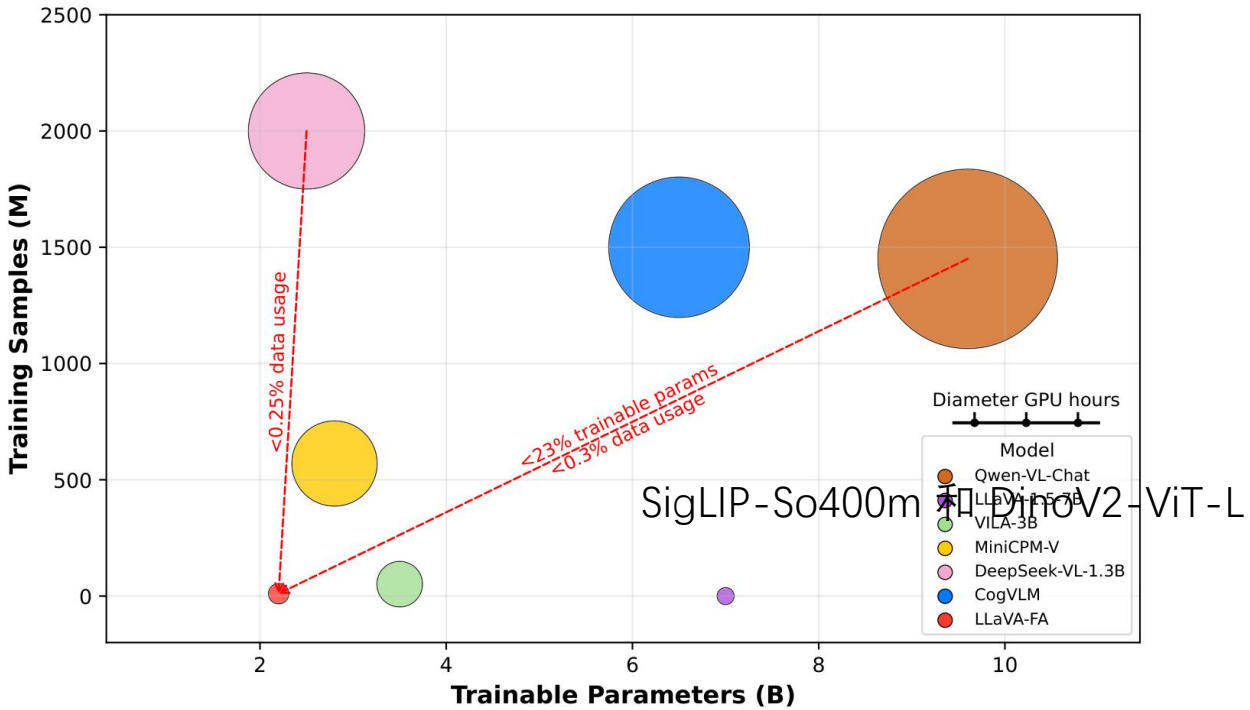
电子科技大学
University of Electronic Science and Technology of China

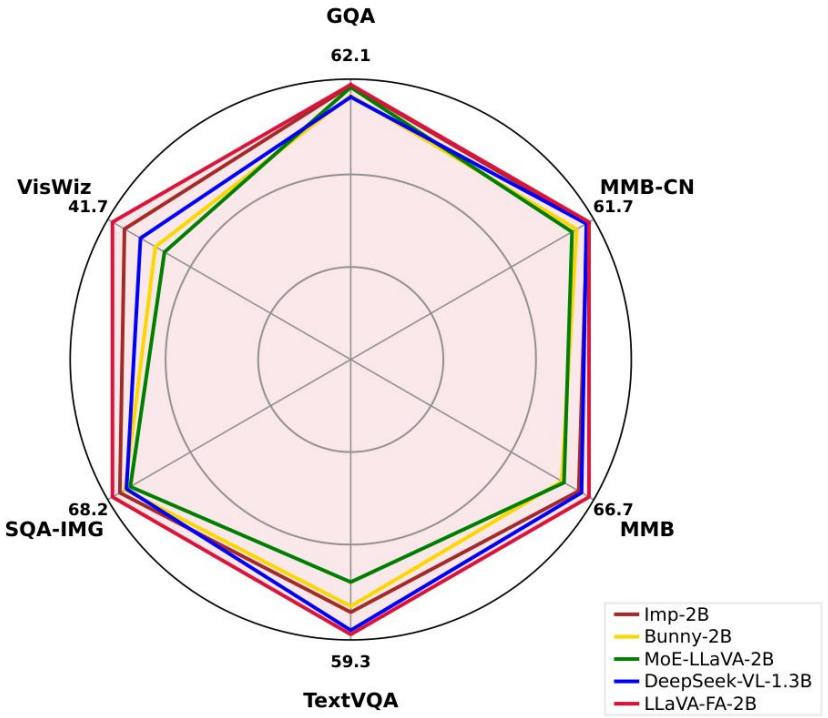
LLaVA-FA: Learning Fourier Approximation For Compressing Large Multimodal Models

