Modality-Aware Block Rotation for Vision-Language-Action Model Quantization

Anonymous Authors[1]

Abstract

Vision-Language-Action (VLA) models enable unified perception, reasoning, and control, but their deployment is constrained by the large memory footprint. Although post-training quantization (PTQ) is a promising solution, existing rotation-based methods fail under 4-bit quantization. We attribute this failure to cross-modal heterogeneity, where vision and language tokens share the same layers but exhibit highly heterogeneous activation statistics, resulting in severe mismatch in both activation scaling and Hessian structure. This mismatch fundamentally breaks the assumptions behind existing rotation-based and Hessian-aware quantization methods. We propose *Modality-Aware Block Rotation* (MABR), which preserves modality-specific channel structure by restricting rotation within modality-consistent groups. This prevents the diffusion of dominant language activations into vision channels and enables stable low-bit quantization. On OpenVLA-7B, MABR substantially bridges the gap to full-precision performance and remains stable where naive 4-bit quantization collapses, incurring only a 3.0% performance drop without any fine-tuning.

1. Introduction

Recently, Vision-Language-Action (VLA) models have emerged as a foundation for general-purpose robot manipulation by unifying visual perception, language understanding, and action generation within a single framework (Brohan et al., 2022; Zitkovich et al., 2023; Black et al., 2024; Bjorck et al., 2025). Despite their strong performance, their massive parameter size often exceeds the memory capacity of resource-constrained edge devices, hindering real-time deployment.

To mitigate this problem, post-training quantization (PTQ) has emerged as an attractive solution because of its ability to compress models without expensive retraining. This advantage is particularly significant for VLA models, where memory consumption is dominated by the large language model (LLM) backbone. For example, in OpenVLA-7B (Kim et al., 2024), the LLM alone occupies approximately 13GB, accounting for nearly 90% of the total memory footprint. Consequently, recent studies have focused on quantizing the LLM backbone of VLA models (Xu et al., 2026b; Zhang et al., 2026; Xu et al., 2026a; Wang et al., 2024).

Among these methods, rotation-based quantization has gained attention because of its efficiency and easy integration with existing PTQ pipelines. However, directly applying such methods to VLA models causes severe performance degradation under 4-bit quantization. This issue is not simply due to reduced numerical precision, but stems from a structural mismatch between the homogeneous assumptions of LLM quantization and the heterogeneous activations observed in VLA models (Liu et al., 2024; Lin et al., 2024a; An et al., 2025).

Our key observation is that VLA layers, especially early MLP layers, suffer from acute *cross-modal heterogeneity*. Before rotation, channels are distinctly separated by modality. However, because language-token activations are orders of magnitude larger than vision activations, global rotation destroys this structural separation by spreading language-dominant energy across all channels. As a result, a small set of outlier channels is transformed into widespread contamination across the activation space, as illustrated in Fig. 1(b). This contamination propagates to the Hessian, distorting its curvature structure such that it no longer reflects the semantic roles of individual channels. Consequently, Hessian-weighted methods such as GPTQ (Frantar et al., 2022) becomes misaligned with the underlying modality structure, prioritizing contaminated language interference while degrading critical vision representations.

To address this cross-modal contamination, we propose *Modality-Aware Block Rotation* (MABR), a quantization strategy that preserves modality-specific channel structure. Instead of applying a homogeneous global rotation, MABR performs modality-aware block-wise rotation that isolates heterogeneous channel groups during transformation. This isolation prevents high magnitude language activations from

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

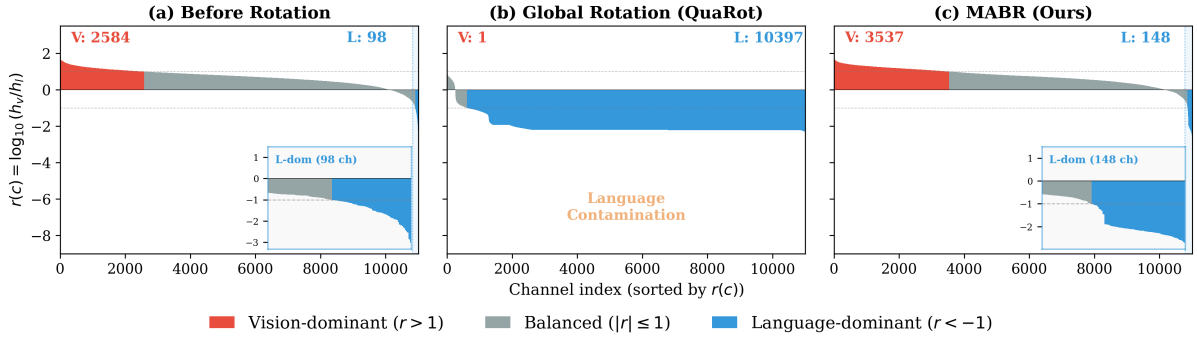

Figure 1. Channel-level modality structure before and after rotation, measured by the dominance ratio $r(c) = \log_{10}(h_v/h_l)$, where h_v and h_l denote the mean squared activation energy over vision and language tokens, respectively. (a) Before rotation, vision-dominant (red) and language-dominant (blue) channels occupy clearly separated regions, with 2584 vision and 98 language-dominant channels. (b) Global rotation collapses this structure: vision-dominant channels drop to 1 while language-dominant channels surge to 10397, spreading language energy across the entire channel space. (c) MABR preserves the modality-separated structure, retaining 3537 vision and 148 language-dominant channels with less cross-modal contamination.

corrupting visual channels, thereby preserving the semantic integrity required for precise and long-horizon embodied control, while still achieving the efficiency benefits of 4-bit quantization. The key insight is that channels in VLA models play distinct functional roles, where vision channels support perception, language channels support reasoning, and action channels support control. Global rotation disrupts this functional separation, whereas MABR preserves separation by balancing activation statistics with reduced cross-modal contamination. Our main contributions are summarized as follows:

- We identify and characterize cross-modal heterogeneity as the primary cause of failure in VLA quantization.
- We propose Modality-Aware Block Rotation (MABR), a modality-aware block-wise rotation scheme that preserves semantic channel structure while enabling effective 4-bit quantization.
- We demonstrate that MABR achieves a 73.5% average success rate on LIBERO under 4-bit quantization of OpenVLA-7B, approaching the full-precision baseline while significantly reducing deployment cost.

