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

Supplementary Material

Describe this image Tokenizer VIT Encoder Large Vision-Language Model Fancy desert Prefill Stage Decode Stage

Figure 2. The inference process of Large VLMs. The blue patches represent language tokens, the red patches represent vision tokens.

6. Additional Preliminaries

6.1. The Inference Process of VLMs

The inference process of VLMs is shown in Fig. 2. The whole inference system consists of three key components:

- 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, and generate language tokens one by one.

Specifically, the transformer-based [42] VLMs have two distinctive stages, including the prefill and the decode stages. Take batch size = 1 as an example.

During the prefill stage, VLMs can take both vision tokens and language tokens as the input prompt. In this stage, VLMs aim to understand both the vision and language information and the connections across each modality. The Key and Value tensors of each attention block in VLMs are stored as the KV Cache to save the computation overhead in the decode stage. The input activation of each layer is typically a large 2D matrix, *making the primary computation operation in the prefill stage the General Matrix Multiply* (*GEMM*), which is compute-bound.

During the decode stage, VLMs take one generated token from step t as the input and use the KV Cache to generate the next token of step t+1. The generation of the current token depends on one previously generated token and the KV Cache. In this case, the input activation of each layer is typically a large 1D vector, and *the main computation operator in the decode stage is the General Matrix-Vector* Multiply (GEMV), which is memory-bound.

7. Related Work

7.1. LLM Quantization

Post-Training Quantization (PTQ) techniques are widely used in LLMs to accelerate the inference process. They employ the low-precision data format and computation to reduce the memory and computation overhead.

To accelerate the memory-bound decode stage of LLMs, existing methods apply weight-only quantization to reduce the memory access overhead. GPTQ [15] quantizes one weight channel at each step and iteratively adjusts the unquantized weights to mitigate reconstruction errors of each transformer block. AWQ [27] searches for proper channelwise equalization factors by minimizing the block-wise reconstruction loss. SpQR and LLM-MQ [14, 23] propose mixed-precision quantization to allocate higher precision for weight outliers, while the rest of the weights are quantized to low-precision. QuIP [7] introduces LDLQ, an optimal adaptive method for a quadratic proxy objective. It reveals that ensuring incoherence between weight and Hessian matrices can enhance the effectiveness of LDLQ. QuIP utilizes LDLQ and achieves incoherence by employing random orthogonal matrix multiplication.

To accelerate the compute-bound prefill stage of LLMs, existing methods propose to use the weight-activation quantization to leverage faster low-precision tensor cores. SmoothQuant [43] employs a channel-wise equalization technique to address the challenges of quantizing activation values. This method expands the data range of weight channels while shrinking the data range of corresponding activation channels to achieve better data equalization. Omniquant [39] optimizes the boundaries for weight clipping and the scaling factor for equivalent transformation to minimize reconstruction errors block-by-block. Atom [53] employs a strategy involving mixed-precision and dynamic quantization for activations. Recently, many studies [3, 33] follow the computational invariance idea, by multiplying rotation matrices to the weight matrices and activation matrices.

However, these methods focus solely on a single language modality without considering the differences between tokens from different modalities in multimodal scenarios, which is the core distinction between MBQ and existing quantization approaches. It is also worth noting that many existing studies search for various hyperparameters by minimizing reconstruction error, where MBQ can be used to achieve performance improvements with these methods on VLMs.

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

Table 10. The number of samples for the five error types in the LLaVA-bench-in-the-wild [31] and LLaVA-bench-wilder [20] dataset. The total number of samples is **188**. "SQ" is short for SmoothQuant.

7.2. Efficient VLM

To improve the efficiency of Large Vision-Language Models, existing work primarily focuses on designing lightweight modules, compressing vision tokens, and employing efficient model architectures.

Firstly, for the lightweight model design, an effective approach is to incorporate efficient components within the VLMs. Some research [5, 55] directly utilizes pre-trained small language models with fewer than 3B parameters as their backbone, while others [12] train a small language model from scratch. For modality alignment, [1, 22] utilizes a lightweight transformer, while subsequent work [11, 21, 28, 31] directly adopts Linear or MLP layers to align the visual modality with the language model's latent space.

Secondly, the number of vision tokens increases with image resolution, imposing a substantial computational burden on VLMs. To address this issue, [6, 38, 45] propose vision token reduction techniques to significantly lower the number of vision tokens, while some approaches [9, 30] remove redundant vision tokens to reduce computational overhead.

Finally, in terms of efficient architectures, some work [26, 35] leverages the Mixture of Experts (MoE) architecture to enhance model performance without increasing active parameter counts, while others [37, 52] adopt efficient Mamba language models as the language backbone.