Pengcheng Zheng, Chaoning Zhang,* Jiarong Mo, GuoHui Li, Jiaquan Zhang, Jiahao Zhang, Sihan Cao, Sheng Zheng, Caiyan Qin, Guoqing Wang, Yang Yang

2026.03.07

Introduction & Motivation





❑ Computational Cost Challenged

- Training LMMs is extremely expensive
- Inference consumes raises environmental concerns
- Massive computational cost limit broader usage

❑ Compression & Performance Challenges

- Separate low-rank and quantization compound errors.
- Extra image encoders make compression harder.
- Performance degradation after compression.

Introduction & Motivation

Methods ⟷ Limitation

Methods	Limitation
Low-rank + Quantization (Separate)	Compounded reconstruction errors
Spatial domain compression	Ignores frequency domain benefits
LLM compression on LMMs	Cross-modal redundancy proliferation
Large-scale calibration datasets	Data accessibility constraints
Full Hessian computation	High computational overhead
Direct weight factorization	Misses de-correlation properties
Token-level LMMs compression	Vision encoder parameter inflation

Core Question:

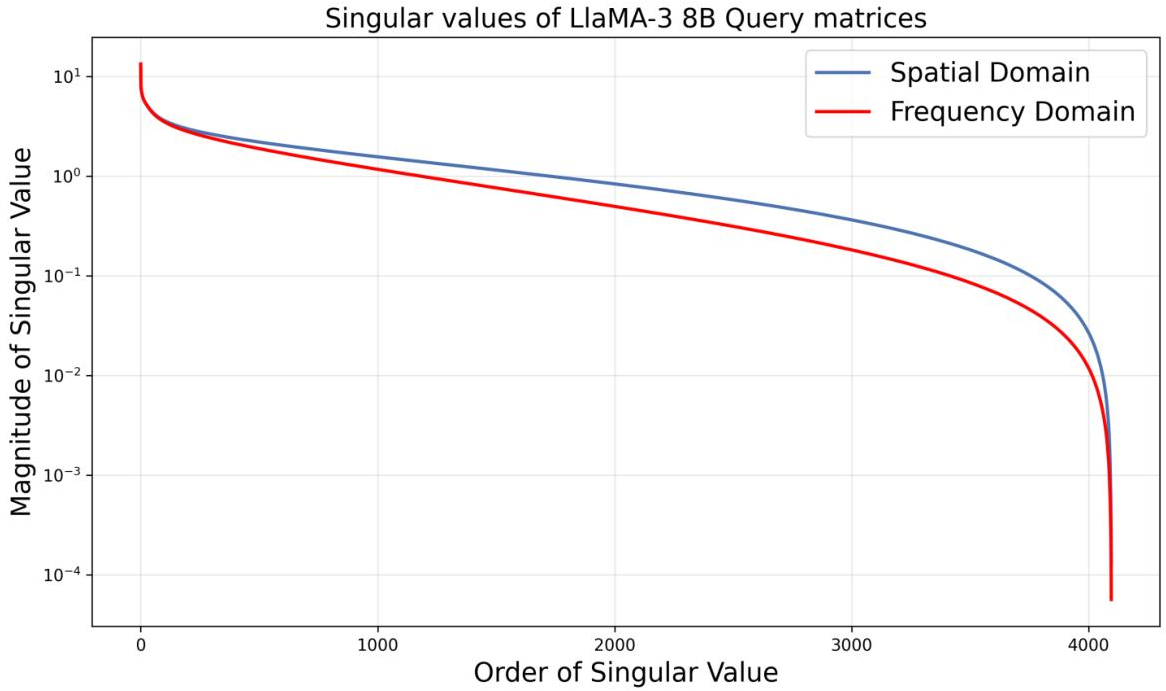
How about leveraging **frequency-domain** properties for more efficient and effective compression?