2. Analysis

In this section, we analyze why existing PTQ methods (Liu et al., 2024; Ashkboos et al., 2024; Frantar et al., 2022; Lin et al., 2024a) fail in VLA models. We attribute this failure to cross-modal heterogeneity, where vision and language tokens share the same layers but exhibit highly heterogeneous activation statistics, resulting in severe mismatch in both activation and Hessian statistics. Existing PTQ methods assume a homogeneous activation geometry designed for unimodal LLMs, whereas VLA models inherently rely on modality-separated hidden representations.This mismatch emerges in the activation space, propagates to the Hessian, and is further amplified by global rotation.

2.1. Cross-Modal Heterogeneity in VLA Activations

Cross-modal heterogeneity in VLA models manifests in two distinct forms: a severe scale disparity in activation magnitude and a clear channel-level separation by modality. Unlike homogeneous LLMs, VLA models process vision and language inputs to generate action tokens through shared transformer layers, making this heterogeneity an inherent structural property of the activation space.

Scale disparity. In layer 1 of OpenVLA-7B, language token activations are $635\times$ larger in magnitude than vision token activations (Tab. 2, L/V $t_{\max}$ ratio). When a single step size is shared across modalities, the language distribution dominates the scale, leaving visual features numerically unresolvable.

Channel separation. As shown in Fig. 1(a), vision-dominant (red) and language-dominant (blue) channels occupy largely disjoint subsets of the channel dimensions. This reflects a functional partitioning induced by the joint VLA training objective.

2.2. Hessian Anisotropy under Cross-Modal Heterogeneity

The cross-modal heterogeneity identified in the activation space manifests directly in the Hessian matrix $\mathbf{H}$, which governs the quantization reconstruction loss:

$$\mathcal{L}(\delta\mathbf{W}) = \delta\mathbf{W}^\top\mathbf{H}, \delta\mathbf{W}, \qquad \mathbf{H} = \mathbb{E}[\mathbf{X}\mathbf{X}^\top]. \quad (1)$$

where $\mathbf{W}$ denotes the layer weights, $\mathbf{X}$ represents the input activations, and $\delta\mathbf{W} = \mathbf{W} - \hat{\mathbf{W}}$ is the quantization error with respect to the quantized weights $\hat{\mathbf{W}}$.

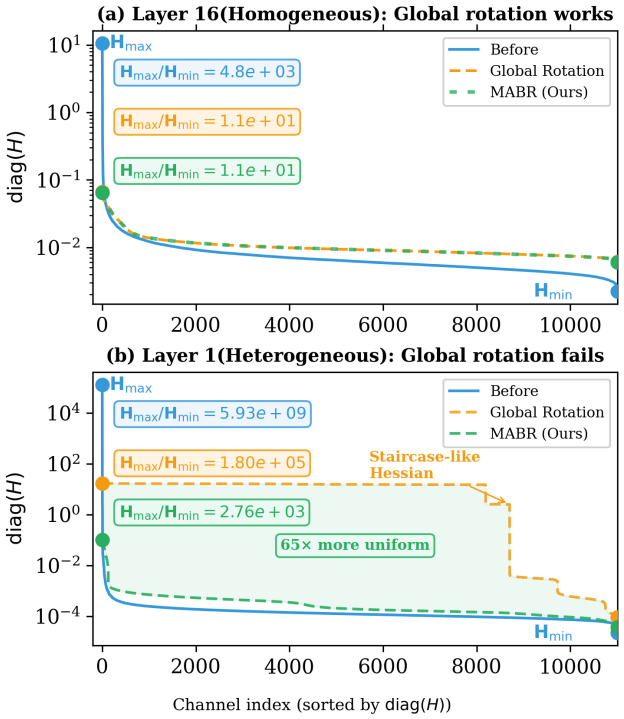

Figure 2. Comparison of Hessian structure in homogeneous and heterogeneous layers. (a) In homogeneous layers, global rotation produces a smooth and well-behaved Hessian. (b) In heterogeneous layers, however, global rotation fails, resulting in a staircase-like Hessian with severe distortion. In contrast, our MABR restores a smooth and stable Hessian even in heterogeneous layers.

This anisotropy is not uniform across entire layers. While middle layers appear relatively homogeneous, early layers exhibit extreme anisotropy. As illustrated in Fig. 2, the Hessian diagonal ratio ($\mathbf{H}_{\max}/\mathbf{H}_{\min}$) at layer 1 reaches 5.93×10^9, compared to only 4.8×10^3 at layer 16.

To mitigate such disparities, PTQ methods like DuQuant (Lin et al., 2024a) apply orthogonal rotations to smooth activations and improve Hessian conditioning. These methods assume that such rotation renders Hessian approximately isotropic ($\mathbf{H} \approx \mathbf{I}$). Under this assumption, minimizing the Euclidean weight error $\|\delta\mathbf{W}\|^2$ serves as a sufficient proxy for minimizing the reconstruction loss $\mathcal{L}(\delta\mathbf{W})$, thereby justifying the use of simple Round-To-Nearest (RTN). However, in VLA models, the sheer magnitude of cross-modal heterogeneity prevents the Hessian from reaching an isotropic state even after rotation. Consequently, Hessian-aware optimization such as GPTQ (Frantar et al., 2022) remains essential even in the rotated space. Furthermore, as analyzed in the next section, global rotation itself disrupts the underlying multimodal structure, necessitating a more specialized, modality-aware approach.

2.3. Failure of Global Rotation under Cross-Modal Heterogeneity

Rotation-based quantization methods such as DuQuant (Lin et al., 2024a) and QuaRot (Ashkboos et al., 2024) apply a global orthogonal rotation $\mathbf{R}$ to activations, transforming $\mathbf{X} \to \mathbf{XR}$ while absorbing $\mathbf{R}^\top$ into the subsequent weight matrix. While this preserves the identity in full precision:

$$\mathbf{XR} \cdot \mathbf{R}^\top\mathbf{W} = \mathbf{X}(\mathbf{R}\mathbf{R}^\top)\mathbf{W} = \mathbf{XW}, \tag{2}$$

it fails in VLA models for two reasons rooted in cross-modal heterogeneity.

