Winter Conference on Applications of Computer Vision, WACV 2026

MR-Pruner: Training-free Multi-resolution Visual Token Pruning for Multi-modal Large Language Models

Seunghoon Han[1], Hyewon Lee[1], SoYoung Park[1],
Jong-Ryul Lee[1,†], Sungsu Lim[1,†]

[1]Chungnam National University

Chungnam National University

Introduction

Multi-modal LLMs (MLLMs)

- Large Language Models extended to multi-modality (e.g., LLaVA[1])
- Achieve strong performance on complex vision-language tasks

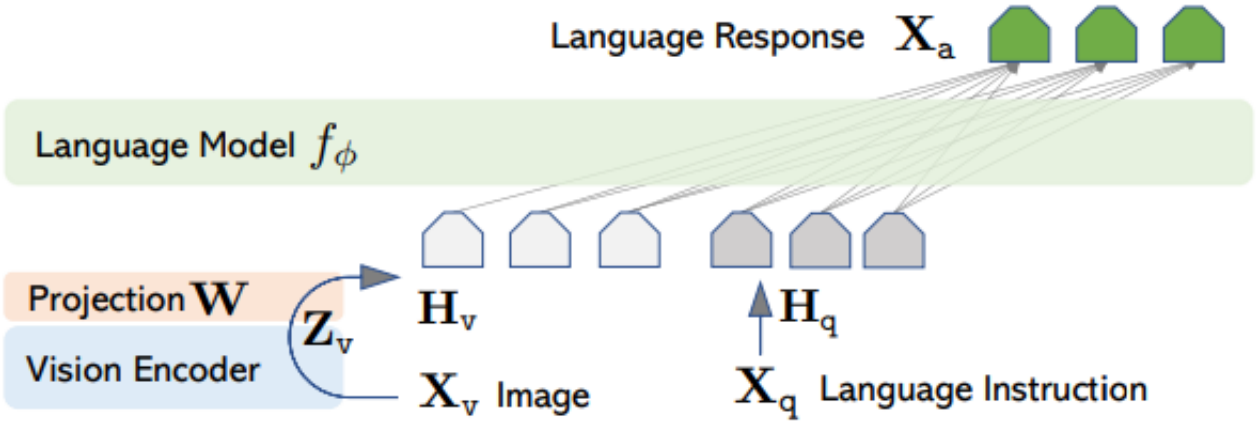


Figure 1. Multi-modal LLM architecture

[1] Visual Instruction Tuning, *NeurIPS 2023*

Introduction

Multi-resolution MLLMs

- Recent MLLMs support **high-resolution inputs** (e.g., LLaVA-NeXT[2])
- Better performance in fine-grained tasks (e.g., OCR)

