVD minimizes reconstruction error solely on text, implicitly suppressing orthogonal visual features (as shown in Sec. 3.3). To prevent this,

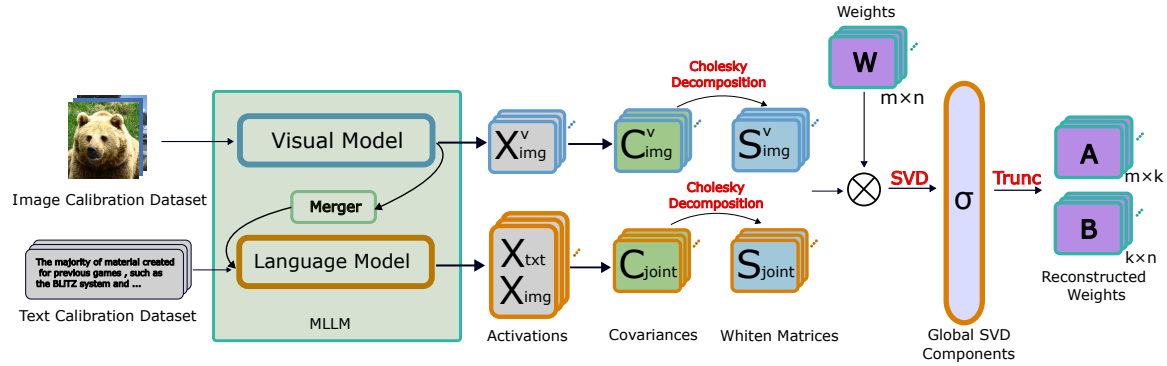

Figure 4: **Overview of the Joint-Whitening SVD (JW-SVD) Framework.** (1) We collect calibration activations from both the Vision Tower (X^v_{img}) and Language Backbone (X_{txt}, X_{img}). (2) *Joint Covariance Whitening* is applied to integrate multimodal statistics into a unified geometric basis via Cholesky decomposition. (3) We perform SVD on the whitened weights and aggregate singular values σ from all layers into a unified pool. (4) **Global Spectrum-Aware Truncation** dynamically determines layer-specific ranks, automatically transferring parameter budget from the redundant Vision Tower to the sensitive Multimodal Backbone before reconstructing low-rank factors $\mathbf{A}$ and $\mathbf{B}$.

we reformulate the compression task as a *Dual-Objective Minimization* problem, seeking a low-rank basis $\hat{\mathbf{W}}$ that minimizes the weighted error for both modalities:

$$\min_{\text{rank}(\hat{\mathbf{W}})\leq k} \mathcal{L}_{joint} = \|\Delta\mathbf{W}\mathbf{X}_T\|_F^2 + \alpha\|\Delta\mathbf{W}\mathbf{X}_I\|_F^2 \tag{6}$$

where $\Delta\mathbf{W} = \mathbf{W} - \hat{\mathbf{W}}$ and α balances modal sensitivity. This formulation ensures that singular vectors are truncated only if redundant for *both* modalities. Directions with low text variance but high image variance (large λ_i) are protected by the second term, forcing the optimization to retain them.

4.1.2 The Joint Covariance

Direct optimization of Eq. 6 is non-trivial due to subspace misalignment. However, we leverage the property that a weighted sum of Frobenius norms is equivalent to a single norm over a concatenated dataset $\tilde{\mathbf{X}} = [\mathbf{X}_T, \sqrt{\alpha}\mathbf{X}_I]$. We define the resulting moment matrix as the *Joint Covariance* $\mathcal{C}_{joint}$:

$$\mathcal{C}_{joint} \triangleq \mathbf{X}_T\mathbf{X}_T^T + \alpha(\mathbf{X}_I\mathbf{X}_I^T) = \mathbf{S}_T\mathbf{S}_T^T + \alpha\mathbf{S}_I\mathbf{S}_I^T \tag{7}$$

This matrix explicitly integrates the covariance structures of both modalities. Crucially, the dense image term $\alpha\mathbf{S}_I\mathbf{S}_I^T$ acts as a spectral regularizer. Unlike text covariances, which are often rank-deficient due to activation sparsity, visual representations are dense, improving the condition number of $\mathcal{C}_{joint}$. Consequently, we avoid the Cholesky decomposition failures reported in prior work (Wang et al., 2025b) without requiring ad-hoc dampening.

Scale-Invariant Energy Balancing. A critical requirement for Eq. 6 is determining the balancing factor α. We observe that a static global α is mathematically flawed because activation energies $\|\mathbf{X}\|_F^2$ vary significantly across layers. As noted in prior work (Li et al., 2025b), $\|\mathbf{X}_T\|_F \gg \|\mathbf{X}_I\|_F$ in many architectures, causing the text modality to dominate the Joint Covariance. This renders the image regularization ineffective due to numerical scale.