Irreversible vision signal loss. The mathematical equivalence of orthogonal rotation breaks once a quantization operator $\mathbf{Q}(\cdot)$ is introduced. Let x_v and x_l ($x_v \ll x_l$) denote the magnitudes of vision and language activations in a d-dimensional space. Dense global rotation approximately distributes activation energy uniformly across dimensions, causing each rotated channel to become

$$\tilde{x}_i \approx \frac{x_l}{\sqrt{d}}, \qquad \Delta \approx \frac{x_l/\sqrt{d}}{2^{b-1}-1}, \tag{3}$$

where Δ denotes the quantization step size and b denotes the bit-width.

For the visual signal to survive quantization, it must satisfy

$$x_v \geq \frac{x_l}{2^b - 2}, \tag{4}$$

as derived in Appendix B. However, at layer 1, the measured activation ratio $\max(x_l)/\max(x_v) = 635$ under 4-bit quantization ($b = 4$) severely violates this condition, as indicated by the L/V $t_{\max}$ ratio in Tab. 2.

As a result, the weak visual signal collapses into the zero quantization bin, yielding

$$\hat{\mathbf{y}} = Q(\mathbf{XR}) \cdot Q(\mathbf{R}^\top\mathbf{W}) \neq \mathbf{XW}, \tag{5}$$

showing that the equivalence in Eq. (2) breaks under low-bit quantization.

Corrupted Hessian and the step-wise spectrum. Beyond activation loss, global rotation distorts the Hessian landscape. In homogeneous layers (Fig. 2(a)), global rotation smooths curvature. However, in heterogeneous layers (Fig. 2(b)), it fails to isotropize the Hessian. Instead, mixing high-magnitude language signals with low-magnitude vision signals produces a staircase-like spectrum. Consequently, Hessian-aware optimization methods like GPTQ are misled to prioritize these noise-inflated directions, allocating insufficient precision to the suppressed vision-critical subspaces.

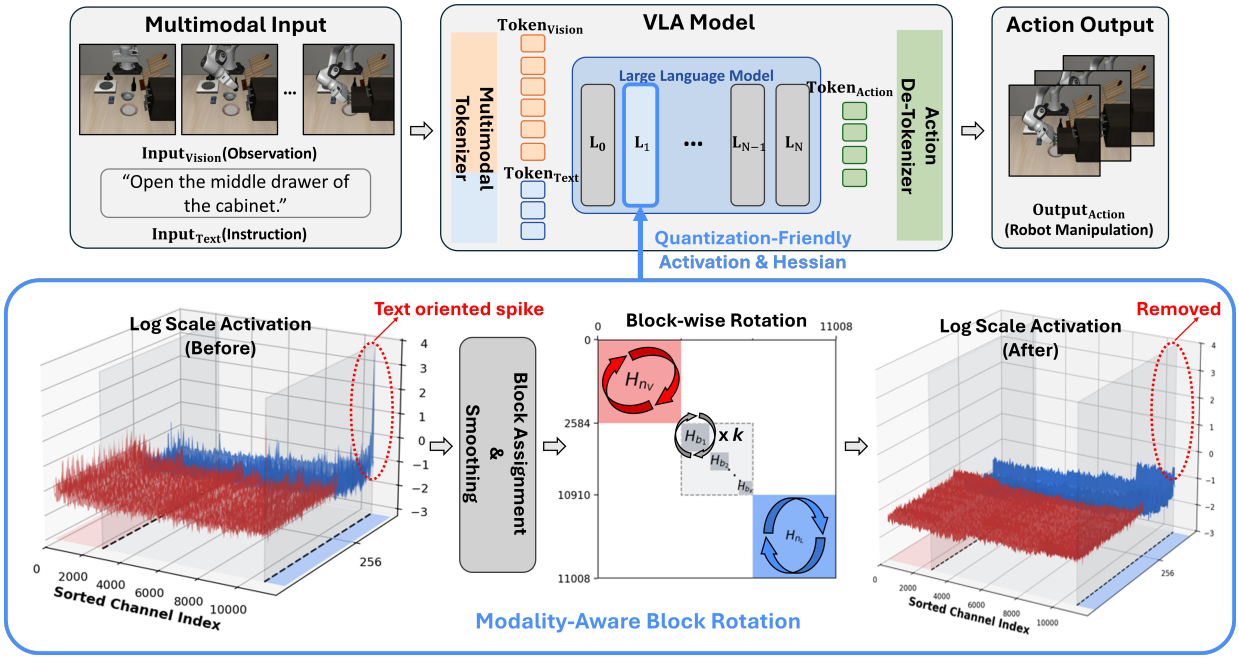

Figure 3. Overview of Modality-Aware Block Rotation (MABR). (*Top*) MABR targets the early layers of the LLM backbone, where cross-modal heterogeneity is most severe. (*Bottom*) Vision (red) and language-dominant (blue) channels are first identified and assigned to modality-consistent blocks. Within each block, a dedicated rotation is applied to smooth activations while preserving the modality structure. The text-oriented activation spike present before rotation is effectively removed without cross-modal contamination.

3. Method

3.1. Overview

Building on the analysis in Section 2, we identify that the failure of existing PTQ frameworks in VLA models stems from cross-modal heterogeneity. Global rotation, while effective in homogeneous LLMs, amplifies this heterogeneity by mixing modality-separated channels, corrupting both activation structure and the Hessian landscape.

To resolve this issue, we propose Modality-Aware Block Rotation (MABR). Rather than applying rotation across all channels, MABR restricts rotation to modality-consistent channel groups, preventing cross-modal contamination while enabling effective outlier smoothing. As illustrated in Fig. 3, MABR consists of four steps: modality heterogeneity evaluation, block assignment, pre-rotation smoothing, and block-wise rotation. The following subsections describe each component in detail. Our method builds on QuaRot and GPTQ as baseline quantization components (Ashkboos et al., 2024; Frantar et al., 2022). We provide additional details in Appendix A.4.

