

Multi-modal LLM에서 그래프를 활용한 효과적인 토큰 프루닝

[Ref] MR-Pruner: Training-free Multi-resolution Visual Token Pruning for Multi-modal Large Language Models, WACV 2026

충남대학교 데이터 인텔리전스 연구실

박사과정 한승훈

- Introduction
- Methodology
- Experiments

Introduction

■ Multi-modal LLMs (MLLMs)

- LLMs이 발전함에 따라 multi-modality로 확장한 대규모 언어 모델 (예: LLaVA [1])
- Vision-language task와 같은 다양하고 복잡한 task에서 우수한 성능을 보여줌
- Multi-modality 처리 과정 (예시: vision-language task)
 - ① Vision encoder를 통해 **visual feature** 생성
 - ② 생성된 visual feature를 **LLM의 semantic space로 projection** (Text feature와 visual feature align)
 - ③ Align 된 text feature와 visual feature를 LLM의 input으로 활용

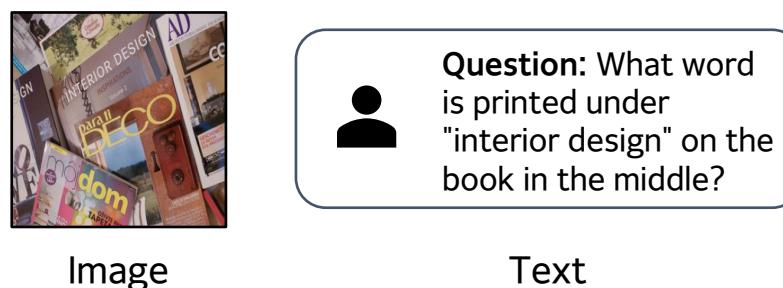


Figure 1. Multi-modal 입력 예시

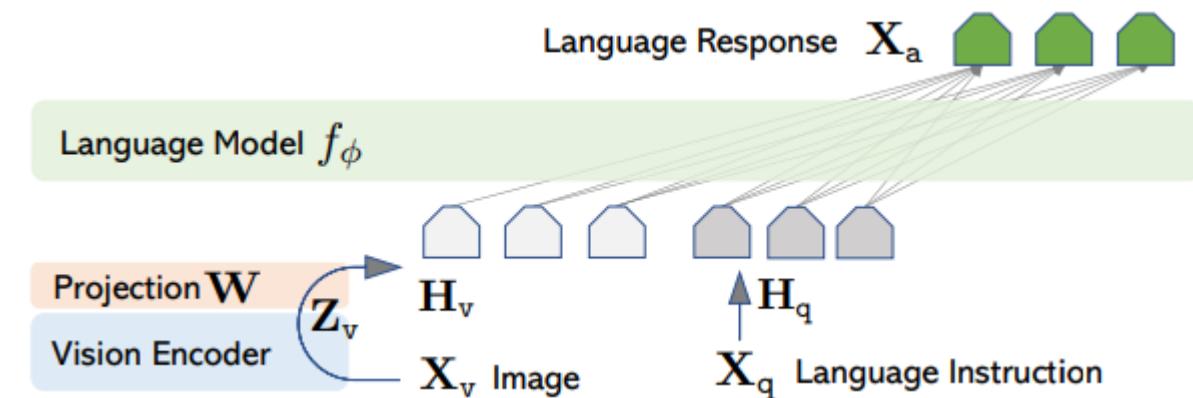


Figure 2. Multi-modal LLM 구조 (예시: LLaVA) [1]

■ Multi-modal LLMs (MLLMs) 문제점

- 일반적으로 text token보다 visual token의 개수가 훨씬 많음
- 한 이미지 내의 visual token들끼리 중복되는 정보를 갖는 경우가 많음
- Video task처럼 수많은 frame이 있는 경우 visual token의 개수가 급격히 증가 → 연산량 급증 $O(N^2)$
 - 중복되는 Visual token의 개수를 줄이는 것 필요
 - Token pruning 필요