To resolve this without manual tuning, we adopt a *Scale-Invariant Energy Balancing* strategy. We define $\alpha^{(l)}$ to equalize the spectral energy contributions of both modalities:

$$\alpha^{(l)} = \frac{\text{Tr}(\mathbf{X}_T^{(l)}(\mathbf{X}_T^{(l)})^T)}{\text{Tr}(\mathbf{X}_I^{(l)}(\mathbf{X}_I^{(l)})^T)} \tag{8}$$

where l denotes the layer index. Substituting Eq. 8 into the Joint Covariance definition ensures that both modalities contribute equally to the geometric structure, preventing numerical dominance by the text modality.

Joint Basis Construction. We compute the *Joint Whitening Matrix* $\mathbf{S}_{joint}$ via the Cholesky decomposition of the Joint Covariance:

$$\mathbf{S}_{joint}\mathbf{S}_{joint}^T = \mathcal{C}_{joint} \tag{9}$$

Algorithm 1 Joint-Whitening SVD (JW-SVD) for MLLM Backbone

Require: Pre-trained layer weights $W \in \mathbb{R}^{m \times n}$, Text activations $X_T \in \mathbb{R}^{n \times s_T}$, Image activations $X_I \in \mathbb{R}^{n \times s_I}$, Global threshold τ.

Ensure: Compressed low-rank matrices $A \in \mathbb{R}^{m \times k_l}$ and $B \in \mathbb{R}^{k_l \times n}$.

$C_T \leftarrow X_T X_T^T$ //Text covariance

$C_I \leftarrow X_I X_I^T$ //Image covariance

$\alpha \leftarrow \text{Tr}(C_T)/\text{Tr}(C_I)$ //Scale-invariant factor

$C_{joint} \leftarrow C_T + \alpha C_I$ //Joint covariance

$S \leftarrow \text{Cholesky}(C_{joint})$ //Cholesky decomposition ($C_{joint} = SS^T$)

$\widetilde{W} \leftarrow WS$ // Whiten the pre-trained weights

$U, \Sigma, V^T \leftarrow \text{SVD}(\widetilde{W})$ // Perform Singular Value Decomposition

$k_l \leftarrow \sum_i \mathbb{I}(\tilde{\sigma}_i \geq \tau)$ // Determine layer rank dynamically via global threshold τ

$U_{k_l}, \Sigma_{k_l}, V_{k_l}^T \leftarrow \text{Truncate}(U, \Sigma, V^T, k_l)$ //Truncate singular components to rank k_l

$A \leftarrow U_{k_l} \Sigma_{k_l}^{1/2}$ //Form left low-rank matrix via symmetric spectral distribution

$B \leftarrow \Sigma_{k_l}^{1/2} V_{k_l}^T S^{-1}$ //Form right matrix and implicitly un-whiten

return A, B

Using $\mathbf{S}_{joint}$, we transform the original weights into the joint-whitened coordinate system $\tilde{\mathbf{W}} = \mathbf{W}\mathbf{S}_{joint}$ and perform Singular Value Decomposition:

$$\tilde{\mathbf{W}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{10}$$

where $\mathbf{\Sigma} = \text{diag}(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_{d_{in}})$.

Resolution of the Mismatch. The efficacy of JW-SVD stems from the physical interpretation of the joint singular values. By substituting Eq. 7, the spectral energy of the i-th component decomposes into textual and visual variances:

$$\begin{aligned} \tilde{\sigma}_i^2 &= \underbrace{\|\mathbf{u}_i^T \mathbf{W}\mathbf{S}_T\|_2^2}_{\sigma_{text}^2} + \alpha \underbrace{\|\mathbf{u}_i^T \mathbf{W}\mathbf{S}_I\|_2^2}_{\sigma_{img}^2} \\ &\approx \sigma_{text,i}^2 (1 + \alpha \cdot \lambda_i) \end{aligned} \tag{11}$$

This derivation explicitly links the new spectrum to the mismatch factor λ_i from Eq. 5. Consequently, directions with high visual importance (large λ_i) are amplified by the $\alpha \cdot \lambda_i$ term. This pushes visual features to the top of the sorting queue, preventing the collapse observed in text-only baselines.

4.2 Global Spectrum-Aware Truncation

Standard compression typically enforces a uniform ratio (e.g., 20%) across all layers. However, our analysis in Section 3.3 reveals that compressibility is highly heterogeneous: "Mismatch Layers" (e.g., backbone MLPs) exhibit slow spectral decay requiring higher rank, whereas other modules are often redundant. To address this, we propose *Global Spectrum-Aware Truncation*, a strategy that dynamically reallocates the parameter budget across the entire MLLM architecture.