3.2. Modality-Aware Block Rotation

3.2.1. Modality Heterogeneity Evaluation

We identify which layers require modality-aware treatment. Given a calibration set, we measure the MLP intermediate activations for vision and language tokens separately and compute the modality ratio:

$$\mathcal{R}(\ell) = \frac{\mathbb{E}_{t\in L}\left[\max_c |a_{t,c}^{(\ell)}|\right]}{\mathbb{E}_{t\in V}\left[\max_c |a_{t,c}^{(\ell)}|\right]}, \tag{6}$$

where $a_{t,c}^{(\ell)}$ denotes the activation of token t at channel c in layer ℓ, and V and L denote the vision and language token sets, respectively. Layers where $\mathcal{R}(\ell) \gg 1$, such as $\mathbf{L_1}$ shown in Fig. 3, are classified as cross-modal heterogeneous layers, indicating a strong imbalance between language and vision activations. These layers are processed using the proposed modality-aware block rotation. All other layers, where the energy distribution remains relatively balanced, follow the standard global rotation baseline.

3.2.2. Block Assignment

For each cross-modal layer, we partition channels by their dominant modality. We compute per-channel activation energy separately for each modality and the modality dominance ratio $r(c)$:

$$r(c) = \log_{10}\left(\frac{\mathbb{E}_{t\in V}\left[a_{t,c}^2\right]}{\mathbb{E}_{t\in L}\left[a_{t,c}^2\right]}\right). \tag{7}$$

We sort all D channels in descending order of $r(c)$. This places vision-dominated channels first, balanced channels in

the middle, and language-dominated channels last, forming a modality-ordered sequence.

We then partition the sorted channels into power-of-2 sized blocks to enable efficient Fast Walsh-Hadamard Transform (FWHT) computation. Let $|\mathcal{C}_V|$ and $|\mathcal{C}_L|$ denote the number of vision and language-dominant channels identified by thresholding $r(c)$. Since FWHT requires power-of-2 block sizes, each group is rounded up to the nearest power of 2: $n_V = 2^{\lceil \log_2 |\mathcal{C}_V| \rceil}$ and $n_L = 2^{\lceil \log_2 |\mathcal{C}_L| \rceil}$. The remaining $D - n_V - n_L$ balanced channels are greedily decomposed into the largest possible power-of-2 sub-blocks:

$$\underbrace{n_V}_{\text{vision}} + \underbrace{b_1 + b_2 + \cdots + b_k}_{\text{balanced}} + \underbrace{n_L}_{\text{language}} = D, \quad \text{where } n_V,\ b_i,\ n_L \in \{2^j\}. \tag{8}$$

This partition confines vision and language-dominant channels to separate rotation groups, minimizing cross-modal contamination.

3.2.3. Pre-Rotation Smoothing

Before applying block rotation, we equalize per-channel activation scales via modality-aware channel smoothing. The per-channel scale factor is computed from the modality-specific activation energies:

$$\tilde{s}(c) = \frac{\mathbb{E}_{t \in L}[a_{t,c}^2]^{\alpha}}{\mathbb{E}_{t \in V}[a_{t,c}^2]^{1-\alpha}} \Big/ \bar{s}, \tag{9}$$

where $\bar{s}$ normalizes the mean to preserve the overall magnitude, and α controls the migration strength between modalities. Each channel is scaled independently by $\mathrm{diag}(1/\tilde{s})$: language-dominant channels are suppressed and vision-dominant channels are amplified, with no cross-channel interaction. This step follows the same dynamic-range transfer intuition as SmoothQuant (Xiao et al., 2023), but applies it within modality-consistent groups rather than globally.

Smoothing must precede rotation. Applying rotation first diffuses localized outliers and collapses inter-channel variance, rendering subsequent scaling ineffective. By smoothing first, we suppress outliers while preserving channel structure, enabling stable block-wise rotation. The scale is absorbed into $\mathbf{W}$ offline ($\mathbf{W} \leftarrow \mathbf{W} \cdot \mathrm{diag}(\tilde{s})$), with the online inverse ($\mathbf{X} \leftarrow \mathbf{X}/\tilde{s}$) applied before rotation, preserving exact equivalence at no runtime cost.

3.2.4. Block-Wise Rotation

After smoothing, we apply the block-diagonal Hadamard rotation that respects the modality partition established in Sec. 3.2.2. Let $\mathbf{P}$ be the permutation matrix that reorders channels into the modality-sorted layout. The block rotation matrix is defined as:

$$\mathbf{R}_{\text{block}} = \mathbf{P}^{\top} \underbrace{\begin{pmatrix} H_{n_V} & & & \\ & H_{b_1} & & \\ & & \ddots & \\ & & & H_{n_L} \end{pmatrix}}_{\mathbf{R}_{\text{sorted}}} \mathbf{P}, \tag{10}$$

where each $H_{(\cdot)}$ is an independent random Hadamard matrix. Unlike global rotation, $\mathbf{R}_{\text{block}}$ is block-diagonal in the sorted space, ensuring that activation mixing is restricted within each modality. This prevents cross-modal energy diffusion while still providing within-block smoothing of outliers.

Since each block is orthogonal, $\mathbf{R}_{\text{block}}$ preserves $\mathbf{R}_{\text{block}}^{\top}\mathbf{R}_{\text{block}} = \mathbf{I}$, and thus maintains the same invariance property as standard rotation-based PTQ. All subsequent quantization is applied on the rotated representation, using per-token symmetric activation quantization and GPTQ-based weight quantization with Hessian statistics estimated in the rotated space.