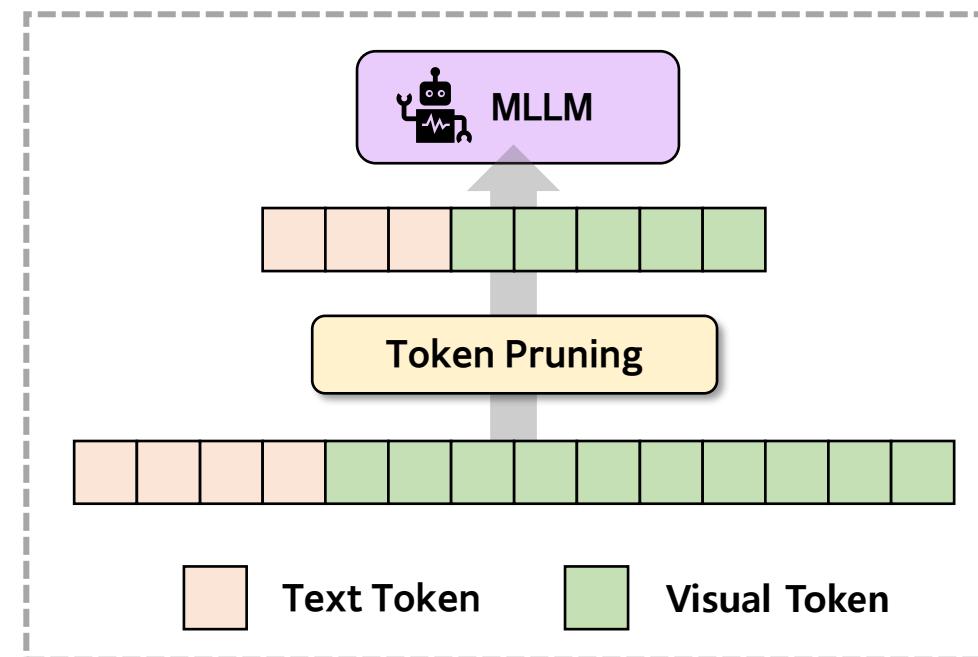


Figure 3. Visual Token Pruning 개념

■ 기존 Token Pruning 연구

- ToME [1]

- 유사한 토큰 쌍을 병합하여 중복되는 토큰을 제거
- 병합 과정에서 정보가 섞이며 정보 손실 발생 가능
- 병합 과정을 위한 처리 시간이 오래 걸릴 수 있음

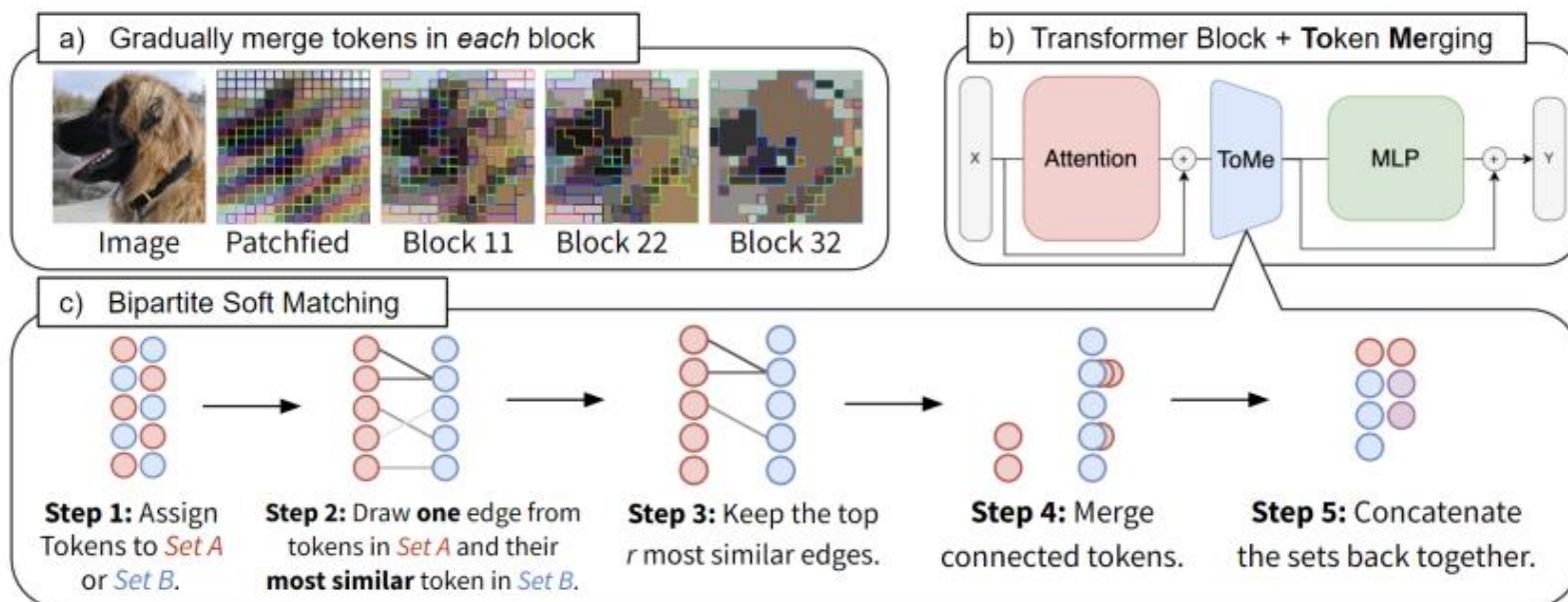


Figure 4. ToME 방법론

■ 기존 Token Pruning 연구

• G-Prune [1] → Target!

- 토큰 프루닝을 그래프로 모델링하여 해결한 방법
- 토큰 간의 관계를 그래프로 구성하고, 그래프 내에서 반복적인 정보 전파를 통해 중요 토큰 선택
- 반복적인 정보 전파를 통한 오버헤드가 커질 수 있음