We model the MLLM as two distinct spectral sources: the Vision Tower $\mathcal{V}$ and the LLM Backbone $\mathcal{B}$.

- **Vision Tower ($\mathcal{V}$):** Processing unimodal data, the encoder is immune to cross-modal mismatch. We therefore apply standard Activation-Aware SVD (Sec. 3.2) using image calibration to derive singular values $\Sigma^{(l)}$ for all $l \in \mathcal{V}$.
- **LLM Backbone ($\mathcal{B}$):** For layers where modalities interact, we apply the proposed JW-SVD (Sec. 4.1.2) to derive joint singular values $\Sigma^{(l)}$ for $l \in \mathcal{B}$, explicitly accounting for cross-modal alignment.

Unified Spectral Pool and Budget Allocation. We aggregate the singular values from all layers into a single *Unified Spectral Pool* $\mathcal{S}_{global}$, enabling a direct comparison of feature importance across disparate architectures:

$$\mathcal{S}_{global} = \bigcup_{l \in \mathcal{V}} \{\sigma_i^{(l)}\} \cup \bigcup_{l \in \mathcal{B}} \{\tilde{\sigma}_j^{(l)}\} \tag{12}$$

Global Thresholding Strategy. Given a target parameter budget B_{target}, we sort the elements of $\mathcal{S}_{global}$ in descending order to determine a global inclusion threshold τ such that $|\{s \in \mathcal{S}_{global} : s \geq \tau\}| = B_{target}$. The layer-specific retention rank k_l is then derived dynamically:

$$k_l = \sum_i \mathbb{I}(\sigma_i^{(l)} \geq \tau) \tag{13}$$

This mechanism automatically executes an *Inter-Module Budget Transfer*. As illustrated in Figure 5, Vision Tower layers exhibit rapid spectral decay, retaining 90% of variance with significantly lower rank ratios (15–25%) compared to the Multimodal Backbone. Consequently, under a unified threshold τ, the optimization naturally assigns lower ranks to

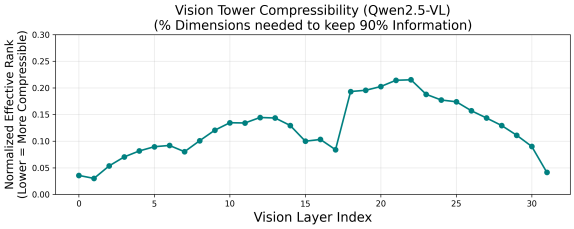

Figure 5: **Vision Tower Spectral Decay.** Normalized effective rank required to preserve 90% of spectral energy across layers. The Vision Tower exhibits high redundancy (sufficient rank $< 20\%$), contrasting with the slower spectral decay of the multimodal backbone. This disparity motivates the inter-module budget transfer.

the redundant vision layers, reallocating the saved parameter capacity to the backbone layers where the cross-modal mismatch (quantified by λ_i) necessitates higher rank retention.

Weight Reconstruction. Upon determining the layer-specific rank k_l, we synthesize the compressed weight matrix $\hat{\mathbf{W}}^{(l)}$. We decompose the truncated approximation into two low-rank factors $\mathbf{A} \in \mathbb{R}^{d_{out} \times k_l}$ and $\mathbf{B} \in \mathbb{R}^{k_l \times d_{in}}$ via symmetric spectral distribution:$\mathbf{A} = \mathbf{U}_{:,1:k_l} \mathbf{\Sigma}_{1:k_l}^{1/2}$, $\mathbf{B} = \mathbf{\Sigma}_{1:k_l}^{1/2} \mathbf{V}_{:,1:k_l}^T \mathbf{S}_{joint}^{-1}$ The term $\mathbf{S}_{joint}^{-1}$ in $\mathbf{B}$ implicitly reverses the joint whitening transformation, mapping the spectral components back to the original parameter space. The final compressed module is implemented as the product $\hat{\mathbf{W}}^{(l)} = \mathbf{AB}$.