4. Experiments

4.1. Experimental Setup

We evaluate OpenVLA-7B (Kim et al., 2024) on the LIBERO benchmark across four task suites: Spatial, Object, Goal, and Long (Liu et al., 2023), reporting success rates under W4A4 and W8A8 quantization. GPTQ is applied with a group size of 128 and calibrated on 512 examples sampled evenly across the four suites. Each suite is evaluated on 10 tasks with 5 trials per task, and images are center-cropped with a scale of 0.9. We compare against FP16, SmoothQuant, OmniQuant, QVLA, and QuaRot+GPTQ (Xiao et al., 2023; Shao et al., 2023; Xu et al., 2026b; Ashkboos et al., 2024; Frantar et al., 2022).

4.2. Main Results

As shown in Tab. 1, naive global rotation (QuaRot + GPTQ) collapses to 51.4% under W4A4, confirming that a global Hadamard transform corrupts the Hessian structure of heterogeneous VLA activations.

MABR achieves the best W8A8 result (**77.5%**) and maintains strong performance under W4A4 (**73.5%**), yielding a substantial **+22.1 pp** improvement over global QuaRot. Under W8A8, the marginal gain over FP16 (76.5%) is within evaluation variance and thus not statistically significant. Although QVLA reports higher W4A4 accuracy, it relies on mixed-precision allocation that is less hardware-friendly.

Overall, MABR provides a practical training-free PTQ solution that preserves uniform W4A4 precision while effectively addressing cross-modal interference in VLA models.

Table 1. LIBERO success rate (%) under various weight-activation quantization settings. W4A4 and W8A8 refer to quantizing weights (W) and activations (A) to 4 and 8 bits, respectively. Results marked with † are cited directly from the QVLA paper.

Method	Spatial	Object	Goal	Long	Avg.
FP16	84.7	88.4	79.2	53.7	76.5
W8A8 Quantization					
SmoothQuant†	84.2	87.8	77.8	53.2	75.8
OmniQuant†	82.6	86.2	74.8	51.7	73.8
QVLA†	85.2	88.0	77.6	54.2	76.3
QuaRot + GPTQ	78.0	74.0	68.0	42.0	65.5
MABR (Ours)	**82.0**	**90.0**	**78.0**	**60.0**	**77.5**
W4A4 Quantization					
SmoothQuant†	69.2	73.2	69.6	40.9	63.2
OmniQuant†	82.2	85.4	75.4	50.3	73.3
QVLA†	84.4	87.6	78.8	53.0	76.0
QuaRot + GPTQ	60.0	55.6	60.0	30.0	51.4
MABR (Ours)	**80.0**	**86.0**	**70.0**	**58.0**	**73.5**

4.3. Effect of Modality-Aware Block Rotation

This ablation studies the effect of modality-aware rotation on activation statistics and Hessian conditioning. We measure modality imbalance using the maximum activation per channel ($t_{\max}$) and Hessian diagonal condition number κ, where $\kappa = \max_i \mathbf{H}_{ii} / \min_i \mathbf{H}_{ii}$.

In the original FP16 model, language channels dominate vision channels by $635\times$, resulting in an ill-conditioned Hessian ($\kappa = 5.93 \times 10^9$). While global rotation reduces the absolute amplitude of language outliers, it diffuses them across nearly all channels. As shown in Tab. 2, language features end up dominating 10,468 channels under global rotation, leading to cross-modal contamination. In contrast, MABR preserves modality separation, limiting language leakage into vision channels. This reduces the L/V $t_{\max}$ ratio to $2.6\times$ and improves Hessian conditioning to $\kappa = 3{,}433$, yielding a more balanced and quantization-friendly representation space.

4.4. Runtime Overhead

For the non-power-of-two intermediate dimension (D=11,008), QuaRot employs a Kronecker-factorized Hadamard rotation $H_{256} \otimes H_{43}$, whereas MABR applies independent FWHTs over modality-aware power-of-two blocks. Further implementation details are summarized in Appendix C.

MABR's higher latency (237 μs vs. 56 μs) stems from multiple independent FWHT kernel launches in the current non-fused implementation, and its larger storage (64.5 KB vs. 4.1 KB) comes from permutation indices used for channel grouping. Over 32 layers, these overheads amount to only $\sim$2.0 MB and $+5.8$ ms ($< 0.02\%$ memory and $< 1.2\%$ inference time), which is negligible relative to the **+22.1%** W4A4 accuracy gain in Tab. 1.

Table 2. Activation statistics for layer 1 down projection. $t_{\max}$ denotes the maximum absolute activation magnitude per channel, and κ represents the condition number of the Hessian diagonal. Lower L/V ratio and κ indicate a more quantization-friendly landscape.

Metric	Original (FP16)	Global Rot.	**Block Rot. (Ours)**
Vision $t_{\max}$	0.4399	0.0606	**0.1028**
Language $t_{\max}$	279.54	3.2104	**0.2715**
L/V $t_{\max}$ ratio	635×	53×	**2.6**×
Language κ	1.10×10^{12}	1.90×10^{7}	$\mathbf{1.47 \times 10^{5}}$
Overall κ	5.93×10^{9}	1.78×10^{5}	**3433**
Vision dom. channels	3842	5	**4417**
Language dom. channels	123	10468	**176**

Table 3. Comparison of per-layer online rotation overhead and quantization performance. QuaRot employs Kronecker-factorized Hadamard rotation ($H_{256} \otimes H_{43}$), whereas MABR uses modality-aware block-diagonal FWHT.

	QuaRot	**MABR (Ours)**
FLOPs / layer	561K	**127K**
Latency / layer	**56 μs**	237 μs
Storage / layer	**4.1 KB**	64.5 KB
W8A8 / W4A4 Avg.	65.5 / 51.4	**77.5 / 73.5**

5. Conclusion

We identify cross-modal heterogeneity as a fundamental obstacle to rotation-based post-training quantization in VLA models. In OpenVLA-7B, certain early layers exhibit a severe mismatch between vision and language activation statistics, making naive global rotation ineffective. To address this, we propose Modality-Aware Block Rotation (MABR), which preserves modality-specific channel structure while mitigating activation outliers and cross-modal contamination. Our approach significantly improves the stability of low-bit quantization. More broadly, our results highlight the importance of designing quantization methods that respect the inherent modality structure of VLAs, rather than inheriting assumptions from homogeneous LLMs.