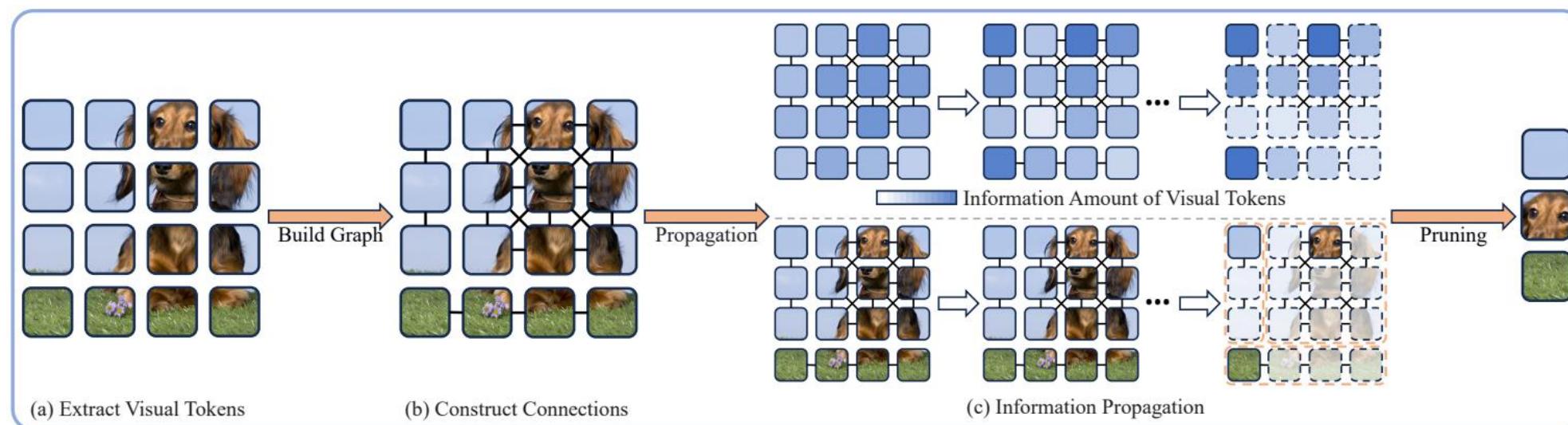


Figure 5. G-Prune 방법론

■ Multi-resolution MLLMs

- MLLMs의 발전에 따라 고해상도를 지원하도록 확장된 멀티모달 대형언어모델 (예: LLaVA-NeXT [1])
 - 기존의 MLLMs보다 OCR과 같은 세밀한 task들을 효과적으로 다룰 수 있음
- 고해상도를 추가로 사용하여 성능 향상을 이뤘지만 기존의 single-resolution MLLMs 보다 훨씬 많은 토큰 필요

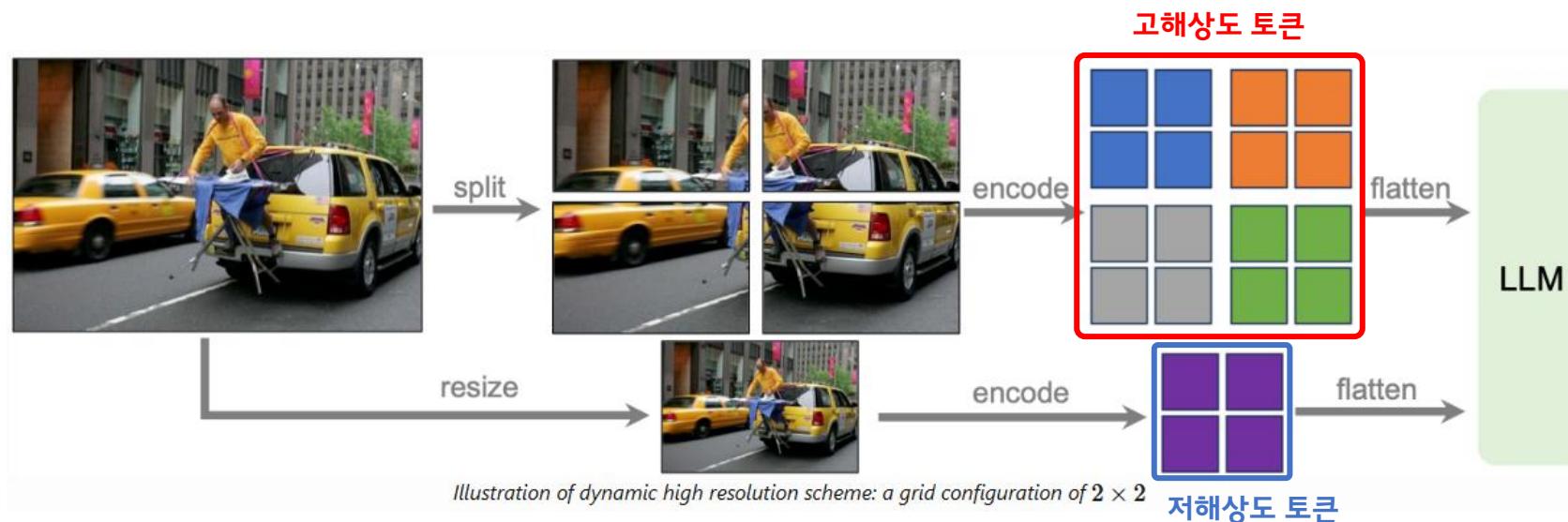


Figure 6. Multi-resolution MLLMs의 visual token 생성 과정

■ 다중 해상도 토큰 특징

- 고해상도 토큰과 저해상도 토큰은 서로 다른 정보량 분포를 가짐 → 정보량을 고려한 프루닝 비율 조절
- 고해상도 토큰과 저해상도 토큰은 mutual complementarity을 가짐 → 교차 해상도 토큰 보존