5 Experiments

5.1 Performance Analysis

Experimental Setup. We evaluate **Qwen2.5-VL-7B** and **Llama-3-Next-8B** on NVIDIA H20 GPUs, utilizing WikiText-2 and random image samples for calibration. We compare JW-SVD against the uncompressed FP16 model and state-of-the-art baselines: SVD-LLM (Wang et al., 2025c) and ASVD (Yuan et al., 2025). Evaluation covers 8 benchmarks across three categories: *Perception* (MME (Fu et al., 2025a), BLINK (Fu et al., 2024), HallusionBench (Guan et al., 2024), OCR-Bench (Liu et al., 2024)); *Reasoning* (MMMU (Yue et al., 2024), MathVista (Lu et al., 2024)); and *General* (SeedBench (Li et al., 2023), ScienceQA (Lu et al., 2022)). To ensure a rigorous evaluation of geometric alignment, we utilized a large-scale calibration set of 1024 randomly sampled images from the COCO(Lin et al., 2015) dataset and 1024 text samples from WikiText-2(Merity et al., 2016). To obtain valid visual activations, each image input includes a generic instruction ('Describe this image'). This is distinct from the Text Calibration Data (WikiText-2), which is unrelated to the images.

Main Results. As detailed in Table 1, JW-SVD consistently outperforms baselines, with significant margins at 40% compression. A critical finding is the divergence in modality retention: while SVD-LLM maintains text reasoning (ScienceQA), it exhibits catastrophic degradation on visual perception tasks (e.g., MME, BLINK). This confirms that text-only objectives prune critical visual variance. In contrast, JW-SVD effectively bridges this alignment gap; by incorporating the Joint Covariance structure, it preserves the high-frequency visual features required for fine-grained tasks without compromising the language backbone.

5.2 Textual Capabilities Retention

To verify that integrating visual constraints does not degrade language modeling, we evaluate perplexity (PPL) on WikiText-103 (Gu et al., 2024). Results in Table 2 demonstrate that JW-SVD maintains textual proficiency comparable to the text-optimized SVD-LLM baseline, with negligible PPL divergence even at 40% compression. This confirms that $\mathbf{S}_{joint}$ successfully identifies a shared subspace that accommodates visual variance without compromising the structural integrity of the language backbone, effectively optimizing the cross-modal Pareto frontier.

5.3 Efficiency and Accuracy Tradeoff

Experimental Setup. We analyze the trade-off between multimodal perception (MME Score) and inference latency (ms/token) on a single NVIDIA A100 GPU. We compare JW-SVD against the uniform SVD-LLM baseline across 80%, 60%, and 40% retention ratios.

Pareto Frontier Analysis. JW-SVD establishes a superior Pareto frontier through strategic budget reallocation. While SVD-LLM suffers catastrophic collapse at 40% retention due to aggressive backbone pruning, JW-SVD trades marginal latency (< 3ms) for a $>30\%$ recovery in perceptual capability. By transferring redundancy from the Vision Tower to the backbone, our method maintains a robust performance profile, effectively mitigating

Method	Ratio	Perception & Fine-Grained				Reasoning		General	
		MME	BLINK	HallB	OCRBench	MMMU	MathVista	SeedBench	SciQA
Original	100%	2154	56.5	52.4	855	55.4	67.1	62.5	70.1
Light Compression									
ASVD	80%	1921	48.2	48.4	812	44.7	62.1	58.2	65.4
SVD-LLM	80%	2086	51.5	50.2	835	47.8	64.5	60.1	**68.2**
JW-SVD (Ours)	80%	**2117**	**52.1**	**51.0**	**840**	**50.1**	**64.8**	**60.5**	68.1
Medium Compression									
ASVD	60%	1653	42.4	40.1	750	43.6	56.5	52.4	58.9
SVD-LLM	60%	1782	45.1	42.5	785	45.2	59.8	55.3	63.5
JW-SVD (Ours)	60%	**2005**	**49.8**	**48.5**	**821**	**48.7**	**63.0**	**59.1**	**64.0**
Heavy Compression									
ASVD	40%	1156	31.5	35.4	625	36.2	48.4	41.5	48.2
SVD-LLM	40%	1312	34.3	39.8	654	36.5	51.2	45.2	57.1
JW-SVD (Ours)	40%	**1854**	**44.5**	**44.2**	**795**	**42.1**	**59.5**	**56.8**	**58.5**

Table 1: **Capability Retention across Compression Ratio.** We compare JW-SVD against baselines on comprehensive MLLM benchmarks. **Visual Mismatch in SVD-LLM:** At the 40% budget, SVD-LLM exhibits a severe degradation in perception tasks . **Recovery via Joint-Whitening:** JW-SVD successfully preserves the visual manifold, recovering 70% of the lost performance while maintaining parity on text-heavy reasoning tasks SciQA.