Introduction & Motivation

❑Why Frequency Domain?

❑Theoretical Advantages:

- **De-correlation**: More compact singular value spread
- **Conjugate Symmetry**: ~50% parameter reduction for real matrices
- **Energy Compaction**: Most energy in few coefficients

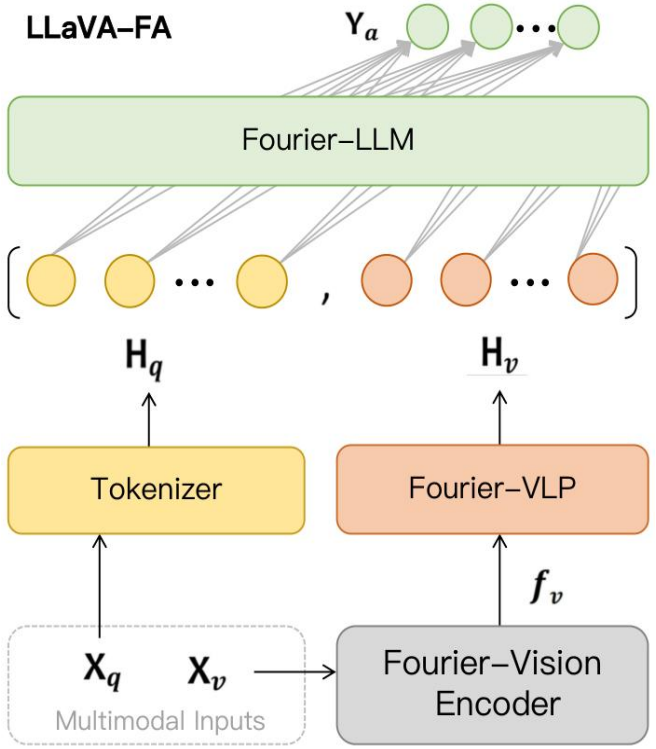


Methodology

❑ Framework Overview

❑ System Architecture

- **Fourier-Vision Encoder**: CLIP-ViT-L/14 compressed via Fourier Approximation
- **Fourier-VLP**: Two-layer MLP for cross-modal alignment
- **Fourier-LLM**: Qwen-2.5 series compressed via Fourier Approximation



❑ Key Innovation

- **Joint optimization**: Unlike existing methods that decouple low-rank and quantization
- **Frequency domain processing**: All weight matrices transformed to complex domain
- **Autoregressive generation**: $p(Y_a|H_v, H_q) = \prod_{i=1}^{L} p(y_i|H_v, H_q, y_{<i})$

Methodology

❑ Core Algorithms

❑ Mathematical Formulation

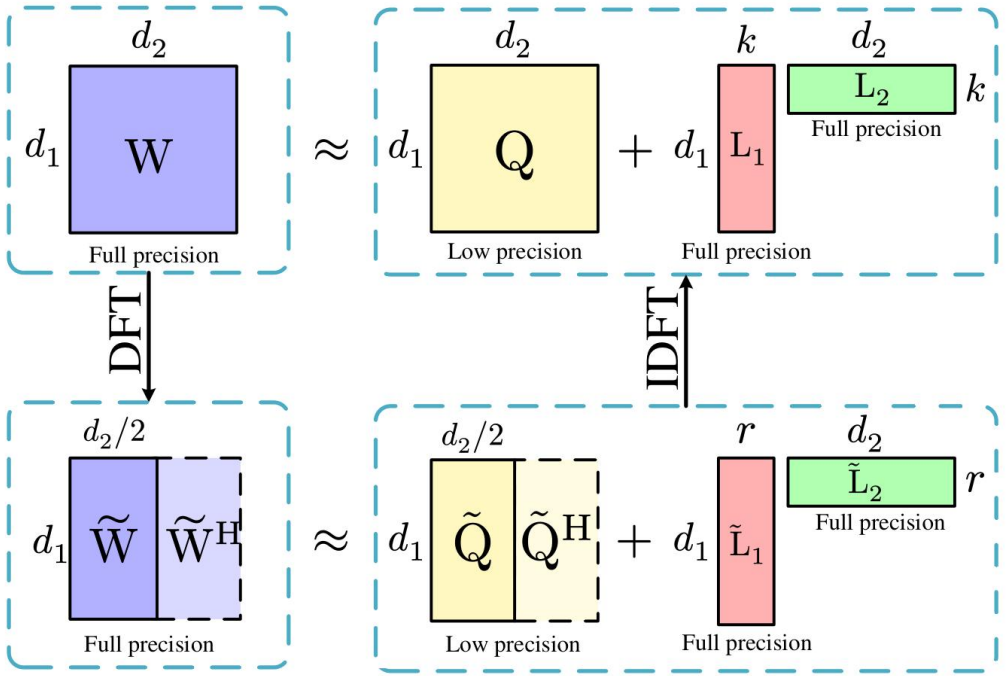
- **Spatial domain**: $\mathbf{W} \approx \mathbf{Q} + \mathbf{L_1 L_2}$
- **Frequency domain**: $\widetilde{\mathbf{W}} \approx \widetilde{\mathbf{Q}} + \widetilde{\mathbf{L}}_1 \widetilde{\mathbf{L}}_2$

