



清华大学电子工程系

Department of Electronic Engineering, Tsinghua University

MBQ: Modality-Balanced Quantization for Large Vision-Language Models

Shiyao Li^{1,2*}, Yingchun Hu^{2,3*}, Xuefei Ning¹⁺, Xihui Liu⁴, Ke Hong^{1,3}, Xiaotao Jia³⁺, Xiuhong Li²,
Yaqi Yan⁵, Pei Ran⁵, Guohao Dai^{6,2}, Shengen Yan², Huazhong Yang¹, Yu Wang¹⁺

¹Tsinghua University ²Infinigence-AI ³Beihang University

⁴University of Hong Kong ⁵ChinaTower ⁶Shanghai Jiao Tong University





Contents

- 1 Background
- 2 Method Design
- 3 Experiments

Background



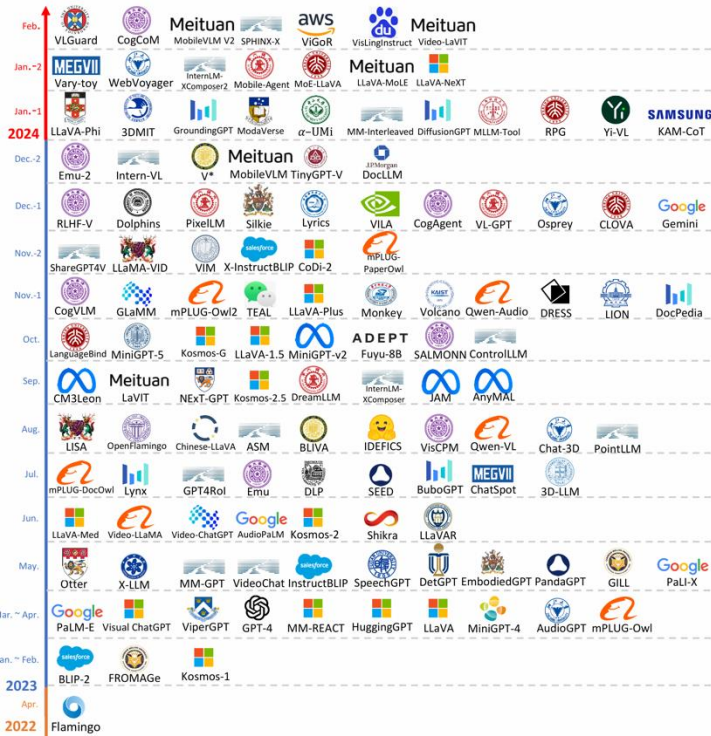
- In recent years, large vision-language models have gained great improvements and enabled various applications.

Vision Language Models

GPT-4o, Gemini, LLaVA-OneVision, InternVL...

Human-Computer Interaction

Autonomous Driving

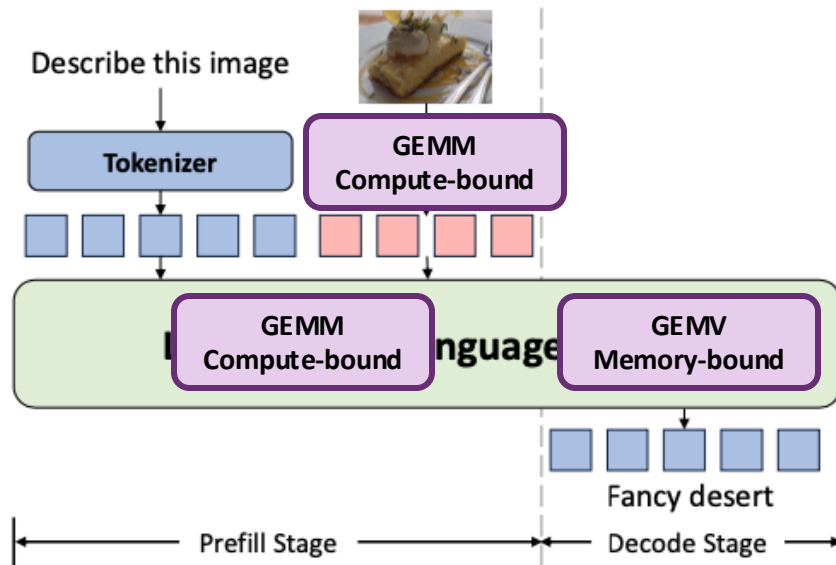


[1] MM-LLMs: Recent Advances in MultiModal Large Language Models (Tencent AI Lab, 2024)