Method	Ratio	WikiText-103 (PPL) ↓	
		Qwen2.5-VL	Llama-Next
Original (FP16)	100%	7.35	8.13
ASVD	80%	8.10	9.55
SVD-LLM	80%	7.85	8.52
JW-SVD (Ours)	80%	**7.98**	**8.65**
ASVD	60%	17.85	18.80
SVD-LLM	60%	13.92	14.40
JW-SVD (Ours)	60%	**15.95**	**17.12**
ASVD	40%	1674.50	2672.20
SVD-LLM	40%	57.15	66.35
JW-SVD (Ours)	40%	**59.23**	**69.58**

Table 2: **Textual Capabilities Retention.** Perplexity scores on WikiText-103. Despite injecting image constraints into the optimization objective, **JW-SVD** maintains perplexity scores nearly identical to the text-optimized **SVD-LLM** baseline. This confirms that our Joint-Whitening approach recovers visual capabilities without compromising the language modeling manifold.

Configuration	Strategy		Perception	Reasoning
	Whitening	Allocation	BLINK	SciQA
JW-SVD (Ours, $\gamma = 1.0$)	**Joint**	**Global (Donor)**	**44.5**	**58.5**
Ablation 1: Whitening Strategy (Fixed Global Allocation)				
Text-Only C	Text	Global	34.8 (-9.7)	58.7 (+0.2)
Image-Only C	Image	Global	43.1 (-1.4)	49.2 (-9.3)
Ablation 2: Allocation Strategy (Fixed Joint Whitening)				
Uniform Ratio	Joint	Uniform (40%)	39.5 (-5.0)	54.2 (-4.3)
Frozen Vision	Joint	Backbone-Only	38.1 (-6.4)	51.5 (-7.0)
Ablation 3: Scale-Invariant Energy Balancing (γ sweep)				
Strong Text Bias ($\gamma = 0.1$)	Joint	Global	35.2 (-9.3)	58.7 (+0.2)
Moderate Text Bias ($\gamma = 0.5$)	Joint	Global	40.1 (-4.4)	58.6 (+0.1)
Moderate Image Bias ($\gamma = 2.0$)	Joint	Global	44.8 (+0.3)	56.2 (-2.3)
Strong Image Bias ($\gamma = 10.0$)	Joint	Global	44.9 (+0.4)	51.0 (-7.5)

Table 3: **Component Analysis at 40% Budget.** *Ablation 1* shows that Text-Only whitening causes visual collapse (BLINK drops 9.7 points). *Ablation 2* confirms that the "Donor Mechanism" (Global) is critical; freezing the Vision Tower forces the Backbone to be over-compressed, hurting both metrics. *Ablation 3* validates the proposed Scale-Invariant Energy Balancing factor (α); scaling it by $\gamma = 1.0$ achieves the Pareto-optimal frontier, whereas heavily down-scaling or up-scaling α collapses perception and reasoning, respectively.

the feature collapse inherent to uniform truncation strategies.

Finally, we demonstrate that JW-SVD is orthogonal to numerical quantization, enabling extreme compression (1.6 bits) when combined with GPTQ. Detailed results are provided in Appendix 5.5.

5.4 Ablation Studies

We conduct component-wise ablation at 40% compression to isolate the impact of dual-objective whitening and global allocation, using BLINK (perception) and SciQA (reasoning) as representative metrics.

Impact of Joint Whitening. Table 3 demonstrates that Joint Whitening is crucial for resolving cross-modal misalignment. Text-only whitening causes a 9.7-point drop in BLINK by suppressing visual features, while image-only whitening degrades SciQA by 9.3 points. JW-SVD achieves a Pareto-optimal balance, preserving the spectral variance required for both manifolds without compromising either modality.

Global Budget Allocation. We validate Global Spectrum-Aware Truncation against uniform and backbone-only baselines. Uniform allocation (40%) proves sub-optimal by over-allocating capacity to the redundant Vision Tower. Furthermore,

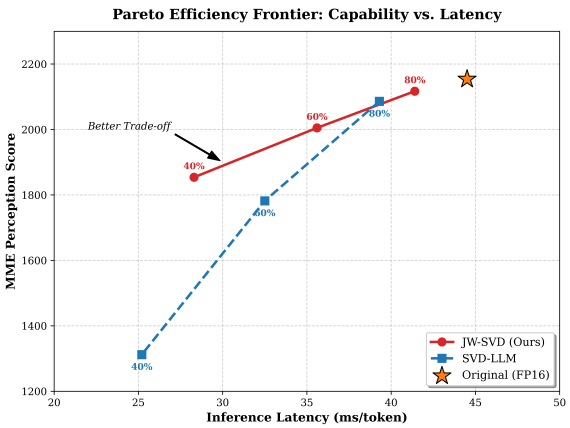

Figure 6: **Pareto Efficiency Frontier.** We plot MME Perception Score against inference latency. While SVD-LLM (Blue) degrades rapidly in capability to gain speed, JW-SVD (Red) maintains high performance, offering a superior trade-off. The Original model (Star) is shown for reference.

a "Backbone-Only" approach forces the sensitive backbone to absorb the full compression burden, degrading perception (-6.4) and reasoning (-7.0). This confirms that transferring parameter budget from the vision encoder to the backbone is essential to prevent feature collapse.