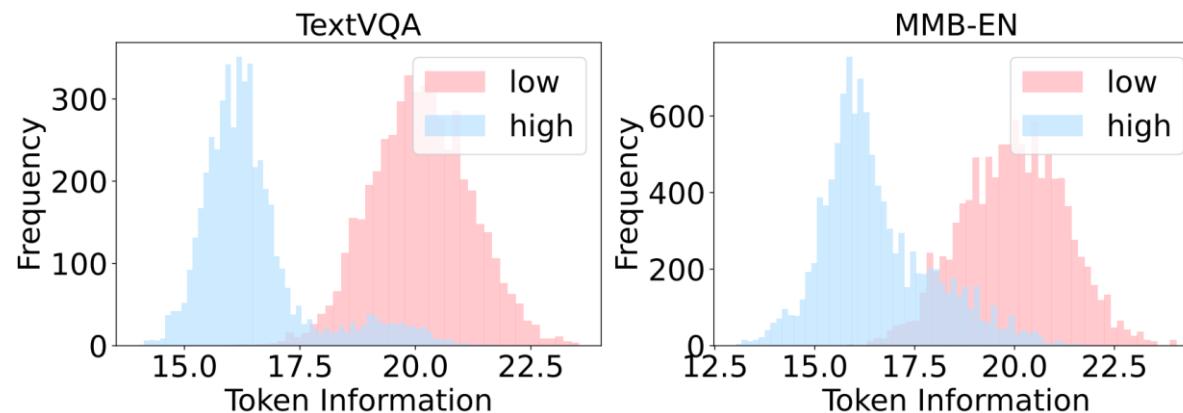


Figure 7. 해상도에 따른 토큰 정보량 비교



Question: What word is printed under "interior design" on the book in the middle?
Answer: Inspiration

Figure 8. Mutual complementarity 예시

■ 제안하는 방법론

- Multi-resolution MLLMs를 위한 **graph-based token pruning** 기법
- **Training-free** 하기 때문에 추가적인 학습 없이 사용 가능 (추가 학습을 위한 자원 소모 X)
- Pre-trained multi-resolution MLLM 모델들에 **plug-and-play 형태로 유연하게 적용 가능**