Background



- The inference process of VLMs is similar to that of LLMs. VLMs also have prefill and decode stages.
 - **Language Tokenizer**: Transform natural language sentences into a series of language tokens
 - **ViT Encoder**: Transform images into a series of vision tokens
 - **Large VLM**: Take the language and vision tokens as input (Prefill Stage), and generate language tokens (Decode Stage).

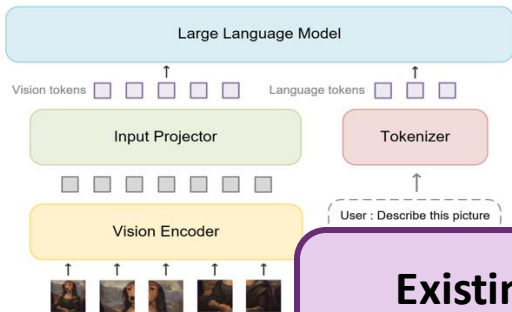


Background

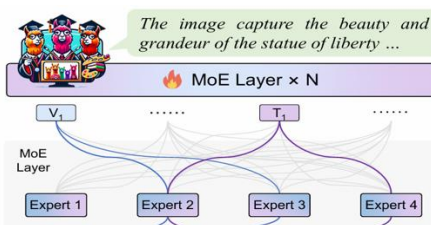


- To accelerate the inference speed and reduce the memory overhead. Various efficiency methods have been developed.

Transformer-based Efficient VLMs



Non-Transformer Architectures



Token Pruning & Merging



Existing Methods do not focus on the traditional compression methods, such as Quantization, Pruning...

- Smaller Vision Encoders:** Train smaller ViTs (~100M level) with MoE instead of ViTs with Large Qformers (~1B level)
- Smaller VLMs:** Develop advanced training skills, such as distillation, Vision-language Supervised FineTuning (SFT), and so on.

Need Training

- knowledge of different domains, only activate a few experts during inference.
- Mamba-based VLMs:** Introduce Mamba as the backbone, significantly reduce the computation overhead of the attention blocks.

Need Training

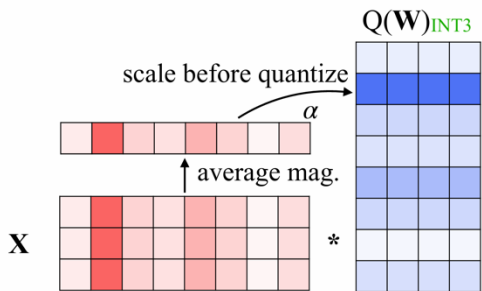
- Compression:** Prune or Merge the similar image tokens before input VLMs
- Dynamic Token Compression:** Dynamically merge the token features in the KV cache during prefill and decode stages.

Reduce Computation overhead in Attention

Background



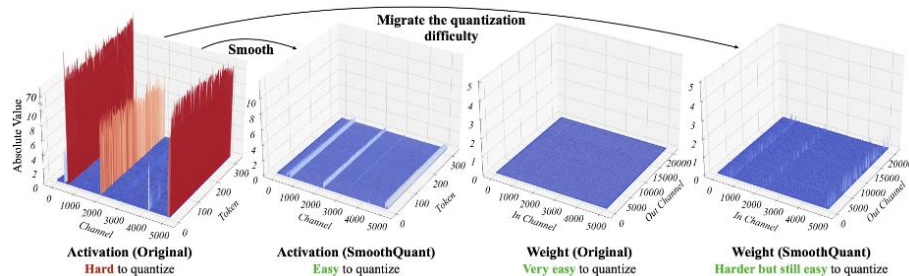
- Quantization has been widely used to reduce memory and computation overhead of LLMs.
- By dealing with massive outlier values in LLMs, SOTA methods like AWQ and SmoothQuant can achieve lossless W4 and W8A8 Quantization.
 - They search for **channel-wise equalization (CWE)** factors for better quantization by minimizing the transformer block-wise reconstruction loss.