Optimality of Scale-Invariant Energy Balancing (α). To verify the optimality of our analytically computed weighting factor α, we conduct a sensitivity analysis by introducing a scaling factor γ such that $\alpha_{scaled} = \gamma \cdot \alpha$. As shown in Table 3, our computed α ($\gamma = 1.0$) lies at the Pareto frontier. Decreasing the weight ($\gamma < 1.0$) triggers the visual collapse predicted in Section 3.2, with BLINK scores dropping by up to 9.3 points. Conversely, over-weighting the visual modality ($\gamma > 1.0$) yields negligible gains in perception (+0.3 to +0.4) but causes significant degradation in textual reasoning (up to -7.5 points on SciQA).

5.5 Compound Compression with Quantitization

Experimental Setup. To examine synergy between geometric and numerical compression, we evaluate JW-SVD combined with 4-bit GPTQ quantization. We compare a standard W4A16 baseline against a compound pipeline: JW-SVD (40% rank) followed by 4-bit GPTQ on the low-rank factors. This yields an extreme effective compression rate of 1.6 bits per parameter (10% of the original FP16 footprint).

Method	Config	Memory (Effective)	Perception BLINK	Reasoning SciQA
Original	FP16	100% (16GB)	56.5	70.1
Single-Mode Compression				
GPTQ	W4A16	25% (4.0GB)	45.8	**68.5**
JW-SVD (Ours)	40% Rank (FP16)	40% (6.4GB)	**44.5**	58.5
Compound Compression (Orthogonality Test)				
JW-SVD + GPTQ	**40% Rank + W4**	**10% (1.6GB)**	**42.8**	57.1

Table 4: **Orthogonality & Compound Compression.** We evaluate the synergy between geometric compression (JW-SVD) and precision quantization (GPTQ)(Frantar et al., 2023). **Single-Mode:** Standard GPTQ (4-bit) degrades visual perception (BLINK: 45.8) significantly compared to the original. **Compound:** By applying GPTQ on top of JW-SVD, we achieve an extreme compression rate (1.6GB memory, ≈1.6-bit effective) while retaining comparable perceptual performance (33.8) to the much larger FP16 low-rank model.

Analysis. Table 4 confirms the orthogonality of JW-SVD to quantization. While standard GPTQ (W4) degrades visual perception (BLINK: 45.8), the compound JW-SVD + GPTQ approach maintains comparable accuracy (42.8) at significantly reduced memory cost (1.6GB vs. 4.0GB). This demonstrates that JW-SVD extracts a robust geometric manifold stable under precision reduction. By targeting rank redundancy prior to quantization, JW-SVD enables extreme compression levels unattainable by either method in isolation.

6 Conclusion

We identify *cross-modal spectral mismatch* as the primary bottleneck in MLLM compression. To address this, we propose **Joint-Whitening SVD (JW-SVD)** and **Global Spectrum-Aware Truncation**. Across 9 diverse benchmarks, JW-SVD demonstrates superior retention of both text and image capabilities, recovering over 30% of the perceptual performance lost by baselines. Furthermore, its synergy with quantization facilitates effective 1.6-bit compression, enabling robust deployment on resource-constrained devices.

Limitations

While JW-SVD effectively mitigates cross-modal mismatch, its Joint Covariance estimation relies heavily on the calibration data's distribution. Consequently, performance on specialized out-of-distribution domains (e.g., medical imaging) may vary. Additionally, extending this framework beyond standard Vision-Tower-plus-LLM architectures to early-fusion or Mixture-of-Experts (MoE) models remains future work.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. *Preprint*, arXiv:2204.14198.

Saleh Ashkboos, Maximilian L. Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. 2024. Slicegpt: Compress large language models by deleting rows and columns. *Preprint*, arXiv:2401.15024.

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

Patrick H. Chen, Hsian-fu Yu, Inderjit S. Dhillon, and Cho-jui Hsieh. 2021. Drone: data-aware low-rank compression for large nlp models. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. Gptq: Accurate post-training quantization for generative pre-trained transformers. *Preprint*, arXiv:2210.17323.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025a. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. *Preprint*, arXiv:2404.12390.

Yao Fu, Runchao Li, Xianxuan Long, Haotian Yu, Xiaotian Han, Yu Yin, and Pan Li. 2025b. Pruning weights but not truth: Safeguarding truthfulness while pruning llms. *Preprint*, arXiv:2509.00096.