❑ Technical Details

- **Optimization objective**: $\min_{\widetilde{\mathbf{Q}}, \widetilde{\mathbf{L}}_1, \widetilde{\mathbf{L}}_2} \left\| \sqrt{\mathbf{C}} \odot \left| \widetilde{\mathbf{W}} - (\widetilde{\mathbf{Q}} + \widetilde{\mathbf{L}}_1 \widetilde{\mathbf{L}}_2) \right| \right\|_F$
- **PolarQuant steps**: $r_{i,j} \leftarrow \sqrt{X_{i,j}^2 + Y_{i,j}^2}$, $\theta_{i,j} \leftarrow \operatorname{atan2}(Y_{i,j}, X_{i,j})$
- **Alternating updates:** FourierSVD ↔ PolarQuant

❑ Key Advantages

- **Conjugate symmetry**: Real matrices → 50% parameter reduction
- **Singular value distribution**: More compact energy in frequency domain
- **PolarQuant**: Separate amplitude and phase for complex matrix quantization



Methodology

Optimization Algorithm for Fourier Approximation

Algorithm 1: Fourier Approximation

Input : Complex weight matrix $\widetilde{\mathbf{W}}$, Calibration matrix $\mathbf{C}$, Target rank r, Amplitude bitwidth b_r, Phase bitwidth b_θ

Initialize $\widetilde{\mathbf{Q}} \leftarrow \mathbf{0}$ and $\epsilon_0 \leftarrow \infty$

for $t \leftarrow 1$ **to** $T-1$ **do**

  # Fourier SVD with rank r

  $\widetilde{\mathbf{L}}_1, \widetilde{\mathbf{L}}_2 \leftarrow \text{FourierSVD}(\widetilde{\mathbf{W}} - \widetilde{\mathbf{Q}}, \mathbf{C}, r)$

  # Polar Quantization

  $\widetilde{\mathbf{Q}} \leftarrow \text{PolarQuant}\left(\widetilde{\mathbf{W}} - \widetilde{\mathbf{L}}_1\widetilde{\mathbf{L}}_2, b_r, b_\theta\right)$

  if $\mathbf{C}$ *is* None **then**

    # Weighted error

    $\epsilon_t = \left\| \left| \widetilde{\mathbf{W}} - (\widetilde{\mathbf{Q}} + \widetilde{\mathbf{L}}_1\widetilde{\mathbf{L}}_2) \right| \right\|_F$

  else

    # Calibration weighted error

    $\epsilon_t = \left\| \sqrt{\mathbf{C}} \odot \left| \widetilde{\mathbf{W}} - (\widetilde{\mathbf{Q}} + \widetilde{\mathbf{L}}_1\widetilde{\mathbf{L}}_2) \right| \right\|_F$

  end

  if $\epsilon_t > \epsilon_{t-1}$ **then**

    break

  end

end

Output : $\widetilde{\mathbf{Q}}, \widetilde{\mathbf{L}}_1, \widetilde{\mathbf{L}}_2, \epsilon_t$

Algorithm 2: FourierSVD

Input : Complex residual matrix $\widetilde{\mathbf{R}} = \widetilde{\mathbf{W}} - \widetilde{\mathbf{Q}}$, Calibration matrix $\mathbf{C}$, Target rank r

if $\mathbf{C}$ *is* None **then**

  # Perform SVD on complex matrix

  $\widetilde{\mathbf{U}}, \mathbf{\Sigma}, \widetilde{\mathbf{V}}^H \leftarrow \text{SVD}(\widetilde{\mathbf{R}})$