$$E^* = \underset{E}{\operatorname{argmin}} \|Q(W * E)(E^{-1} * X) - WX\|^2,$$

OPT / PPL↓		125M	1.3B	2.7B	6.7B	13B	30B	66B
FP16	-	31.95	16.41	14.32	12.29	11.5	10.67	10.09
INT4 g128	RTN	35.51	17.70	15.12	13.02	11.89	11.00	10.44
	GPTQ	34.23	16.92	14.69	12.51	11.60	10.74	10.24
	AWQ	33.96	16.85	14.61	12.44	11.60	10.75	10.16

Weight-only Quantization (AWQ)



$$E^* = \underset{E}{\operatorname{argmin}} \|Q(W * E)Q(E^{-1} * X) - WX\|^2,$$

OPT-175B	LAMBADA	HellaSwag	PIQA	WinoGrande	OpenBookQA	RTE	COPA	Average↑	WikiText↓
FP16	74.7%	59.3%	79.7%	72.6%	34.0%	59.9%	88.0%	66.9%	10.99
W8A8	0.0%	25.6%	53.4%	50.3%	14.0%	49.5%	56.0%	35.5%	93080
ZeroQuant	0.0%*	26.0%	51.7%	49.3%	17.8%	50.9%	55.0%	35.8%	84648
LLM.int8()	74.7%	59.2%	79.7%	72.1%	34.2%	60.3%	87.0%	66.7%	11.10
Outlier Suppression	0.00%	25.8%	52.5%	48.6%	16.6%	53.4%	55.0%	36.0%	96151
SmoothQuant-O1	74.7%	59.2%	79.7%	71.2%	33.4%	58.1%	89.0%	66.5%	11.11
SmoothQuant-O2	75.0%	59.0%	79.2%	71.2%	33.0%	59.6%	88.0%	66.4%	11.14
SmoothQuant-O3	74.6%	58.9%	79.7%	71.2%	33.4%	59.9%	90.0%	66.8%	11.17

Weight-Activation Quantization (SmoothQuant)

Background



- However, directly applying the SOTA quantization methods to VLMs causes significant accuracy loss.
- On MMMU, the performance of AWQ/SmoothQuant is even worse than naive RTN.

Model	Bitwidth	Method	MMMU
LLaVA-onevision-7B	FP16	--	46.0
	W3	RTN	34.7
		AWQ	36.6
		RTN	33.9
LLaVA-onevision-70B	W3	RTN	33.9
		AWQ	33.4
	W4A8	RTN	54.8
		SmoothQuant	51.6

What is the core reason that causes existing quantization methods to fail for VLMs?



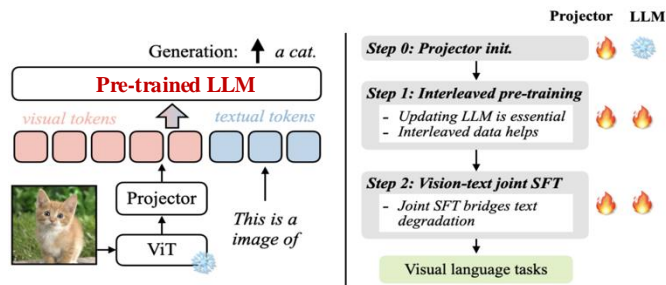
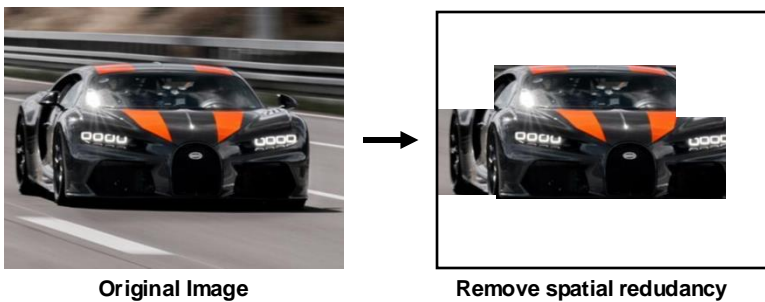
Contents