Yao Fu, Xianxuan Long, Runchao Li, Haotian Yu, Mu Sheng, Xiaotian Han, Yu Yin, and Pan Li. 2025c. Quantized but deceptive? a multi-dimensional truthfulness evaluation of quantized LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30435–30458, Suzhou, China. Association for Computational Linguistics.

Albert Gu, Karan Goel, and Christopher R´e. 2024. Wikitext-103 dataset.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *Preprint*, arXiv:2310.14566.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In *Computer Vision – ECCV 2020*, pages 709–727, Cham. Springer International Publishing.

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

Yuxuan Jiang and Francis Ferraro. 2026. Beyond math: Stories as a testbed for memorization-constrained reasoning in llms. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5590–5607.

Yuxuan Jiang, Dawei Li, and Frank Ferraro. 2025. Drp: Distilled reasoning pruning with skill-aware step decomposition for efficient large reasoning models. *arXiv preprint arXiv:2505.13975*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Kunxi Li, Zhonghua Jiang, Zhouzhou Shen, Zhaode Wang, Chengfei Lv, Shengyu Zhang, Fan Wu, and Fei Wu. 2025b. Madakv: Adaptive modality-perception kv cache eviction for efficient multimodal long-context inference. *Preprint*, arXiv:2506.15724.

Runchao Li, Yao Fu, Mu Sheng, Xianxuan Long, Haotian Yu, and Pan Li. 2025c. FAEDKV: Infinite-window Fourier transform for unbiased KV cache compression. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16856–16866, Suzhou, China. Association for Computational Linguistics.

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. In *Proceedings of Machine Learning and Systems*, volume 6, pages 87–100.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2025. World model on million-length video and language with blockwise ringattention. *Preprint*, arXiv:2402.08268.

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

Xianxuan Long, Yao Fu, Runchao Li, Mu Sheng, Haotian Yu, Xiaotian Han, and Pan Li. 2025. When truthful representations flip under deceptive instructions? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16315–16335, Suzhou, China. Association for Computational Linguistics.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Pan Lu, Swaroop Mishra, Tony 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. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. Llm-pruner: On the structural pruning of large language models. *Preprint*, arXiv:2305.11627.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Sharath Turuvekere Sreenivas, Saurav Muralidharan, Raviraj Joshi, Marcin Chochowski, Ameya Sunil Mahabaleshwarkar, Gerald Shen, Jiaqi Zeng, Zijia Chen, Yoshi Suhara, Shizhe Diao, Chenhan Yu, Wei-Chun Chen, Hayley Ross, Oluwatobi Olabiyi, Ashwath Aithal, Oleksii Kuchaiev, Daniel Korzekwa, Pavlo Molchanov, Mostofa Patwary, and 3 others. 2024. Llm pruning and distillation in practice: The minitron approach. *Preprint*, arXiv:2408.11796.

Mart van Baalen, Andrey Kuzmin, Ivan Koryakovskiy, Markus Nagel, Peter Couperus, Cedric Bastoul, Eric Mahurin, Tijmen Blankevoort, and Paul Whatmough. 2025. Gptvq: The blessing of dimensionality for llm quantization. *Preprint*, arXiv:2402.15319.

Qinsi Wang, Jinghan Ke, Masayoshi Tomizuka, Yiran Chen, Kurt Keutzer, and Chenfeng Xu. 2025a. Dobi-svd: Differentiable svd for llm compression and some new perspectives. *Preprint*, arXiv:2502.02723.

Xin Wang, Samiul Alam, Zhongwei Wan, Hui Shen, and Mi Zhang. 2025b. Svd-llm v2: Optimizing singular value truncation for large language model compression. *Preprint*, arXiv:2503.12340.

Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. 2025c. Svd-llm: Truncation-aware singular value decomposition for large language model compression. *Preprint*, arXiv:2403.07378.

Ningning Xu, Yuxuan Jiang, Shubhashis Roy Dipta, and Zhang Hengyuan. 2025. Learning how to use tools, not just when: Pattern-aware tool-integrated reasoning. *MATH-AI @ NeurIPS 2025*.

Zhihang Yuan, Yuzhang Shang, Yue Song, Dawei Yang, Qiang Wu, Yan Yan, and Guangyu Sun. 2025. Asvd: Activation-aware singular value decomposition for compressing large language models. *Preprint*, arXiv:2312.05821.

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, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Hengyuan Zhang, Shiping Yang, Xiao Liang, Chenming Shang, Yuxuan Jiang, Chaofan Tao, Jing Xiong, Hayden Kwok-Hay So, Ruobing Xie, Angel X Chang, and 1 others. 2025. Find your optimal teacher: Personalized data synthesis via router-guided multi-teacher distillation. *arXiv preprint arXiv:2510.10925*.