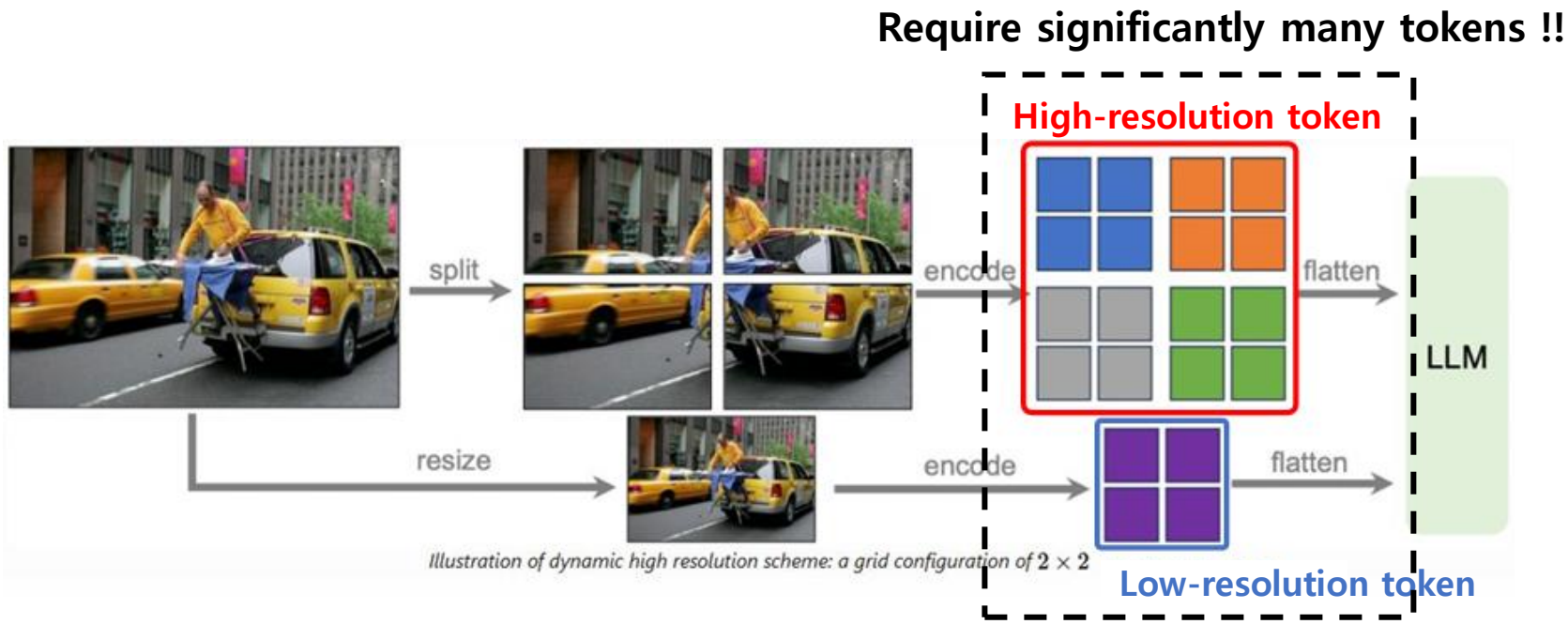


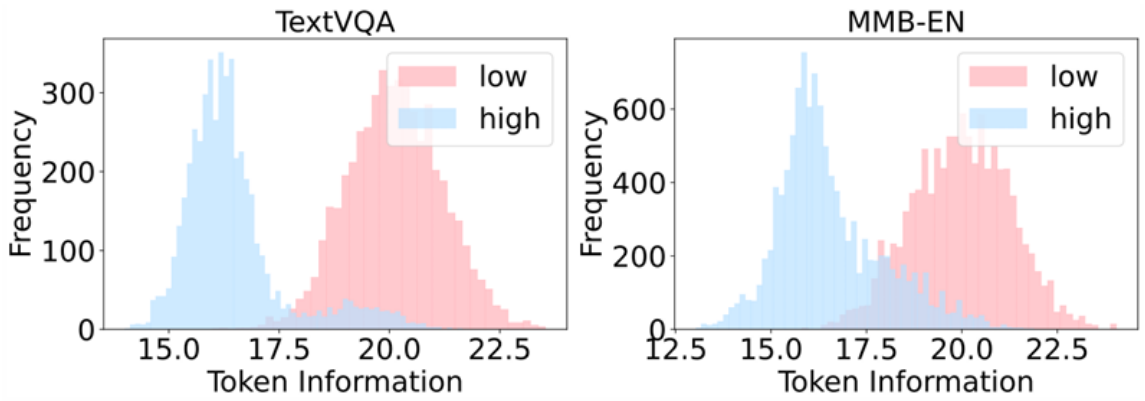
Figure 2. The overview of Multi-resolution MLLMs

[2] LLaVA-NeXT: Improved reasoning, OCR, and world knowledge (https://llava-vl.github.io/blog/2024-01-30-llava-next), 2024

Key Observations

Characteristics of Multi-resolution Tokens

- **Different informativeness distributions** (Left)
- **Mutual Complementarity** (Right)



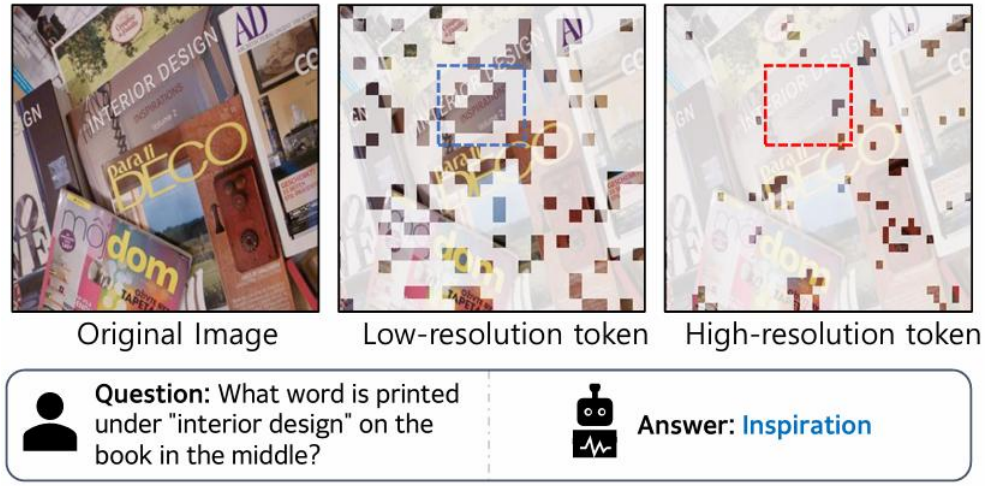


Figure 3. Properties of multi-resolution visual tokens

Challenge and Solution

Challenge. Using the same pruning strategy / pruning ratio without considering resolution types.