- 1 Background
- 2 Method Design**
- 3 Experiments

Sensitivity Varies Across Modalities



- Intuitively, we speculate that the significant performance degradation when applying SOTA LLM quantization methods to VLMs stems from **treating different modalities equally**.
- This is because **errors in vision tokens might have a smaller impact on the output context** compared to introducing the same errors in language tokens.
- **From the data perspective:** Visual data generally contains a high degree of redundancy and might be more fault tolerant for small perturbations.
- **From the model perspective:** Zhang et al. [1] discover that the generated content of current VLMs is primarily biased by the pre-trained LLMs rather than the input image.
 - Perhaps it is because the VLM has encountered more language data during training.

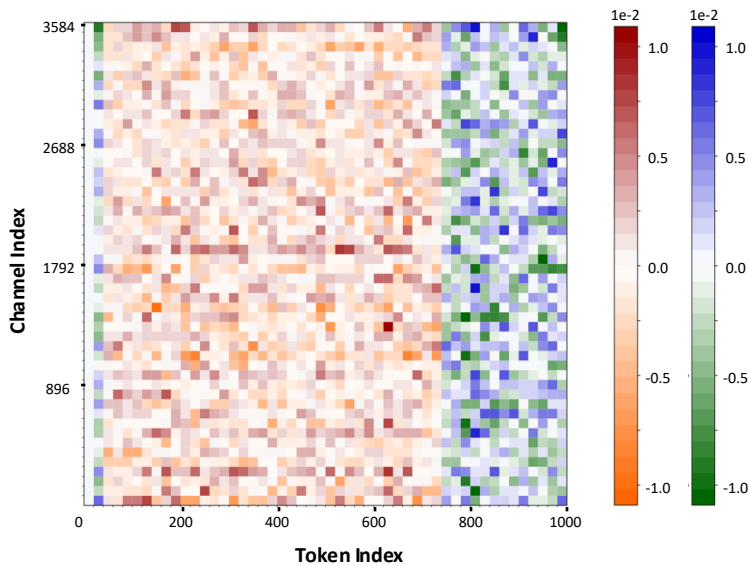


[1] Yi-Fan Zhang, et al. Debiasing multimodal large language models. CoRR, 2024.

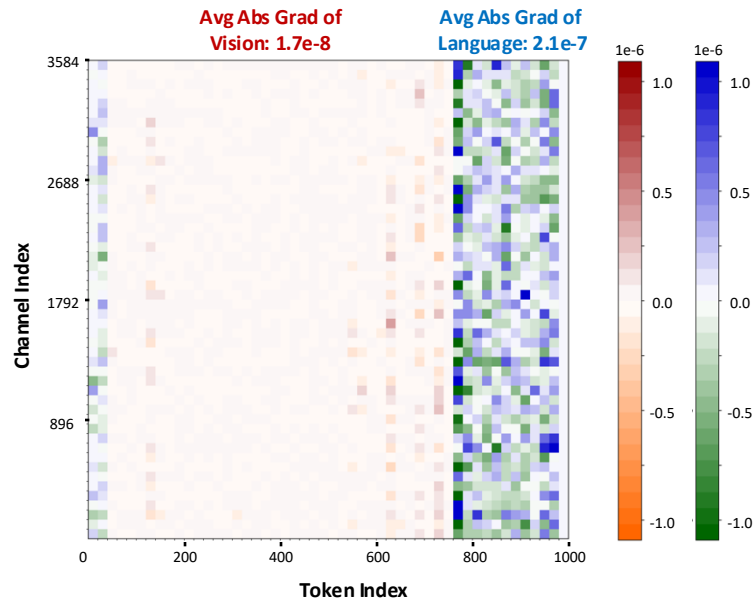
Sensitivity Varies Across Modalities



- Key observation:
 - The activation values of vision and language tokens are similar.
 - However, the average absolute gradient of language tokens is 10x larger than that of vision tokens.
For a similar perturbation, a language token's impact on the loss function might be 10x that of a vision token.