  # Keep only top r singular values

  $\mathbf{\Sigma}_r \leftarrow \mathbf{\Sigma}[1:r, 1:r]$,

  $\widetilde{\mathbf{U}}_r \leftarrow \widetilde{\mathbf{U}}[:, 1:r]$,

  $\widetilde{\mathbf{V}}_r^H \leftarrow \widetilde{\mathbf{V}}^H[:, 1:r]$

  $\widetilde{\mathbf{L}}_1 \leftarrow \widetilde{\mathbf{U}}_r\sqrt{\mathbf{\Sigma}_r}$, $\widetilde{\mathbf{L}}_2 \leftarrow \sqrt{\mathbf{\Sigma}_r}\widetilde{\mathbf{V}}_r^H$

else

  $\mathbf{D}_{\text{row}} \leftarrow \text{RowAverage}(\mathbf{C})$

  $\mathbf{D}_{\text{col}} \leftarrow \text{ColAverage}(\mathbf{C})$

  $\widetilde{\mathbf{U}}, \mathbf{\Sigma}, \widetilde{\mathbf{V}}^H \leftarrow \text{SVD}(\mathbf{D}_{\text{row}}\widetilde{\mathbf{R}}\mathbf{D}_{\text{col}})$

  # Keep only top r singular values

  $\mathbf{\Sigma}_r \leftarrow \mathbf{\Sigma}[1:r, 1:r]$,

  $\widetilde{\mathbf{U}}_r \leftarrow \widetilde{\mathbf{U}}[:, 1:r]$,

  $\widetilde{\mathbf{V}}_r^H \leftarrow \widetilde{\mathbf{V}}^H[:, 1:r]$

  $\widetilde{\mathbf{L}}_1 \leftarrow \mathbf{D}_{\text{row}}^{-1}\widetilde{\mathbf{U}}_r\sqrt{\mathbf{\Sigma}_r}$

  $\widetilde{\mathbf{L}}_2 \leftarrow \sqrt{\mathbf{\Sigma}_r}\widetilde{\mathbf{V}}_r^H\mathbf{D}_{\text{col}}^{-1}$

end

Output : $\widetilde{\mathbf{L}}_1, \widetilde{\mathbf{L}}_2$

Algorithm 3: PolarQuant

Input : Complex residual matrix $\widetilde{\mathbf{R}} = \widetilde{\mathbf{W}} - \widetilde{\mathbf{L}}_1\widetilde{\mathbf{L}}_2$, Amplitude bitwidth b_r, Phase bitwidth b_θ

Extract real and imaginary parts

$\mathbf{C} \leftarrow \text{Re}(\widetilde{\mathbf{R}})$, $\mathbf{Y} \leftarrow \text{Im}(\widetilde{\mathbf{R}})$

Convert to polar coordinates

for *each element* (i, j) **do**

  $r_{i,j} \leftarrow \sqrt{X_{i,j}^2 + Y_{i,j}^2}$

  $\theta_{i,j} \leftarrow \text{atan2}(Y_{i,j}, X_{i,j})$

end

Compute quantization parameters

$r_{\max} \leftarrow \max(r_{i,j})$

$\Delta_r \leftarrow r_{\max}/(2^{b_r} - 1)$

$\Delta_\theta \leftarrow 2\pi/(2^{b_\theta})$

Quantize amplitude and phase

for *each element* (i, j) **do**

  $q_{r,i,j} \leftarrow \text{round}(r_{i,j}/\Delta_r)$

  $q_{\theta,i,j} \leftarrow \text{round}((\theta_{i,j} + \pi)/\Delta_\theta)$

  $\hat{r}_{i,j} \leftarrow q_{r,i,j} \cdot \Delta_r$

  $\hat{\theta}_{i,j} \leftarrow q_{\theta,i,j} \cdot \Delta_\theta - \pi$

end

Reconstruct complex matrix

for *each element* (i, j) **do**

  $\widetilde{\mathbf{Q}}_{i,j} \leftarrow \hat{r}_{i,j} \cdot e^{i\hat{\theta}_{i,j}}$