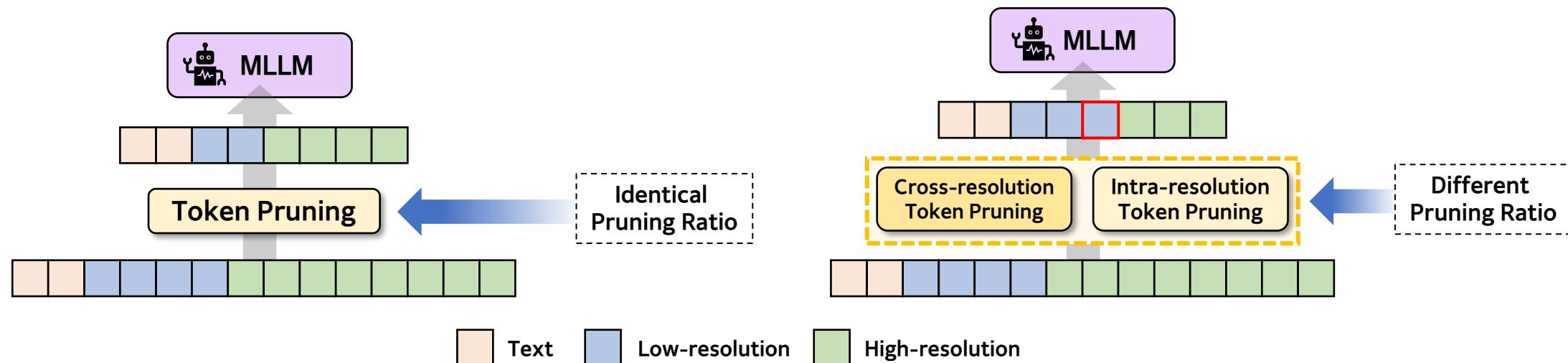


Figure 9. single-resolution token pruning 기법과 제안하는 방법론 비교

Methodology

■ MR-Pruner

- Intra-resolution Token Scoring

- 같은 해상도 내의 토큰들 간의 중요도 점수 측정

- Cross-resolution Token Scoring

- 서로 다른 해상도에 속하는 토큰들 간의 중요도 점수 측정

- Mutual complementarity 특성을 반영 가능