The **data distribution** of the output activation of the 13th transformer block in the LLaVA-onevision-7B



The **gradient distribution** of the output activation of the 13th transformer block in the LLaVA-onevision-7B

Sensitivity Varies Across Modalities



- If we treat language and vision tokens equally (like AWQ and SmoothQuant), we will **miss the opportunity to leverage the VLM's lower sensitivity to vision tokens to achieve higher performance**
 - The naïve CWE optimization objective:

$$\mathbf{E}^* = \operatorname{argmin}_{\mathbf{E}} \|Q(\mathbf{W} * \mathbf{E})(\mathbf{E}^{-1} * \mathbf{X}) - \mathbf{W}\mathbf{X}\|^2,$$

- We propose balanced CWE, by applying a modality-balancing factor of 0.1 to the vision tokens' reconstruction loss:

$$\mathbf{E}^* = \operatorname{argmin}_{\mathbf{E}} [\|Q(\mathbf{W} * \mathbf{E})(\mathbf{E}^{-1} * \mathbf{X}_l) - \mathbf{W}\mathbf{X}_l\|^2 + 0.1\|Q(\mathbf{W} * \mathbf{E})(\mathbf{E}^{-1} * \mathbf{X}_v) - \mathbf{W}\mathbf{X}_v\|^2],$$

Model	FP16	W3		
		RTN	CWE	Balanced CWE
LLaVA-ov-7B	46.00	34.67	36.56	40.22
InternVL2-8B	48.00	43.67	44.78	46.33

Table 1. The accuracy results (%) on MMMU benchmark after W3 quantization. The calibration dataset is selected from COCO caption dataset [10]. The LLaVA-ov-7B is short for LLaVA-onevision-7B.

* CWE is short for Channel-wise Equalization.

Modality-Balanced Quantization (MBQ)



- **Modality-Balanced Quantization**

- Given that the sensitivity differences between vision and language tokens may **vary across layers and VLM families**, exploring an **automatic modality-balancing approach** could further enhance the performance of the quantized VLMs.
- We aim to derive an approach for allocating the optimal Modality-Balanced factors to each layer by directly minimizing the change in the SFT loss function:

$$L(\mathbf{Y} + \Delta) \simeq L(\mathbf{Y}) + \mathbf{g}^T * \Delta,$$

- The change in SFT loss caused by quantization can be expressed as the following equation:

$$\begin{aligned} \|L(\hat{\mathbf{Y}})\| &\simeq \|\mathbf{g}^T * \Delta\| \\ &\leq |\bar{\mathbf{g}}| * \|\Delta\| \\ &\simeq |\bar{\mathbf{g}}_v| * \|\Delta_v\| + |\bar{\mathbf{g}}_l| * \|\Delta_l\| \\ &= |\bar{\mathbf{g}}_v| * \|\mathbf{Y}_v - \hat{\mathbf{Y}}_v\| + |\bar{\mathbf{g}}_l| * \|\mathbf{Y}_l - \hat{\mathbf{Y}}_l\|, \end{aligned}$$

Modality-Balanced Quantization (MBQ)



- **Modality-Balanced Quantization**

- In order to accelerate the prefill stage of VLMs, we quantize both the weights and input activations of each linear layer to leverage fast low-precision tensor cores. The objective is shown in the following:

$$\min_{\mathbf{E}} [|\bar{\mathbf{g}}_v| * \|\mathbf{W}\mathbf{X}_v - Q(\mathbf{W} * \mathbf{E})Q(\mathbf{E}^{-1} * \mathbf{X}_v)\| + |\bar{\mathbf{g}}_l| * \|\mathbf{W}\mathbf{X}_l - Q(\mathbf{W} * \mathbf{E})Q(\mathbf{E}^{-1} * \mathbf{X}_l)\|],$$

- To accelerate the decode stage of VLMs, we only quantize weights to reduce the memory overhead, aiming to minimize the balanced reconstruction error by the following objective:

$$\min_{\mathbf{E}} [|\bar{\mathbf{g}}_v| * \|\mathbf{W}\mathbf{X}_v - Q(\mathbf{W} * \mathbf{E})(\mathbf{E}^{-1} * \mathbf{X}_v)\| + |\bar{\mathbf{g}}_l| * \|\mathbf{W}\mathbf{X}_l - Q(\mathbf{W} * \mathbf{E})(\mathbf{E}^{-1} * \mathbf{X}_l)\|],$$

- Notably, unlike directly using the heuristically selected MSE-based balanced CWE loss, our derived reconstruction error loss function relies on Mean Absolute Error (MAE). Our ablation also demonstrates that minimizing MAE-based reconstruction loss in MBQ yields better results.