8. Additional Experiments

8.1. W4A16 and W8A8 Results on Large VLMs

As shown in Tab. 11, we present the evaluation results for W4A16 and W8A8 quantized VLMs from the LLaVAonevision, InternVL2, and Qwen2-VL families. In most cases, the proposed MBQ achieves accuracy comparable to the AWQ and SmoothQuant baselines under W4A16 and W8A8 quantization. Furthermore, the average accuracy of the quantized VLMs is very close to that of the original FP16 VLMs, indicating that quantization under W4A16 and W8A8 is nearly lossless.

A notable different case arises during the W4A16 quantization of LLaVA-onevision-72B, where AWQ significantly degrades the VLM's accuracy, with the average accuracy falling more than 10% below that of MBQ and RTN. A similar phenomenon also occurs during W3A16 quantization of LLaVA-onevision-72B in Tab. 10, demonstrating that the modality-balancing concept proposed by MBQ can more consistently maintain high model performance compared to SOTA quantization baselines, whether in high-bitwidth or low-bitwidth quantization scenarios.

8.2. Case Studies

To evaluate the open-ended conversational ability of the quantized VLMs, we evaluate the proposed MBQ alongside state-of-the-art baselines on the LLaVA-onevision family using conversation benchmarks. As shown in Tab. 10, we find that MBQ can significantly outperform SOTA baseline methods when applied to weight-activation quantization for LLaVA-onevision-7B and weight-only quantization for LLaVA-onevision-72B. Therefore, we analyze the conversation results of VLMs under these two quantization schemes.

Specifically, we manually evaluated the responses of the quantized VLM to each question and identified the following five frequently occurring error types:

- 1. **No output**: The quantized VLM generates no or only a few valid tokens, as shown in Example 2;
- 2. **Randomness**: The quantized VLM randomly generates meaningless symbols, as shown in Example 4;
- 3. **Repetition**: The quantized VLM keeps repeating some certain tokens, as shown in Example 1 and Example 6;
- 4. **Condition Missing**: The quantized VLM misses the key points in the questions as shown in Example 3;
- 5. **Semantic Error**: The quantized VLM can understand the questions but still generates wrong answers with meaningful and fluent language, as shown in Example 5.

We summarize the number of samples corresponding to the above five error types for each quantized VLM, the results are shown in Tab. 10.

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. Specifically, the most common type of bad case of SmoothQuant is "No Output". For instance, in Example 2, the W4A8