Limitations Although we conduct an in-depth analysis of cross-modal heterogeneity in VLA models, this study is currently limited to OpenVLA, an autoregressive architecture. We plan to extend our method to flow-matching-based and hybrid architectures, such as π_0, GR00T N1, and other emerging VLA foundations. In addition, our evaluation is restricted to simulation environments, and we intend to validate our approach on real-robot platforms in future work.

References

An, Y., Zhao, X., Yu, T., Tang, M., and Wang, J. Systematic outliers in large language models. *arXiv preprint arXiv:2502.06415*, 2025.

Ashkboos, S., Mohtashami, A., Croci, M. L., Li, B., Cameron, P., Jaggi, M., Alistarh, D., Hoefler, T., and Hensman, J. Quarot: Outlier-free 4-bit inference in rotated llms. *Advances in Neural Information Processing Systems*, 37:100213–100240, 2024.

Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., Fang, Y., Fox, D., Hu, F., Huang, S., et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al. π_0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*, 2024.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jackson, T., Jesmonth, S., Joshi, N. J., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, K., Levine, S., Lu, Y., Malla, U., Manjunath, D., Mordatch, I., Nachum, O., Parada, C., Peralta, J., Perez, E., Pertsch, K., Quiambao, J., Rao, K., Ryoo, M. S., Salazar, G., Sanketi, P., Sayed, K., Singh, J., Sontakke, S., Stone, A., Tan, C., Tran, H. T., Vanhoucke, V., Vega, S., Vuong, Q., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., and Zitkovich, B. Rt-1: Robotics transformer for real-world control at scale. *CoRR*, abs/2212.06817, 2022.

Dong, Z., Yao, Z., Gholami, A., Mahoney, M. W., and Keutzer, K. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 293–302, 2019.

Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., et al. $\pi_{0.5}$: A Vision-Language-Action Model with Open-World Generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Li, Y., Gong, R., Tan, X., Yang, Y., Hu, P., Zhang, Q., Yu, F., Wang, W., and Gu, S. Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*, 2021.

Lin, H., Xu, H., Wu, Y., Cui, J., Zhang, Y., Mou, L., Song, L., Sun, Z., and Wei, Y. Duquant: Distributing outliers via dual transformation makes stronger quantized llms. *Advances in Neural Information Processing Systems*, 37: 87766–87800, 2024a.

Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.-M., Wang, W.-C., Xiao, G., Dang, X., Gan, C., and Han, S. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6:87–100, 2024b.

Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., and Stone, P. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

Liu, Z., Zhao, C., Fedorov, I., Soran, B., Choudhary, D., Krishnamoorthi, R., Chandra, V., Tian, Y., and Blankevoort, T. Spinquant: Llm quantization with learned rotations. *arXiv preprint arXiv:2405.16406*, 2024.

Shao, W., Chen, M., Zhang, Z., Xu, P., Zhao, L., Li, Z., Zhang, K., Gao, P., Qiao, Y., and Luo, P. Omniquant: Omnidirectionally calibrated quantization for large language models. *arXiv preprint arXiv:2308.13137*, 2023.

Team, O. M., Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Kreiman, T., Xu, C., et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

Wang, C., Wang, Z., Xu, X., Tang, Y., Zhou, J., and Lu, J. Q-vlm: Post-training quantization for large vision-language models. *Advances in Neural Information Processing Systems*, 37:114553–114573, 2024.

Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., and Han, S. Smoothquant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099, 2023.

Xu, S., Wang, T., Li, F., Zhu, L., and Shen, H. T. Da-ptq: Drift-aware post-training quantization for efficient vision-language-action models. *arXiv preprint arXiv:2604.11572*, 2026a.

Xu, Y., Yang, Y., Fan, Z., Liu, Y., Li, Y., Li, B., and Zhang, Z. Qvla: Not all channels are equal in vision-language-action model's quantization. *arXiv preprint arXiv:2602.03782*, 2026b.

Zhang, J., Hsieh, Y., Wan, Z., Lin, H., Wang, X., Wang, Z., Lei, Y., and Zhang, M. Quantvla: Scale-calibrated post-training quantization for vision-language-action models. *arXiv preprint arXiv:2602.20309*, 2026.

Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

A. Preliminaries

A.1. Post-Training Quantization

Post-training quantization (PTQ) compresses a pretrained neural network by representing weights and activations with low-bit integers without additional retraining. In uniform symmetric quantization, floating-point values are mapped onto uniformly spaced fixed-point grids:

$$\mathcal{Q}_b^{\text{sym}} = \Delta \times \left\{-2^{b-1}, \ldots, 0, \ldots, 2^{b-1}-1\right\}, \quad (11)$$

where b denotes the bit-width and Δ is the quantization step size, also referred to as the scaling factor.

Given a floating-point tensor $\mathbf{X}$, the scaling factor is typically determined by

$$\Delta = \frac{\max(|\mathbf{X}|)}{2^{b-1}-1}. \quad (12)$$

The quantization function $q(\cdot) : \mathbb{R} \to \mathcal{Q}_b^{\text{sym}}$ maps floating-point values into integer grids:

$$\mathbf{X}_q = \text{clip}\left(\left\lfloor \frac{\mathbf{X}}{\Delta} \right\rceil, -2^{b-1}, 2^{b-1}-1\right), \quad (13)$$

where $\lfloor \cdot \rceil$ denotes round-to-nearest quantization and $\text{clip}(\cdot)$ truncates values outside the representable range.

During inference, dequantization reconstructs the approximate floating-point tensor:

$$\hat{\mathbf{X}} = \Delta \cdot \mathbf{X}_q. \quad (14)$$

A.2. Hessian-Aware Quantization

Quantization can be interpreted as a weight perturbation $\delta\mathbf{W}$. Via second-order Taylor expansion, the induced loss degradation is:

$$\mathbb{E}[\mathcal{L}(\mathbf{W}+\delta\mathbf{W})] - \mathbb{E}[\mathcal{L}(\mathbf{W})] \approx \frac{1}{2}\delta\mathbf{W}^\top \bar{\mathbf{H}} \delta\mathbf{W}, \quad (15)$$

where $\bar{\mathbf{H}} = \mathbb{E}[\nabla^2_{\mathbf{W}}\mathcal{L}]$ is the Hessian matrix. Since pretrained models are near local minima, the gradient term is negligible, making the Hessian the dominant factor governing quantization sensitivity. This motivates Hessian-aware methods such as GPTQ.