Contents

- 1 Background
- 2 Method Design
- 3 Experiments**

Experiments



- **Main results on various benchmarks and 3 VLM families.**
 - For weight-only quantization, the proposed MBQ can significantly outperform the AWQ baseline across different families. Especially within the LLaVA-onevision family.
 - For Weight-activation quantization, MBQ can significantly outperform SmoothQuant, with improvements of up to 11.6%.

Model	Bitwidth	Method	MMMU	SEED	OCRBench	VizWiz	ScienceQA	TextVQA	Average (↑)
LLaVA-onevision-7B	FP16	-	46.0	71.1	62.2	60.4	85.4	76.1	66.9
		RTN	34.7	10.4	35.9	59.2	86.2	60.9	47.9
	W3	AWQ	36.6	51.5	59.3	58.5	83.2	73.0	60.4
		MBQ	42.0	66.4	61.1	60.7	85.0	73.3	64.8
	W4A8	RTN	38.2	48.4	40.1	58.2	88.3	61.5	55.8
		MBQ	42.6	64.4	52.3	58.9	88.5	68.3	62.5
InternVL2-8B	FP16	-	48.0	71.6	76.5	61.1	96.2	77.0	71.7
		RTN	43.7	70.3	74.0	56.0	95.6	74.6	69.0
	W3	AWQ	44.8	70.4	74.7	58.9	95.5	74.2	69.8
		MBQ	46.9	70.8	75.1	58.7	95.6	75.1	70.4
	W4A8	RTN	44.3	69.5	72.0	57.1	95.5	73.1	68.6
		MBQ	45.6	69.9	73.0	56.5	95.8	72.3	68.9
Qwen2-VL-7B	FP16	-	50.6	71.9	80.7	68.3	85.1	82.0	73.1
		RTN	44.9	69.8	60.0	65.2	81.5	71.2	65.4
	W3	AWQ	44.7	70.4	76.9	68.0	82.5	79.5	70.3
		MBQ	47.9	70.2	76.8	67.7	82.8	79.9	70.9
	W4A8	RTN	43.8	70.1	60.3	58.9	78.9	71.0	63.8
		MBQ	47.2	70.3	72.8	59.3	81.2	75.0	67.6

7B-level VLMs

Model	Bitwidth	Method	MMMU	SEED	OCRBench	VizWiz	ScienceQA	TextVQA	Average (↑)
LLaVA-onevision-72B	FP16	-	56.1	78.1	73.2	69.2	90.0	79.3	74.3
		RTN	53.9	77.4	68.2	66.1	89.5	77.4	72.1
	W3	AWQ	33.4	71.2	48.7	49.3	69.2	58.8	55.1
		MBQ	54.4	77.6	71.6	69.0	90.3	78.5	73.6
	W4A8	RTN	54.8	76.6	64.5	64.7	89.0	74.5	70.7
		MBQ	55.6	76.5	64.4	65.7	89.2	73.3	70.8
InternVL2-26B	FP16	-	47.1	76.8	77.9	66.2	97.5	82.1	74.6
		RTN	46.6	75.7	75.9	64.7	96.4	80.6	73.3
	W3	AWQ	46.4	76.2	76.4	64.5	96.7	81.0	73.5
		MBQ	47.1	76.3	76.5	64.5	97.3	81.1	73.8
	W4A8	RTN	44.7	76.0	76.4	62.6	96.7	79.6	72.7
		MBQ	44.0	75.7	77.5	62.0	97.1	80.0	72.7
Qwen2-VL-72B	FP16	-	61.1	77.6	79.9	76.0	91.6	82.5	78.1
		RTN	57.7	77.5	70.4	74.8	89.7	79.7	75.0
	W3	AWQ	59.6	77.6	79.6	75.4	90.4	82.4	77.5
		MBQ	59.6	77.7	79.4	75.6	90.5	82.5	77.6
	W4A8	RTN	58.1	76.6	66.2	71.3	90.1	77.0	73.2
		MBQ	57.7	76.3	77.5	73.6	89.6	80.5	75.8

26B & 70B VLMs

Experiments



- **The Effect of calibration dataset**

- Directly apply the vision-language dataset as calibration **does not consistently improve the performance of the quantized VLMs**, especially **W4A8**.
- We speculate that this is because weight-activation quantization requires considering both weight and activation quantization errors.