- Informativeness-aware Token Pruning

- 해상도에 따라 갖는 토큰 정보량에 따라 프루닝 비율을 조절 가능

- 서로 다른 해상도에 속하는 토큰의 정보량 분포 차이를 반영 가능

- MR-Pruner

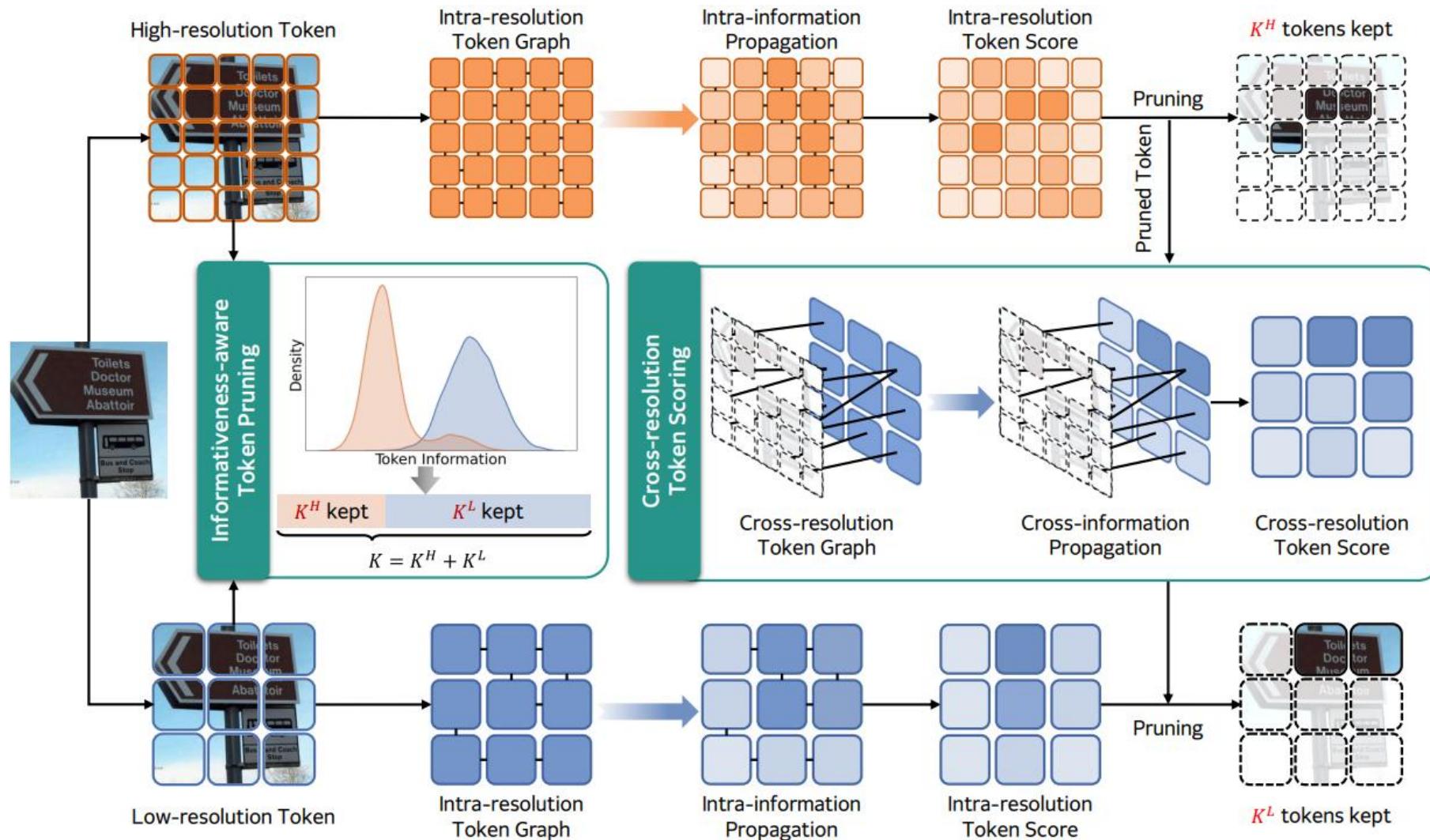


Figure 10. MR-Pruner overview

- MR-Pruner

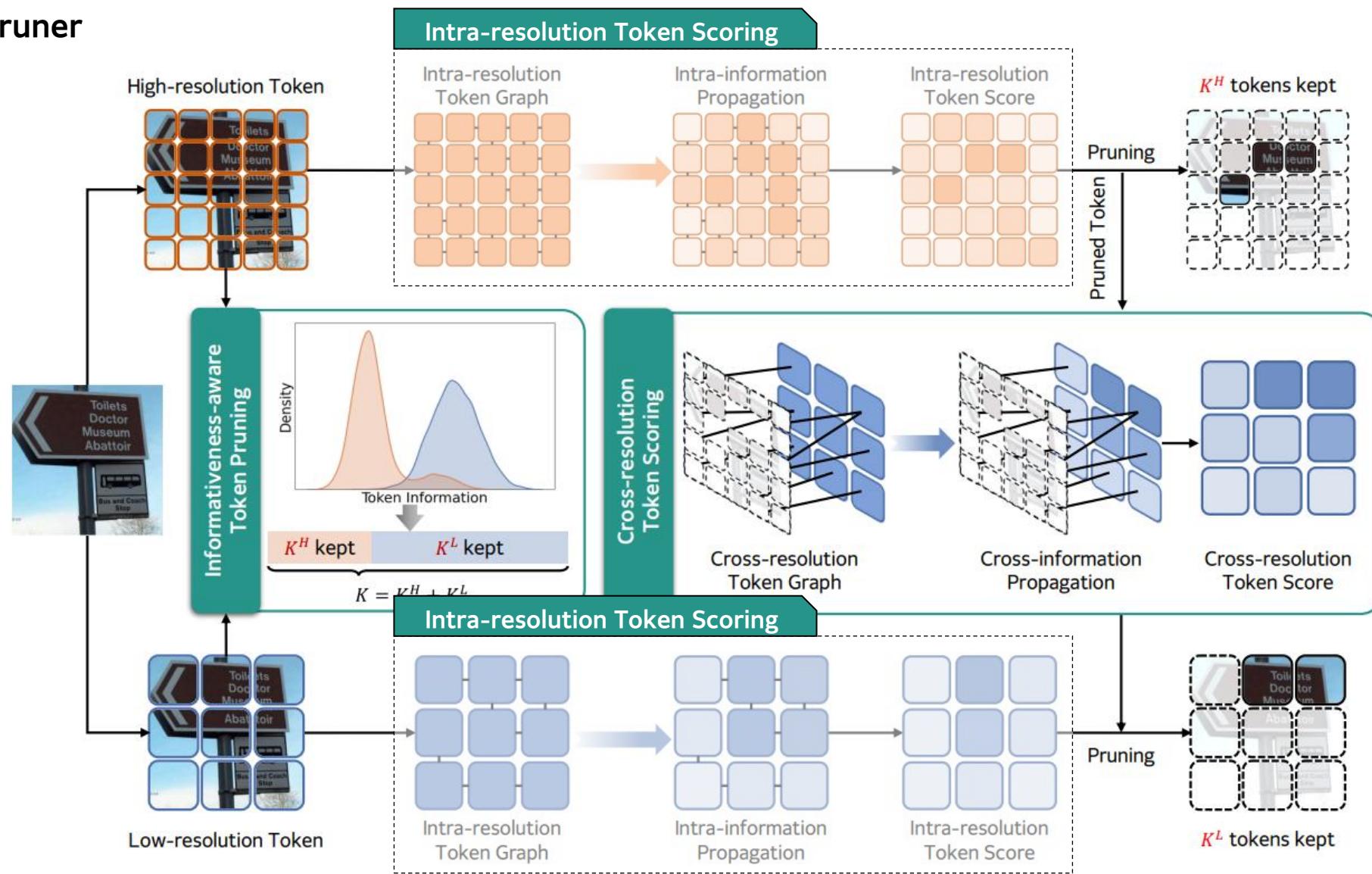
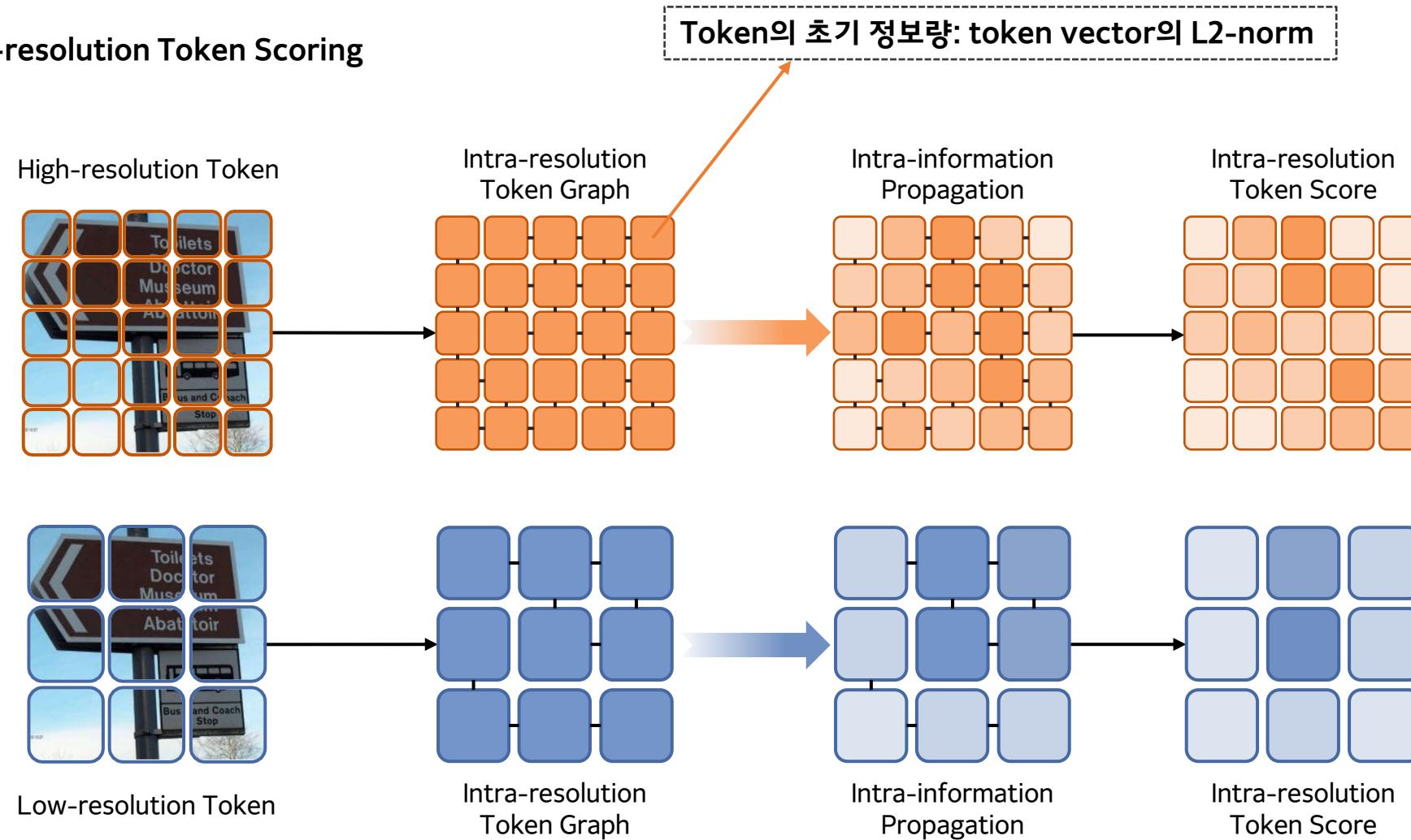