end

Output : Quantized complex matrix $\widetilde{\mathbf{Q}}$

Experiments

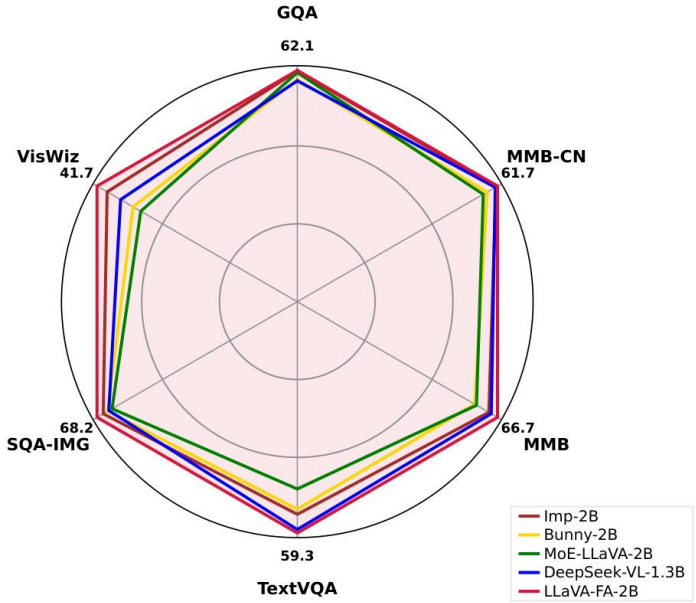
❑Main Results

❑Key Highlights

- 60.9% average accuracy across benchmarks
- +5.5% improvement over MoE-LLaVA-2B
- +3.7% improvement over Mini-Gemini-2B
- Competitive with 7B models while using 2B parameters

❑Training Efficiency

- **Training samples:** 5M (vs 15M-2000M for competitors)
- **Parameters:** ~2B (77% reduction from base model)
- **Data usage**: <0.3% compared to large models
- 23% trainable parameters of full model
- Maintained performance with aggressive compression



Method	LLM	#Sample	#Param	GQA	VisWiz	SQA^I	VQA^T	MME	MMB	MMB^{CN}	AVG
BLIP-2	Vicuna-13B	129M		41.0	19.6	61.0	42.5	64.7	-	-	-
VILA-7B	LLaMA-7B	50M		62.3	57.8	68.2	64.4	76.7	68.9	61.7	65.7
CogVLM	Vicuna-7B	1500M		64.9	-	65.6	**78.2**	71.8	63.7	53.8	-
InstructBLIP	Vicuna-13B	130M	≥7B	49.5	33.4	63.1	50.7	60.6	-	-	-
Qwen-VL-Chat	Qwen-7B	1450M		57.5	38.9	68.2	61.5	74.4	60.6	56.7	56.7
Deepseek-VL-7B	DLLM-7B	2000M		61.3	49.9	74.0	64.7	73.4	74.1	**72.8**	67.2
LLaVA-1.5-7B	Vicuna-1.5-7B	1.2M		62.0	50.0	66.8	58.2	75.5	64.3	58.3	62.1
LLaVA-NeXT	Vicuna-1.5-13B	1.3M		65.4	60.5	73.6	67.1	**78.7**	70.4	64.4	68.5
LLaVA-FA-7B	InternLM-2-20B	5M		**68.5**	**62.0**	**76.0**	68.0	74.5	**74.5**	69.5	**70.4**
Imp-3B	Phi-2-2.7B	1.6M		63.5	54.1	72.8	59.8	72.3	**72.9**	46.7	63.2
Bunny-3B	Phi-2-2.7B	2.7M		62.5	43.8	70.9	56.7	**74.4**	68.6	37.2	59.2
VILA-3B	LLaMA-2.7B	51M		61.5	53.5	69.0	60.4	72.1	63.4	52.7	61.8
MobileVLM	MLLaMA-2.7B	1.3M		59.0	-	61.0	47.5	64.4	59.6	-	-
$MobileVLM^{v2}$	MLLaMA-2.7B	3.6M	~3B	61.1	-	70.0	57.5	72.0	63.2	-	-
MoE-LLaVA-3B	Phi-2-2.7B	2.2M		61.4	43.9	68.5	51.4	71.1	65.2	41.8	57.6
MiniCPM-V	MiniCPM-2.4B	570M		51.5	50.5	74.4	56.6	68.9	64.0	62.7	61.2
MiniCPM-V-2	MiniCPM-2.4B	570M		52.1	60.2	76.3	**73.2**	70.5	68.5	67.2	66.9
LLaVA-FA-3B	LLaMA-3-8B	5M		**65.0**	**62.5**	**77.0**	64.0	71.0	70.5	**68.0**	**68.3**
Imp-2B	Qwen-1.5-1.8B	1.6M		61.9	39.6	66.1	54.5	65.2	63.8	61.2	58.9
Bunny-2B	Qwen-1.5-1.8B	2.7M		59.6	34.2	64.6	53.2	65.0	59.1	58.5	56.3
Mini-Gemini-2B	Gemma-2B	2.7M	~2B	60.7	41.5	63.1	56.2	**67.0**	59.8	51.3	57.1
MoE-LLaVA-2B	Qwen-1.5-1.8B	2.2M		61.5	32.6	63.1	48.0	64.6	59.7	57.3	55.3
DeepSeek-VL-1.3B	DLLM-1.3B	2000M		59.3	36.8	64.2	58.4	65.3	64.6	61.0	58.5
LLaVA-FA-2B	Qwen-2.5-7B	5M		**62.1**	**41.7**	**68.2**	**59.3**	66.6	**66.7**	**61.7**	**60.9**
SPHINX-Tiny	TLLaMA-1.1B	15M	~1B	**58.0**	49.2	21.5	57.8	63.1	56.6	37.8	49.2
LLaVA-FA-1B	Qwen-2.5-3B	5M		56.7	**49.7**	**61.3**	**57.9**	**63.3**	**58.3**	**49.4**	**56.7**