BitWidth	Method	Components				MMMU (↑)	SEED (↑)
		Channel-wise Equalization	COCO Calib.	Modality-Balance (MSE)	Modality-Balance (MAE)		
FP16	-	-	-	-	-	46.0	71.1
W3	RTN	✗	✗	✗	✗	34.7	10.4
	AWQ	✓	✗	✗	✗	36.6	51.5
	-	✓	✓	✗	✗	38.7	61.8
	-	✓	✓	✓	✗	40.8	64.8
	MBQ	✓	✓	✗	✓	42.0	66.4
W4A8	RTN	✗	✗	✗	✗	38.2	48.4
	SQ	✓	✗	✗	✗	30.9	41.6
	-	✓	✓	✗	✗	29.2	10.2
	-	✓	✓	✓	✗	41.9	63.5
	MBQ	✓	✓	✗	✓	42.6	64.4

Table 3. The ablation study on LLaVA-onevision-7B with W3 and W4A8 quantization. “SQ” is short for SmoothQuant.

Experiments



- **The Effect of Modality-Balance**

- Modality-Balancing plays a crucial role in weight-activation quantization and can also improve the performance of weight-only quantization.
- Modality-Balancing (MAE) can consistently outperform Modality-Balancing (MSE) in both weight-only and weight-activation quantization.

BitWidth	Method	Components				MMMU (\uparrow)	SEED (\uparrow)
		Channel-wise Equalization	COCO Calib.	Modality-Balance (MSE)	Modality-Balance (MAE)		
FP16	-	-	-	-	-	46.0	71.1
W3	RTN	✗	✗	✗	✗	34.7	10.4
	AWQ	✓	✗	✗	✗	36.6	51.5
	-	✓	✓	✗	✗	38.7	61.8
	-	✓	✓	✓	✗	40.8	64.8
	MBQ	✓	✓	✗	✓	42.0	66.4
W4A8	RTN	✗	✗	✗	✗	38.2	48.4
	SQ	✓	✗	✗	✗	30.9	41.6
	-	✓	✓	✗	✗	29.2	10.2
	-	✓	✓	✓	✗	41.9	63.5
	MBQ	✓	✓	✗	✓	42.6	64.4

Table 3. The ablation study on LLaVA-onevision-7B with W3 and W4A8 quantization. “SQ” is short for SmoothQuant.

Experiments



- **The performance on language-only benchmark**

- The main idea of the proposed MBQ is to consider the sensitivity across different modalities during quantization, **aiming to enhance performance in both vision-language and language-only tasks**.
- Accordingly, we evaluated the performance of the quantized LLaVA-onevision-7B VLM on the MMLU benchmark with different quantization methods.
- With W3 and W4A8 quantization, MBQ achieves a performance improvement of 0.9% and 2%, compared to AWQ and SmoothQuant.

BitWidth	Method	MMLU (\uparrow)
FP16	-	65.9
W3	AWQ (Pile)	61.9
	AWQ (COCO)	62.0
	MBQ	62.9
W4A8	SQ (Pile)	59.8
	SQ (COCO)	59.0
	MBQ	61.8

Table 5. The results of quantized LLaVA-onevision-7B on the MMLU benchmark. “SQ” is short for SmoothQuant.

Experiments



- **The performance on open-ended generation tasks**

- To evaluate the open-ended conversational ability of the quantized VLMs, we evaluate the proposed MBQ alongside SOTA baselines on the LLaVA-onevision family using conversation benchmarks.
- We manually evaluated the responses of the quantized VLM to each question and identified the following five frequently occurring error types:
 - **No output:** The quantized VLM generates no or only a few valid tokens;
 - **Randomness:** The quantized VLM randomly generates meaningless symbols;
 - **Repetition:** The quantized VLM keeps repeating some certain tokens;
 - **Condition Missing:** The quantized VLM misses the keypoints in the questions;
 - **Semantic Error:** The quantized VLM can understand the questions but still generates wrong answers with meaningful and fluent language.