| Model | Bitwidth | Method | MMMU | SEED | OCRBench | VizWiz | ScienceQA | TextVQA | Average (†) |
|---------------------|----------|--------|------|------|----------|--------|-----------|---------|-------------|
| LLaVA-onevision-7B | FP16 | - | 46.0 | 74.9 | 62.2 | 60.4 | 85.4 | 76.1 | 67.5 |
| | | RTN | 44.9 | 74.6 | 61.7 | 59.6 | 89.8 | 75.3 | 67.6 |
| | W4A16 | AWQ | 44.6 | 74.7 | 61.8 | 59.1 | 90.1 | 75.8 | 67.7 |
| | | MBQ | 44.4 | 74.7 | 62.1 | 59.3 | 90.2 | 75.6 | 67.7 |
| | W8A8 | RTN | 46.3 | 74.8 | 63.5 | 60.5 | 90.3 | 75.9 | 68.6 |
| | | SQ | 46.0 | 74.9 | 63.2 | 60.7 | 90.3 | 75.7 | 68.5 |
| | | MBQ | 45.6 | 74.7 | 62.6 | 61.0 | 90.2 | 75.7 | 68.3 |
| InternVL2-8B | FP16 | - | 48.0 | 76.0 | 76.5 | 61.1 | 96.2 | 77.0 | 72.5 |
| | W4A16 | RTN | 47.6 | 75.9 | 75.6 | 60.1 | 96.0 | 76.2 | 71.9 |
| | | AWQ | 47.1 | 75.8 | 76.7 | 60.1 | 96.3 | 76.4 | 72.1 |
| | | MBQ | 48.9 | 75.9 | 76.7 | 60.8 | 96.3 | 76.5 | 72.5 |
| | W8A8 | RTN | 47.4 | 76.2 | 77.3 | 61.0 | 96.2 | 76.9 | 72.5 |
| | | SQ | 48.0 | 76.1 | 77.1 | 61.0 | 96.1 | 76.9 | 72.5 |
| | | MBQ | 48.0 | 76.0 | 77.4 | 61.0 | 96.4 | 77.0 | 72.6 |
| Qwen2-VL-7B | FP16 | - | 50.6 | 76.4 | 80.7 | 68.3 | 85.1 | 82.0 | 73.8 |
| | W4A16 | RTN | 50.2 | 76.0 | 80.1 | 67.4 | 84.5 | 81.2 | 73.2 |
| | | AWQ | 50.1 | 76.1 | 80.4 | 68.4 | 85.0 | 81.7 | 73.6 |
| | | MBQ | 50.0 | 76.3 | 80.8 | 68.6 | 84.6 | 81.4 | 73.6 |
| | | RTN | 49.4 | 76.3 | 80.9 | 68.2 | 84.5 | 81.2 | 73.4 |
| | W8A8 | SQ | 50.1 | 76.3 | 80.6 | 68.5 | 85.0 | 81.5 | 73.7 |
| | | MBQ | 50.1 | 76.4 | 80.7 | 68.3 | 85.4 | 81.8 | 73.8 |
| LLaVA-onevision-72B | FP16 | - | 56.1 | 78.1 | 73.2 | 69.2 | 90.0 | 79.3 | 74.3 |
| | | RTN | 56.2 | 77.9 | 72.1 | 68.8 | 90.4 | 78.9 | 74.1 |
| | W4A16 | AWQ | 39.1 | 75.9 | 58.1 | 59.9 | 80.4 | 69.1 | 63.8 |
| | | MBQ | 56.4 | 77.9 | 72.3 | 69.0 | 90.3 | 79.3 | 74.2 |
| | W8A8 | RTN | 56.8 | 78.0 | 73.1 | 69.4 | 90.3 | 79.2 | 74.5 |
| | | SQ | 56.3 | 78.0 | 72.7 | 69.2 | 89.7 | 79.0 | 74.2 |
| | | MBQ | 56.2 | 78.1 | 73.1 | 69.2 | 89.8 | 79.1 | 74.3 |
| InternVL2-26B | FP16 | - | 47.1 | 76.8 | 77.9 | 66.2 | 97.5 | 82.1 | 74.6 |
| | W4A16 | RTN | 48.2 | 76.8 | 78.0 | 64.6 | 97.1 | 81.8 | 74.4 |
| | | AWQ | 47.4 | 76.8 | 77.1 | 65.9 | 97.3 | 82.0 | 74.4 |
| | | MBQ | 47.2 | 76.8 | 77.5 | 65.4 | 97.5 | 82.1 | 74.4 |
| | W8A8 | RTN | 47.4 | 76.5 | 78.4 | 65.1 | 97.3 | 81.7 | 74.4 |
| | | SQ | 48.1 | 76.7 | 78.3 | 65.5 | 97.4 | 82.0 | 74.7 |
| | | MBQ | 47.9 | 76.8 | 78.1 | 66.2 | 97.5 | 82.0 | 74.8 |
| Qwen2-VL-72B | FP16 | - | 61.1 | 77.6 | 79.9 | 76.0 | 91.6 | 82.5 | 78.1 |
| | W4A16 | RTN | 59.8 | 77.7 | 79.6 | 75.8 | 91.3 | 82.6 | 77.8 |
| | | AWQ | 61.9 | 77.5 | 79.5 | 75.8 | 92.2 | 82.4 | 78.2 |
| | | MBQ | 60.8 | 77.6 | 79.9 | 75.8 | 92.8 | 82.3 | 78.2 |
| | W8A8 | RTN | 60.4 | 77.5 | 79.5 | 75.7 | 90.9 | 82.2 | 77.7 |
| | | SQ | 60.3 | 77.5 | 79.7 | 76.1 | 91.3 | 82.0 | 77.8 |
| | | MBQ | 60.7 | 77.6 | 79.9 | 75.9 | 91.4 | 82.4 | 78.0 |

Table 11. Additional results on LLaVA-onevision, InternVL2, and Qwen2-VL families. "SQ" is short for SmoothQuant.

VLM using SmoothQuant fails to generate any meaningful tokens in response to the provided question. In contrast, with MBQ, the W4A8 VLM experiences only 3 "No Output" bad cases. Additionally, 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 potential for "Semantic Error" cases.

For the LLaVA-onevision-72B with W3A16 quantization, a similar trend is observed as with the LLaVAonevision-7B using W4A8 quantization. MBQ results in only 38 bad cases, which is 48 fewer than the 86 bad cases reported for AWQ. Specifically, the most frequent type of bad case for the AWQ baseline is still "No Output". This indicates that AWQ can still lead to VLM failing to generate meaningful and fluent responses in many situations, while MBQ significantly alleviates this type of bad case, narrowing the issues down to "Semantic Errors".

Example 1. LLaVA-onevision-7B-W4A8

Question: Who painted this artwork, and what is its significance in the art world?