Experiments

❑Hallucination & Ablation Results

Model	LLM	#Param	Object HalBench		POPE	MMHal-Bench	
			Resp ↓	Ment ↓	F1 ↑	Score ↑	Hall ↓
Qwen-VL-Chat	Qwen-7B	9.6B	40.4	20.7	74.9	2.76	38.5
LLaVA-1.5-7B	Vicuna-7B	7B	53.6	25.2	86.1	2.36	51.0
VCD	Vicuna-1.5-7B	7B	48.8	24.3	84.5	2.12	54.2
OPERA	Vicuna-1.5-7B	7B	45.1	22.3	85.4	2.15	54.2
HA-DPO	Vicuna-1.5-7B	7B	39.9	19.9	86.8	1.98	60.4
POVID	Vicuna-1.5-7B	7B	48.1	24.4	86.3	2.08	56.2
LLaVA-RLHF	Vicuna-1.5-13B	13B	38.1	18.9	82.7	2.02	62.5
LURE	Vicuna-1.5-7B	7B	27.7	17.3	-	1.64	60.4
RLHF-V	Vicuna-13B	13B	12.2	7.5	86.2	2.45	51.0
RLAIF-V	Vicuna-1.5-7B	7B	8.5	4.3	-	3.06	29.2
MiniCPM-V	MiniCPM-2.4B	2.8B	21.6	11.5	79.5	3.70	24.9
MiniCPM-V-2	MiniCPM-2.4B	2.8B	14.5	7.8	86.3	**4.09**	18.2
Mini-Gemini-2B	Gemma-2B	2B	29.7	21.1	85.6	2.83	18.8
Bunny-2B	Qwen-1.5-1.8	2.2B	50.2	23.4	85.8	2.72	19.3
LLaVA-FA-2B	Qwen-2.5-7B	2.2B	**11.2**	**7.7**	**87.5**	2.79	**17.5**

Table 3: Performance comparison with different rank.

Model	Rank	Avg Acc ↑
LLaVA-FA-1B	64	55.3
	128	56.3
	256	**56.6**
LLaVA-FA-2B	64	58.9
	128	59.7
	256	**60.8**

Table 4: Impact of b_r and b_θ on model performance and compression ratio (CR).

b_r	b_θ	Avg Acc ↑	CR ↑	Hall ↓
2	2	56.3	**14.5**	18.7
3	3	59.1	12.8	13.4
4	4	60.9	11.2	11.2
5	4	61.0	10.1	10.9
4	5	60.9	10.8	11.0
6	6	**61.2**	8.9	**10.5**

❑Key Highlights

- 11.2% response-level hallucination (best in class)
- 87.5% POPE F1 score (highest accuracy)
- Outperforms larger 7B models