A Derivation of Cross-Modal Reconstruction Loss

In this section, we provide the formal derivation for the reconstruction loss under covariance mismatch.

Setup. Let $\mathbf{W} \in \mathbb{R}^{d_{out} \times d_{in}}$ be the original weight matrix. We apply Singular Value Decomposition (SVD) on the weight matrix whitened by text statistics $\mathbf{S}_{\text{text}}$:

$$\mathbf{W}\mathbf{S}_{\text{text}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \tag{14}$$

We approximate the weights by keeping the top-k singular values. The approximation error matrix in the whitened space is $\mathbf{\Delta} = \sum_{i=k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$.

Derivation. We seek to measure the Frobenius reconstruction loss on the image modality, characterized by the image covariance matrix $\mathbf{X}_{\text{img}}\mathbf{X}_{\text{img}}^T \approx \mathbf{S}_{\text{img}}\mathbf{S}_{\text{img}}^T$. The loss is defined as:

$$\mathcal{L}_{\text{img}} = \|\mathbf{\Delta}\mathbf{S}_{\text{text}}^{-1}\mathbf{X}_{\text{img}}\|_F^2 \tag{15}$$

Using the trace property $\|\mathbf{A}\|_F^2 = \text{Tr}(\mathbf{A}\mathbf{A}^T)$ and defining the mismatch matrix $\mathbf{M} = \mathbf{S}_{\text{text}}^{-1}\mathbf{S}_{\text{img}}$, we expand the term:

$$\mathcal{L}_{\text{img}} = \text{Tr}\left((\mathbf{\Delta}\mathbf{S}_{\text{text}}^{-1}\mathbf{X}_{\text{img}})(\mathbf{\Delta}\mathbf{S}_{\text{text}}^{-1}\mathbf{X}_{\text{img}})^T\right) \tag{16}$$

$$= \text{Tr}\left(\mathbf{\Delta}(\mathbf{S}_{\text{text}}^{-1}\mathbf{S}_{\text{img}}\mathbf{S}_{\text{img}}^T\mathbf{S}_{\text{text}}^{-T})\mathbf{\Delta}^T\right) \tag{17}$$

$$= \text{Tr}\left(\mathbf{\Delta}\mathbf{M}\mathbf{M}^T\mathbf{\Delta}^T\right) \tag{18}$$

Substituting the spectral expansion $\mathbf{\Delta} = \sum_{i=k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ yields a double summation over indices $i, j \in \{k+1, \ldots, r\}$:

$$\mathcal{L}_{\text{img}} = \text{Tr}\left(\sum_i \sum_j \sigma_i \sigma_j (\mathbf{u}_i \mathbf{v}_i^T)(\mathbf{M}\mathbf{M}^T)(\mathbf{v}_j \mathbf{u}_j^T)^T\right) \tag{19}$$

Rearranging terms using the cyclic property of the Trace operator ($\text{Tr}(\mathbf{ABC}) = \text{Tr}(\mathbf{BCA})$):

$$\mathcal{L}_{\text{img}} = \sum_i \sum_j \sigma_i \sigma_j \,\text{Tr}\left(\mathbf{u}_i (\mathbf{v}_i^T \mathbf{M}\mathbf{M}^T \mathbf{v}_j)\mathbf{u}_j^T\right) \tag{20}$$

$$= \sum_i \sum_j \sigma_i \sigma_j (\mathbf{v}_i^T \mathbf{M}\mathbf{M}^T \mathbf{v}_j)\,\text{Tr}(\mathbf{u}_j^T \mathbf{u}_i) \tag{21}$$

Since the left singular vectors $\mathbf{U}$ are orthonormal, $\text{Tr}(\mathbf{u}_j^T \mathbf{u}_i) = \mathbf{u}_j^T \mathbf{u}_i = \delta_{ij}$. This orthogonality eliminates all cross-terms where $i \neq j$:

$$\mathcal{L}_{\text{img}} = \sum_{i=k+1}^{r} \sigma_i^2 (\mathbf{v}_i^T \mathbf{M}\mathbf{M}^T \mathbf{v}_i) \tag{22}$$

Recognizing that $\mathbf{v}_i^T \mathbf{M}\mathbf{M}^T \mathbf{v}_i = \|\mathbf{v}_i^T \mathbf{M}\|_2^2$, we arrive at the final expression:

$$\mathcal{L}_{\text{img}} = \sum_{i=k+1}^{r} \sigma_i^2 \|\mathbf{v}_i^T \mathbf{M}\|_2^2 \tag{23}$$