FP16 Response: This artwork was painted by Vincent van Gogh, a post-impressionist painter known for his vibrant use of color and emotional intensity. "The Starry Night" is significant in the art world as it exemplifies Van Gogh's unique style and his ability to evoke emotion through his paintings. It has become one of the most recognized and celebrated works of post-impressionism, symbolizing hope and the beauty of the night sky.

SmoothQuant Response: This artwork was painted by Vincent van Gogh titled Nightsky Landscape With Bright Starry Sky Above Sleeping Townscape Surrounding Wind Blowing Trees And Hills In Background Van Gogh Postcard Print Poster Blanket Pillowcase Pillow Box Pil

MBQ Response: This artwork is painted by Vincent van Gogh, titled "The Starry Night." It's significant in the art world because it exemplifies Van Gogh's unique post-impressionist style and emotional intensity. The swirling patterns in the sky are iconic and have become synonymous with Van Gogh's expressive technique. This painting also signifies his fascination with the cosmos and has been interpreted as an expression of hope amidst darkness.

Example 2. LLaVA-onevision-7B-W4A8

Question: Describe this photo in detail.



FP16 Response: The image is a digitally created artwork that combines elements of Renaissance art with modern humor. It features a dog's head, specifically that of a dachshund, seamlessly integrated into the iconic pose and attire of Leonardo da Vinci's "Mona Lisa." The dog's face is superimposed onto the original painting, maintaining the Mona Lisa's enigmatic expression and posture. The background remains consistent with the original, depicting a distant landscape with a river and mountains under a cloudy sky. The dog's fur color matches the original's flesh tones, and its ears are styled to resemble the Mona Lisa's hair. The overall effect is a playful fusion of classical art and contemporary animal imagery.

SmoothQuant Response: The image presents itself uniquely ** Error Type: No Output. The quantized VLM can only generate a few meaningless tokens.

MBQ Response: This image is a playful rendition of Leonardo da Vinci's famous painting, the Mona Lisa. However, instead of the original human subject, this version features a dog, specifically a brown dachshund, in place of the Mona Lisa. The dog is portrayed with a slight smile, adding a touch of whimsy to the piece. The background maintains the original's serene landscape of mountains and a calm river. The art style is a fusion of Renaissance painting techniques with a modern twist, reflecting the artist's creative flair.

Example 3. LLaVA-onevision-7B-W4A8

Question: What is the atmosphere and setting depicted in the image, and how do the elements contribute to creating a relaxing environment?



FP16 Response: The atmosphere depicted in the image is one of tranquility and relaxation. The setting is a cozy living room bathed in soft, warm light that suggests it's either dawn or dusk. The presence of the person sitting cross-legged on the floor with headphones on, along with the calm posture and the cup of coffee, implies they are enjoying some quiet time, possibly listening to music. The room's decor, including framed pictures and plants, adds to the homely and peaceful ambiance. The sunlight streaming through the window enhances the serene mood, making it an ideal space for unwinding.

SmoothQuant Response: The atmosphere depicted suggests tranquility amidst warmth—an ideal setting conducive to relaxation—one might even say therapeutic listening experience.

** Error Type: Condition Missing. The quantized VLM fails to address the question: "How do the elements contribute to creating a relaxing environment?"

MBQ Response: The atmosphere depicted in the image is one of tranquility and relaxation. The setting appears to be a cozy living room designed for comfort and leisure, with warm lighting that suggests either sunrise or sunset. Elements such as plush cushions on the floor, soft lighting from the pendant lamp, and the presence of calming artwork contribute to this relaxing ambiance. Additionally, the person's posture while listening to music through headphones indicates immersion in the music, suggesting that they are enjoying a peaceful moment alone.

Example 4. LLaVA-onevision-72B-W3

Question: What model is this car, and can you provide some specifications and performance details about it?



FP16 Response: This car is a Bugatti Chiron, known for its exceptional performance and luxury. It features an 8.0-liter, 16-cylinder engine that produces 1,500 horsepower and 1,300 lb-ft of torque. The Chiron can accelerate from 0 to 60 mph in just 2.4 seconds and has a top speed electronically limited to 261 mph. Its advanced engineering and craftsmanship make it one of the fastest and most exclusive cars in the world.

AWQ Response: This is a Bugatti Veyron,EBu,Chirona/EB110,EB110/EB1

** Error Type: Randomness. The quantized VLM generates random symbols such as "EBu" and "/EB100".

MBQ Response: This is a Bugatti Chiron, known for its incredible performance. It boasts an 8.0-liter, 16-cylinder engine that produces 1,479 horsepower and 1,300 lb-ft of torque. The Chiron can accelerate from 0 to 60 mph in just 2.4 seconds and has a top speed electronically limited to 261 mph.

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.