Solution. Applying different token pruning strategies / pruning ratios that reflect the characteristics of tokens according to their resolution types.

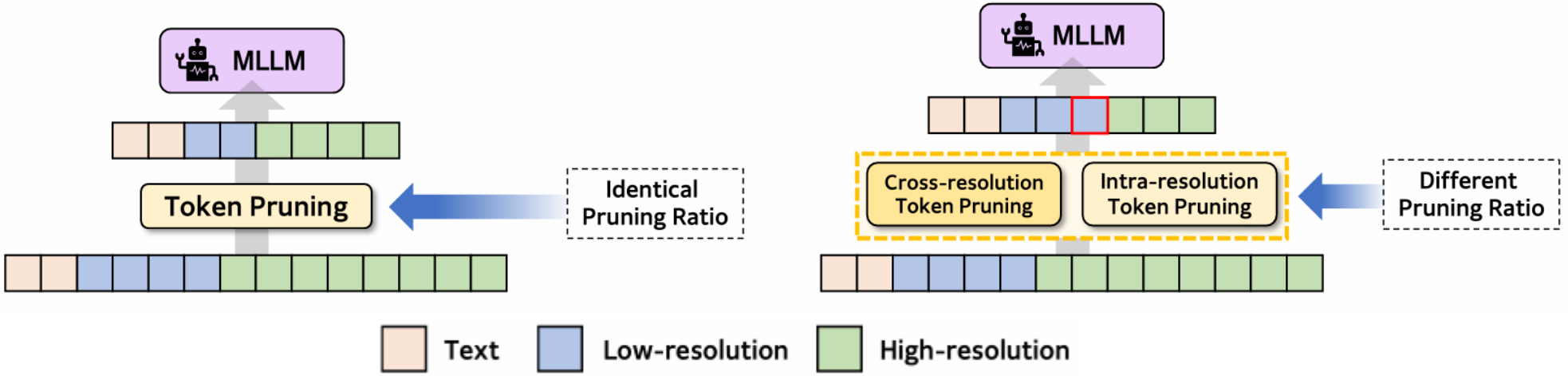


Figure 4. Comparison between single-resolution pruning methods and proposed method.