Figure 10. MR-Pruner overview

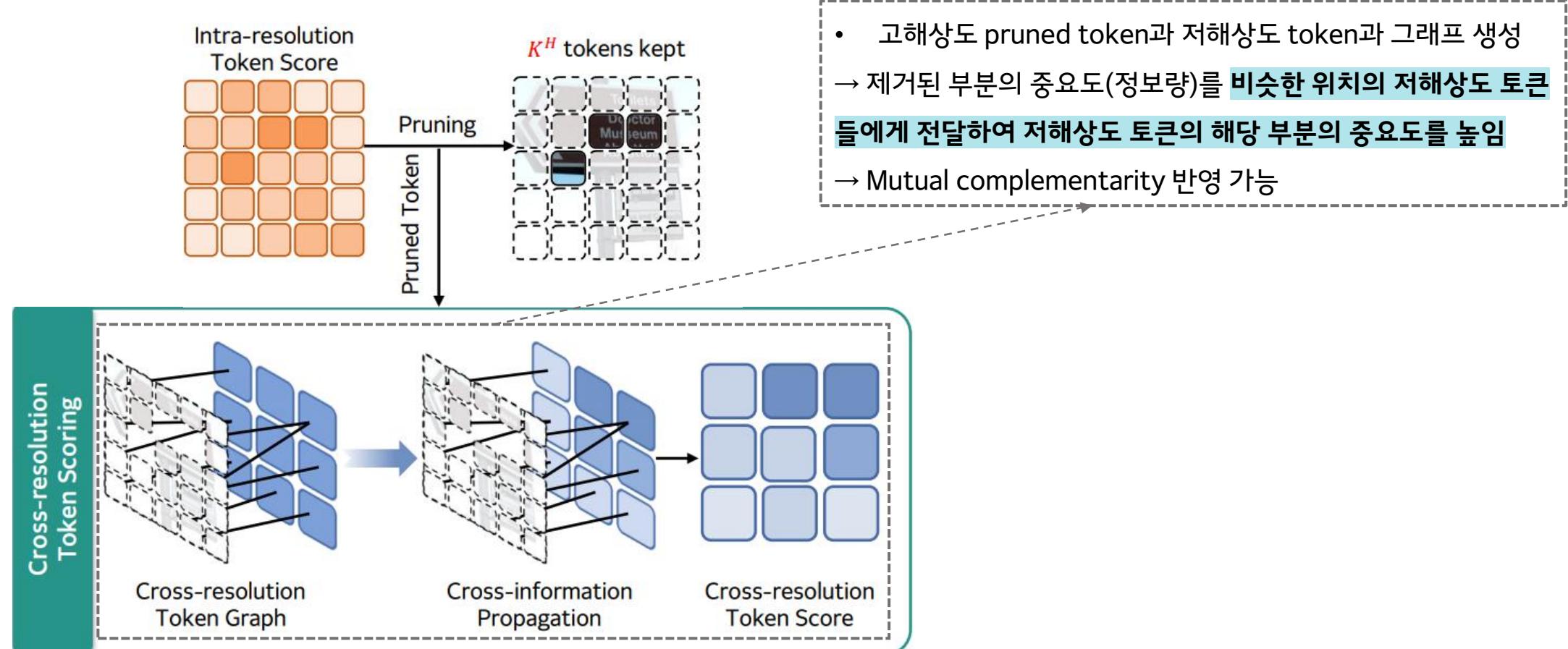
- MR-Pruner

- Intra-resolution Token Scoring



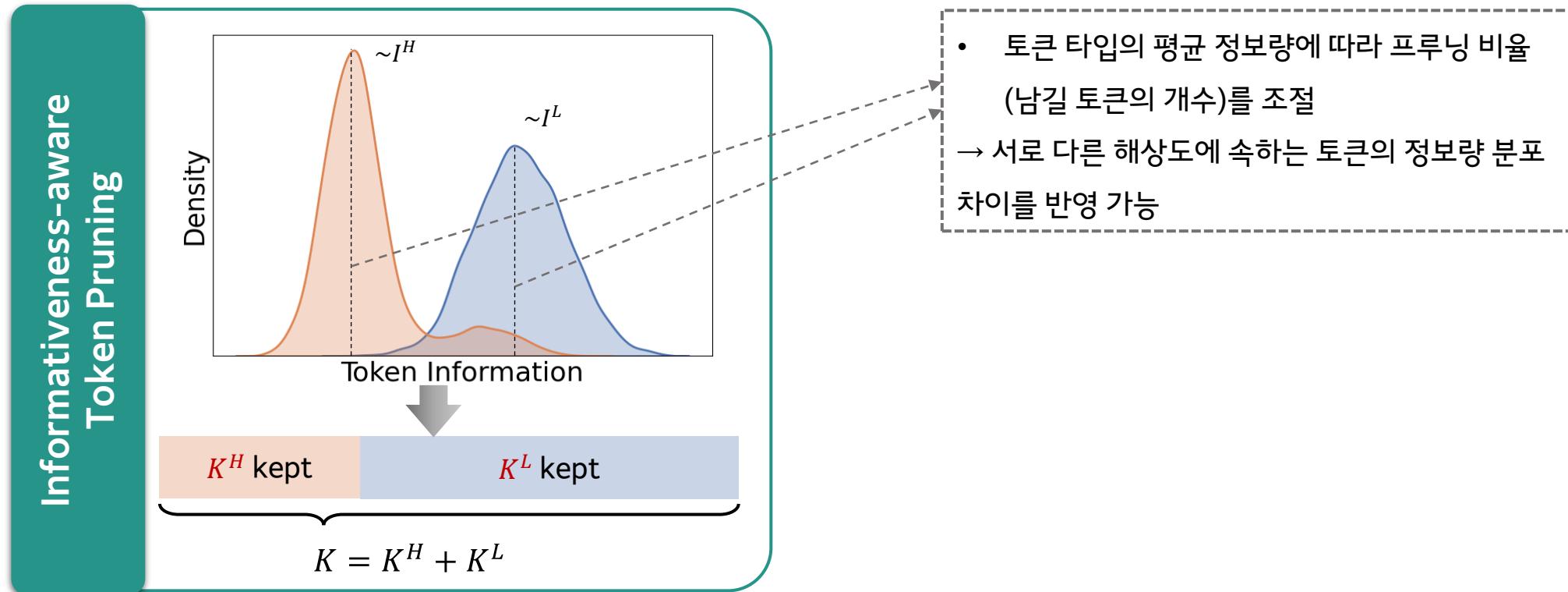
- MR-Pruner

- Cross-resolution Token Scoring



- MR-Pruner

- Informativeness-aware Token Pruning



Experiments

- 실험 설정

- MLLM: *LLaVA-NeXT-8B*
- 데이터셋
 - 총 8개의 MLLM benchmark 데이터셋에서 성능 비교
 - GQA, VQA 2.0, MME, POPE, MMB-EN, MMB-CN, TextVQA, SQA-IMG
- 실험 구현
 - LMMs-eval [1]: MLLMs 모델들의 공정한 평가를 위해 만들어진 Toolkit
- Pruning ratio
 - 50~95% pruning ratio에서 실험 비교
- 평가 지표
 - 정확도 및 처리율

- 메인 실험 결과

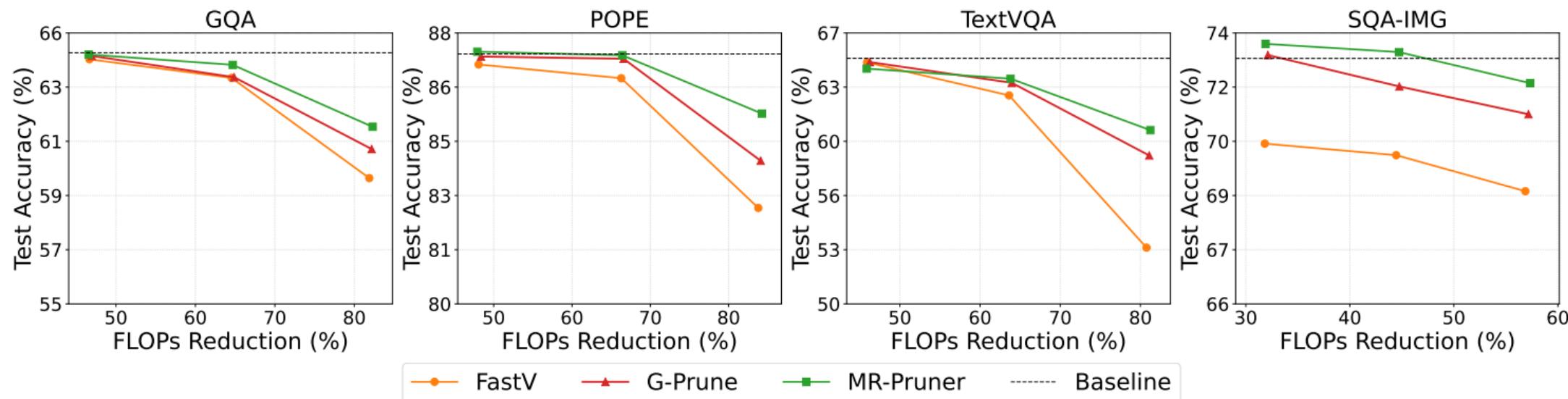
Method	Pruning Ratio	GQA	VQA 2.0	MME	POPE	MMB-EN	MMB-CN	TextVQA	SQA-IMG	Throughput
<i>Upper Bound Model</i>										
LLaVA-NeXT-8B	0%	65.38	82.70	1587.72	87.84	72.08	67.18	65.41	73.43	1.46
<i>Single-resolution Pruning Methods</i>										
Random	50%	64.95	81.61	1605.43	86.47	70.27	63.83	58.21	73.48	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×)
<i>Multi-resolution Pruning Method</i>										
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×)

■ 실험 결과

- 극단적 프루닝 시나리오 (95% pruning ratio) 실험 결과

Method	GQA	MMB-EN	MME	TextVQA
FastV	55.10	59.01	1301.74	49.10
G-Prune	56.42	59.11	1299.95	52.15
MR-Pruner	56.81	62.29	1312.11	54.47

- FLOPs 감소에 따른 성능 변화



감사합니다.