A.3. Hadamard Transformations

Rotation-based PTQ methods employ orthogonal transformations to redistribute activation outliers before quantization. An orthogonal matrix $\mathbf{Q}$ satisfies $\mathbf{Q}\mathbf{Q}^\top = \mathbf{I}$, preserving the original computation in full precision: $\mathbf{X}\mathbf{Q} \cdot \mathbf{Q}^\top\mathbf{W} = \mathbf{X}\mathbf{W}$. Walsh-Hadamard matrices are recursively defined as:

$$\mathbf{H}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}, \qquad \mathbf{H}_{2^n} = \mathbf{H}_2 \otimes \mathbf{H}_{2^{n-1}}, \quad (16)$$

enabling $O(d\log_2 d)$ matrix-vector products via the Fast Walsh-Hadamard Transform (FWHT). Randomized variants $\tilde{\mathbf{H}} = \mathbf{H}\,\text{diag}(\mathbf{s})$, where $\mathbf{s} \in \{+1, -1\}^d$, remain orthogonal and further improve outlier smoothing.

A.4. Baseline

A.4.1. QuaRot

QuaRot mitigates activation outliers by applying an orthogonal rotation R to activations and absorbing it into the weight matrix:

$$XW^\top = (XR)(WR)^\top. \quad (17)$$

In practice, QuaRot employs randomized Walsh-Hadamard transformations to redistribute activation energy across channels and produce a more quantization-friendly distribution. The base Hadamard matrix is recursively defined as

$$H_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}, \qquad H_{2^n} = H_2 \otimes H_{2^{n-1}}, \quad (18)$$

where $\otimes$ denotes the Kronecker product.

For non-power-of-2 hidden dimensions, QuaRot constructs a Kronecker-factored rotation matrix. For example, the OpenVLA intermediate dimension D=11,008 is factorized as

$$11{,}008 = 256 \times 43, \quad (19)$$

yielding the rotation

$$R = H_{256} \otimes H_{43}. \quad (20)$$

This decomposition enables efficient approximate full-channel mixing through a dense 43×43 matrix multiplication followed by a Fast Walsh-Hadamard Transform (FWHT) of length 256.

A.4.2. GPTQ

GPTQ is a post-training quantization method that leverages the inverse Hessian H^{-1} to compensate quantization error. It quantizes weights column-wise and updates the remaining columns as:

$$W_{:,c'} \leftarrow W_{:,c'} - \frac{W_{:,c} - Q(W_{:,c})}{[H^{-1}]_{c,c}}[H^{-1}]_{c,c'}. \quad (21)$$

This redistributes quantization error according to second-order curvature, reducing overall reconstruction loss.

B. Derivation of Visual Signal Preservation Condition

We derive the condition under which low-magnitude visual signals vanish after global rotation and low-bit quantization.

Let x_v and x_l denote the typical magnitudes of vision and language activations, respectively, where

$$x_v \ll x_l. \tag{22}$$

Given a dense orthogonal rotation matrix $\mathbf{R} \in \mathbb{R}^{d\times d}$, global rotation approximately distributes activation energy uniformly across dimensions. Therefore, each rotated channel becomes dominated by the language activation:

$$\tilde{x}_i \approx \frac{x_l}{\sqrt{d}}. \tag{23}$$

Under symmetric uniform b-bit quantization, the quantization step size is

$$\Delta = \frac{\max(|\tilde{\mathbf{x}}|)}{2^{b-1}-1}. \tag{24}$$

Since the rotated channels are dominated by language activations, the quantization step size becomes

$$\max(|\tilde{\mathbf{x}}|) \approx \frac{x_l}{\sqrt{d}}, \qquad \Delta \approx \frac{x_l/\sqrt{d}}{2^{b-1}-1}. \tag{25}$$

For the visual component to survive round-to-nearest quantization, its magnitude after rotation must exceed half of a quantization bin:

$$\frac{x_v}{\sqrt{d}} \geq \frac{\Delta}{2} = \frac{1}{2}\cdot\frac{x_l/\sqrt{d}}{2^{b-1}-1}. \tag{26}$$

Finally,

$$x_v \geq \frac{x_l}{2^b-2}. \tag{27}$$

Therefore, if

$$x_v < \frac{x_l}{2^b-2}, \tag{28}$$

the visual signal collapses into the zero quantization bin after rotation and quantization.

C. Additional Details on Runtime Overhead

C.1. Block-Diagonal FWHT in MABR

Unlike QuaRot, MABR avoids dense global mixing across modalities to prevent cross-modal contamination. Instead, we partition the sorted channels into power-of-two sized blocks to enable efficient Fast Walsh-Hadamard Transform (FWHT) computation while preserving modality boundaries.

Block Construction Logic. Let $|\mathcal{C}_V|$ and $|\mathcal{C}_L|$ denote the number of vision and language-dominant channels identified by the dominance ratio $r(c)$. To ensure modality-consistent smoothing, we first assign these dominant groups to their own blocks. Since FWHT requires power-of-two dimensions, each group is rounded up to the nearest power of 2:

$$n_V = 2^{\lceil \log_2 |\mathcal{C}_V| \rceil}, \qquad n_L = 2^{\lceil \log_2 |\mathcal{C}_L| \rceil}. \tag{29}$$

The remaining $D_{balanced} = D - n_V - n_L$ channels are greedily decomposed into the largest possible power-of-2 sub-blocks $\{b_1, b_2, \ldots, b_k\}$. The final block-diagonal rotation matrix is defined as:

$$R_{sorted} = \mathrm{diag}\left(H_{n_V}, H_{b_1}, \ldots, H_{b_k}, H_{n_L}\right). \tag{30}$$

Numerical Example for OpenVLA. For the OpenVLA intermediate layer with $D = 11,008$, the blocks are derived as follows:

1. **Dominant Groups:** Based on our thresholding, we identify $|\mathcal{C}_V| = 2,584$ and $|\mathcal{C}_L| = 98$ for layer 1 as illustrated in Fig. 1 (a).
2. **Rounding to Power-of-2:**
   - Vision block: $n_V = 2^{\lceil \log_2 2584 \rceil} = 2^{12} = 4,096$.
   - Language block: $n_L = 2^{\lceil \log_2 98 \rceil} = 2^7 = 128$.
3. **Greedy Decomposition of Remaining Subspace:**The remaining $11,008 - (4,096 + 128) = 6,784$ channels are decomposed into:
   - $b_1 = 4,096$ (remaining $2,688$)
   - $b_2 = 2,048$ (remaining 640)
   - $b_3 = 512$ (remaining 128)
   - $b_4 = 128$ (remaining 0)

The resulting block structure is $[4096, 4096, 2048, 512, 128, 128]$, totaling exactly $11,008$. This block-wise formulation substantially reduces arithmetic complexity:

$$\sum_i d_i \log_2 d_i \ll D \log_2 D, \tag{31}$$

yielding $4.4\times$ fewer FLOPs than QuaRot's Kronecker-factored rotation ($H_{256} \otimes H_{43}$) in our implementation.

D. Related Work

D.1. Vision-Language-Action Models

Vision-Language-Action (VLA) models unify perception, language reasoning, and action generation within a single policy architecture. Existing approaches can largely be categorized by their action generation strategy. Autoregressive VLA models such as RT-2, OpenVLA (Zitkovich et al., 2023; Kim et al., 2024) formulate robot control as token prediction, leveraging large language or vision-language backbones for open-vocabulary reasoning and semantic grounding. In contrast, continuous-action approaches based on diffusion or flow matching, including Octo, π_0 (Team et al., 2024; Black et al., 2024) generate temporally smooth trajectories directly in continuous action space, improving control fidelity and robustness. Recent hybrid architectures further combine language reasoning with diffusion-based control, exemplified by GR00T N1 and OPENPI $\pi_{0.5}$ (Bjorck et al., 2025; Intelligence et al., 2025).

D.2. Post-Training Quantization

Post-training quantization (PTQ) has been widely studied for efficient deployment of large neural networks. Weight-only approaches such as GPTQ (Frantar et al., 2022), AWQ (Lin et al., 2024b), and HAWQ (Dong et al., 2019) reduce memory usage while preserving high accuracy, whereas weight-activation quantization methods additionally quantize activations to improve end-to-end inference efficiency.

Recent PTQ methods focus on mitigating activation outliers, which become particularly problematic under low-bit quantization. SmoothQuant (Xiao et al., 2023) reduces activation difficulty through channel-wise activation-to-weight scale migration, while OmniQuant (Shao et al., 2023) jointly optimizes quantization parameters for weights and activations. Rotation-based approaches such as QuaRot (Ashkboos et al., 2024), SpinQuant (Liu et al., 2024), and DuQuant (Lin et al., 2024a) further redistribute outliers through orthogonal transformations, enabling robust 4-bit quantization in LLMs. In parallel, Hessian-aware methods including HAWQ, GPTQ, and BRECQ (Li et al., 2021) exploit second-order curvature information to reduce reconstruction error during quantization.

D.3. VLA Quantization

Recent works have begun exploring quantization for vision-language-action (VLA) and multimodal embodied transformers. QVLA (Xu et al., 2026b) highlights channel-wise sensitivity differences in VLA models and introduces channel-aware allocation strategies for OpenVLA. QuantVLA (Zhang et al., 2026) improves robustness through scale calibration for multimodal activations, while DA-PTQ (Xu et al., 2026a) addresses distribution shift during post-training compression of VLA backbones. Beyond robotics, Q-VLM (Wang et al., 2024) studies post-training quantization for vision-language transformers and shows that multimodal models exhibit significantly different sensitivity patterns from language-only LLMs.

However, existing VLA quantization methods mainly operate in moderate precision regimes such as W8A8 and W4A8, or rely on calibration heuristics without explicitly addressing cross-modal interference inside shared transformer layers. Furthermore, while QVLA demonstrates W4A4 quantization, it requires mixed-precision allocation, which complicates efficient deployment on practical hardware accelerators.

In contrast, our work targets aggressive uniform W4A4 deployment without relying on mixed precision, and identifies cross-modal heterogeneity between vision, language, and action representations as the fundamental obstacle to stable low-bit quantization in VLA models. Our method addresses this challenge through modality-aware block rotation that preserves modality-specific structure while still benefiting from activation smoothing.

E. Qualitative Results

Qualitative rollout videos are provided through the project page at anonymous project page. The project page is hosted anonymously and contains no author-identifying information. The page is designed for side-by-side comparison of FP16, global rotation, and MABR results.

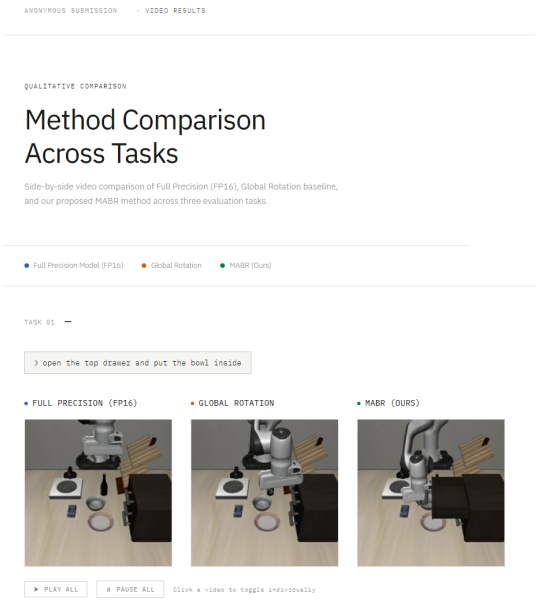

Figure 4. Anonymous project page for side-by-side video comparison of FP16, Global Rotation, and MABR (Ours).