Methodology

MR-Pruner Framework

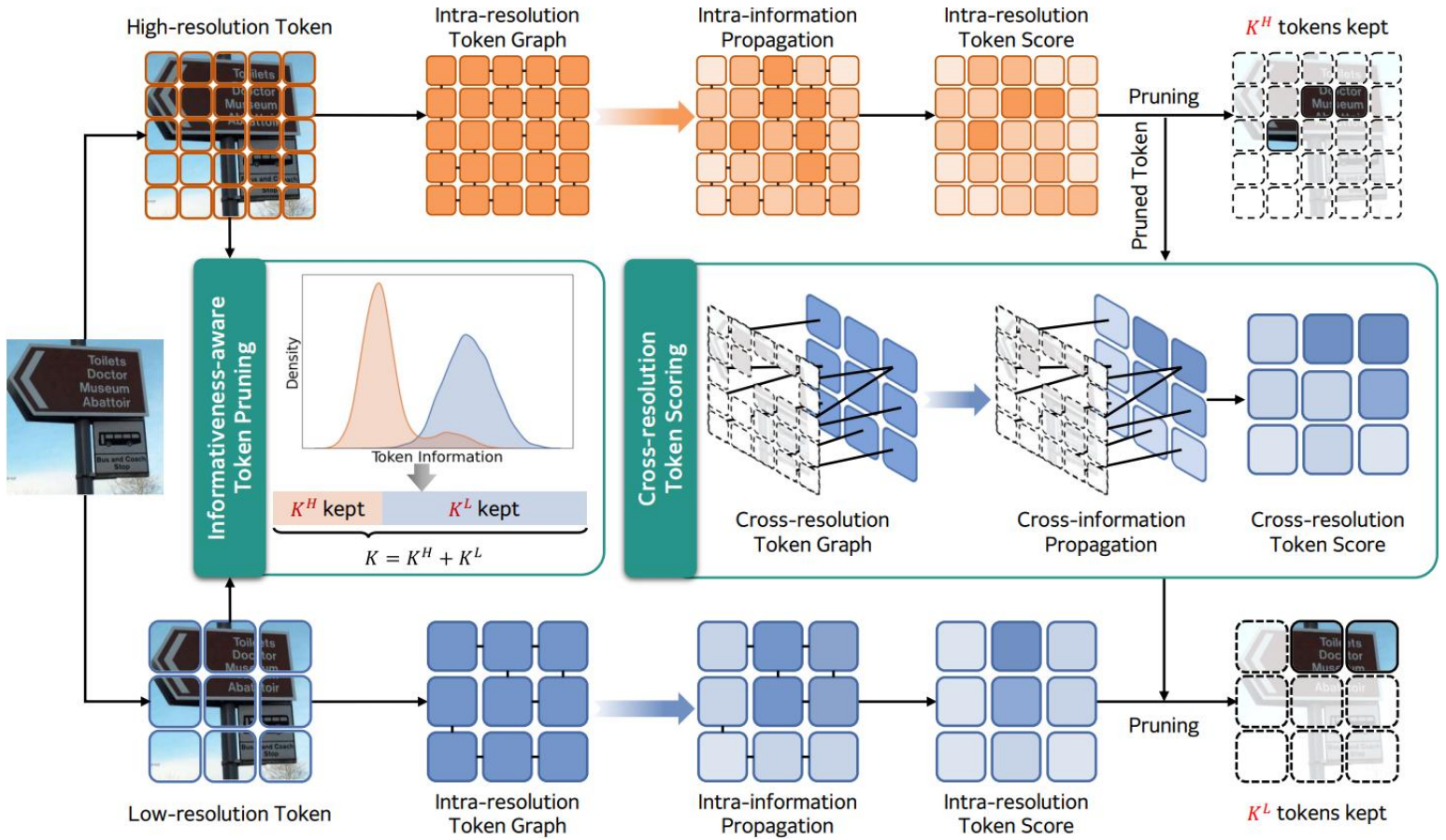


Figure 5. The overview of MR-Pruner.

Experiment

Main Results

Method	Pruning Ratio	GQA	VQA 2.0	MME	POPE	MMB-EN	MMB-CN	TextVQA	SQA-IMG	Throughput
Upper Bound Model										
LLaVA-NeXT-8B	0%	65.38	82.70	1587.72	87.84	72.08	67.18	65.41	73.43	1.46
Single-resolution Pruning Methods										
Random	50%	64.95	81.61	1605.43	86.47	70.27	63.83	58.21	73.48	2.25 (1.54×)
	70%	64.22	80.17	1576.33	84.98	69.42	61.34	49.48	73.53	2.36 (1.61×)
	90%	60.55	74.23	1475.07	79.63	61.77	51.46	31.73	72.43	2.54 (1.74×)
ToMe	50%	65.07	81.82	1566.60	87.56	70.88	64.43	59.07	72.52	0.59 (0.41×)
	70%	64.07	80.56	1564.36	87.33	68.21	61.91	52.19	70.88	0.64 (0.44×)
	90%	59.72	76.36	1453.13	84.29	61.77	53.14	38.36	69.98	0.72 (0.49×)
FastV	50%	65.11	82.51	1604.14	87.51	71.82	65.91	65.15	72.85	2.07 (1.41×)
	70%	64.34	81.83	1600.83	87.08	68.35	62.56	63.08	71.50	2.19 (1.50×)
	90%	60.20	77.21	1488.16	83.01	67.23	56.91	53.53	69.41	2.25 (1.54×)
G-Prune	50%	65.25	82.54	**1623.27**	87.76	71.91	**66.15**	**65.17**	73.53	2.13 (1.45×)
	70%	64.37	81.91	**1604.86**	87.69	70.19	63.74	63.87	72.58	2.26 (1.54×)
	90%	61.40	77.51	1456.14	84.49	67.27	58.59	59.31	71.74	2.31 (1.58×)
Multi-resolution Pruning Method										
MR-Pruner	50%	**65.31**	**82.62**	1597.31	**87.91**	**72.16**	65.98	64.76	**73.87**	**2.24** (1.53×)
	70%	**64.88**	**82.10**	1595.79	**87.80**	**70.96**	**65.12**	**64.13**	**73.62**	**2.35** (1.60×)
	90%	**62.32**	**78.47**	**1530.87**	**85.97**	**68.04**	**60.14**	**60.90**	**72.68**	**2.52** (1.72×)

Thank you