❑**Finding:** Higher rank enables better adaptation, especially for larger base models (+1.9% gain for 2B model)

❑**Optimal Configuration**:

- br = bθ = 4 balances quality and compression
- 11.2× compression ratio maintained

Experiments

❑Qualitative Analysis

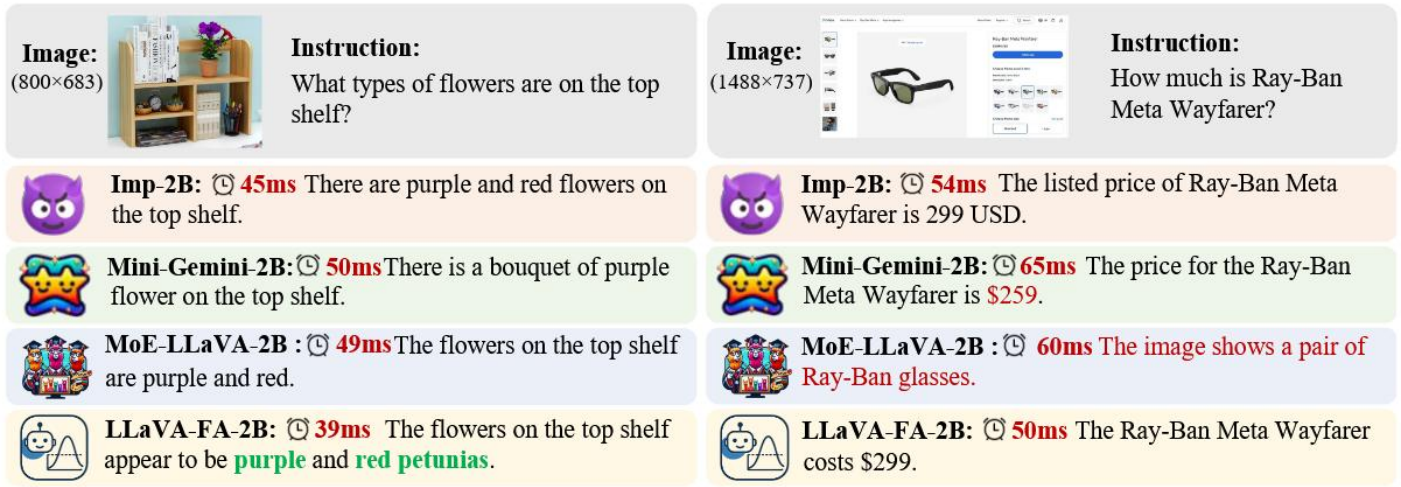


Figure 5: Qualitative comparison. LLaVA-FA-2B demonstrates superior precision and efficiency on fine-grained recognition (left) and OCR tasks (right). It correctly identifies specific details (*e.g.*, "petunias") and prices with the lowest latency among all baselines.

❑Key Highlights

- **Higher Precision**: Delivers specific, fine-grained visual recognition.
- **Superior OCR**: Accurately extracts complex text and numerical data.
- **Faster Inference**: Maintains the lowest latency, achieving a 13–28% speedup over competitors.

❑Comparison of data efficiency

Table 15: Comparison of data efficiency. LLaVA-FA-2 outperforms baseline models even when restricted t only 1M samples, which demonstrates its superior da efficiency.

Model	Training samples	AVG↑
Imp-2B	1.6M	58.9
Bunny-2B	2.7M	56.3
Mini-Gemini-2B	2.7M	57.1
MoE-LLaVA-2B	2.2M	55.3
DeepSeek-VL-1.3B	2000M	58.5
LLaVA-FA-2B	5M	60.9
LLaVA-FA-2B	2.5M	59.7
LLaVA-FA-2B	1M	**59.2**

❑Key Highlights

- **Data Efficiency**: Outperforms rivals using significantly fewer training samples.
- **Architectural Advantage**: Gains stem from Fourier-domain compression rather than data volume.

Conclusions

- **LLaVA-FA** leverages joint low-rank + quantization approximation in **frequency** domain
  - PolarQuant: polar-coordinate quantization for complex matrices
- **LLaVA-FA** Leverages Fourier properties for **compact and accurate** representations
- Practical solution for deploying LMMs in resource-constrained scenarios
  - Optional diagonal calibration eliminates need for large-scale calibration data
- Opens new direction for frequency-domain neural compression