Model	Bitwidth	Method	Error Types					Total
			No Output	Randomness	Repetition	Condition Missing	Semantic Error	Bad Cases
LLaVA-onevision-7B	W4A8	SQ	61	16	15	32	25	149
		MBQ	3	0	2	3	37	45
LLaVA-onevision-72B	W3	AWQ	30	0	3	11	42	86
		MBQ	0	0	1	0	37	38

Table 9. The number of samples for the five error types in the LLaVA-bench-in-the-wild [29] and LLaVA-bench-wilder [19] dataset. The total number of samples is **188**. “SQ” is short for SmoothQuant.

Experiments



- **The performance on open-ended generation tasks**

- For the LLaVA-onevision-7B with W4A8 quantization, **MBQ results in only 45 total bad cases, which is 104 fewer than the 149 bad cases observed with SmoothQuant.**
 - While MBQ leads to more “Semantic Error” cases than SmoothQuant, this is because SmoothQuant often produces outputs that are not even fluently generated, whereas MBQ enables the quantized VLM to generate fluent responses, thereby increasing the “Semantic Error”.
- For the LLaVA-onevision-72B with W3 quantization, a similar trend is observed as with the LLaVA-onevision-7B using W4A8 quantization. **MBQ results in only 38 bad cases, which is 48 fewer than the 86 bad cases reported for AWQ.**

Model	Bitwidth	Method	Error Types					Total
			No Output	Randomness	Repetition	Condition Missing	Semantic Error	Bad Cases
LLaVA-onevision-7B	W4A8	SQ	61	16	15	32	25	149
		MBQ	3	0	2	3	37	45
LLaVA-onevision-72B	W3	AWQ	30	0	3	11	42	86
		MBQ	0	0	1	0	37	38

Table 9. The number of samples for the five error types in the LLaVA-bench-in-the-wild [29] and LLaVA-bench-wilder [19] dataset. The total number of samples is **188**. “SQ” is short for SmoothQuant.

Experiments



- The performance on open-ended generation tasks

Example 5. LLaVA-onevision-72B-W3

Question: Imagine the fragrance of the fruits in the image. How would you describe this to someone who has never had this fruit before?



FP16 Response: The fragrance of the mangosteen is a delightful blend of sweet and slightly tart notes, reminiscent of a combination of strawberries, kiwi, and pineapple. It's a fresh and inviting aroma that hints at the juicy and flavorful experience one is about to encounter when tasting the fruit.

AWQ Response: The **mangos** fruit in the image has a unique fragrance that is both sweet and slightly tart. The sweetness is reminiscent to a combination of honey and a hint of citrus, while the tartness comes from a subtle tang of earthy undertones. When you bite into it, the initial sweetness gives way to a tangy, lingering taste that lingers on your tongue, leaving you with a pleasant and refreshing aftertaste.

**** Error Type: Semantic Error. The quantized VLM mistakenly identified mangosteens as mangoes.**

MBQ Response: The fragrance of the mangosteen is a delightful blend of tropical sweetness and subtle floral notes. Imagine a scent that is both refreshing and rich, akin to a combination of strawberries and kiwi with a hint of rose petals. It's a fragrance that promises a burst of flavor and freshness, inviting you to take a bite and experience its unique taste.



清华大学电子工程系

Department of Electronic Engineering, Tsinghua University



Thank You !

MBQ: Modality-Balanced Quantization for Large Vision-Language Models

Shiyao Li^{1,2*}, Yingchun Hu^{2,3*}, Xuefei Ning¹⁺, Xihui Liu⁴, Ke Hong^{1,3}, Xiaotao Jia³⁺, Xiuhong Li²,
Yaqi Yan⁵, Pei Ran⁵, Guohao Dai^{6,2}, Shengen Yan², Huazhong Yang¹, Yu Wang¹⁺

¹Tsinghua University ²Infinigence-AI ³Beihang University

⁴University of Hong Kong ⁵ChinaTower ⁶Shanghai Jiao Tong University

