LLaVA-Mini: Efficient Image and Video Large Multimodal Models with One Vision Token

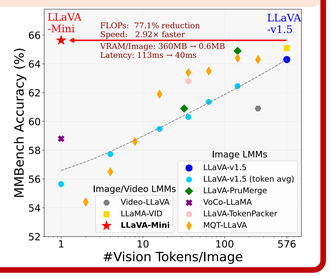
ICLR

Shaolei Zhang[1,3], *Qingkai Fang*[1,3], *Zhe Yang*[1,3], *Yang Feng*[1,2,3,*]

[1]Institute of Computing Technology, Chinese Academy of Sciences [2]Key Laboratory of AI Safety, Chinese Academy of Sciences [3]University of Chinese Academy of Sciences

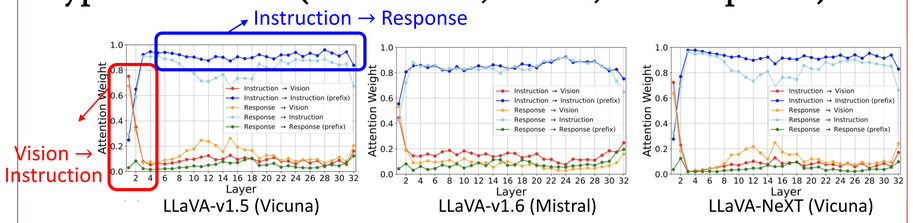
LLaVA-Mini: compress vision tokens that fed into LLM backbone

- Performance
  - Comparable to LLaVA-v1.5
- High efficiency
  - FLOPs: 8.55T → **1.96T**
  - Latency: 113ms → **40ms**
  - VRAM/Image: 360MB → **0.6MB**

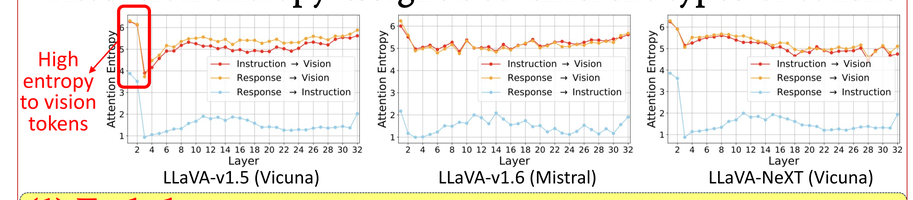


How LMMs Understand Vision Tokens?

- Layer-wise variation of attention weights to different types of tokens (instruction, vision, and response)



- Attention entropy assigned to different types of tokens

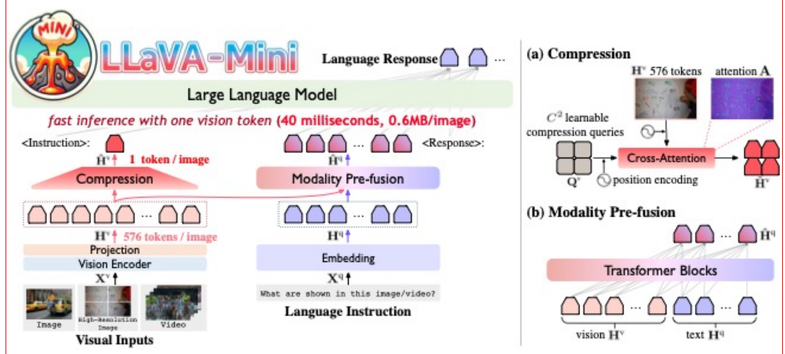


(1) Early layers:
instruction tokens fuse visual information from vision tokens
(2) Later layers:
rely more on instructions tokens to generate responses

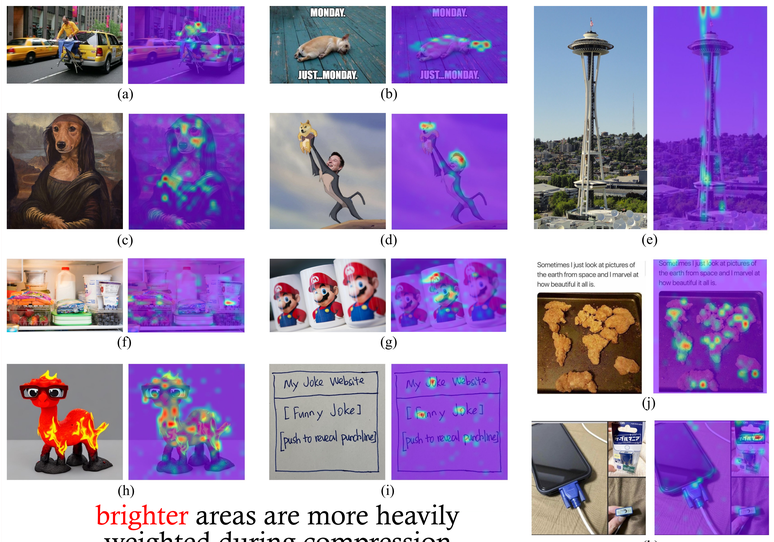
if fuse visual information into instruction in advance, we can reduce vision tokens fed in to LLM

LLaVA-Mini

- **Compression**: compress N^2 vision tokens (i.e., 576) to C^2 tokens (i.e., C=1) with *query-based compression*
- **Modality Pre-fusion**: fuse vision information into instruction tokens in advance



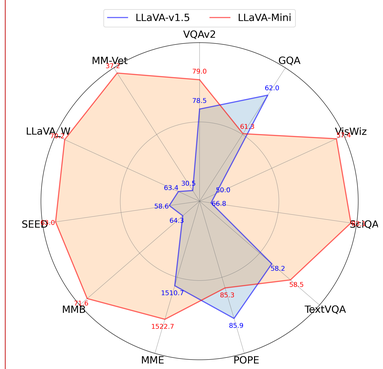
- Compression Visualization

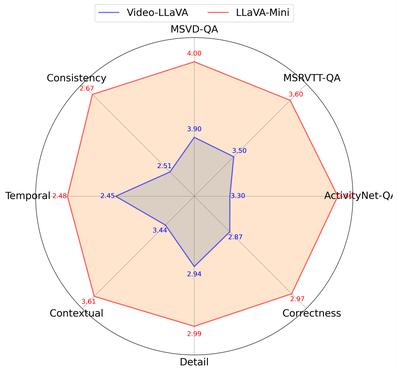


brighter areas are more heavily weighted during compression

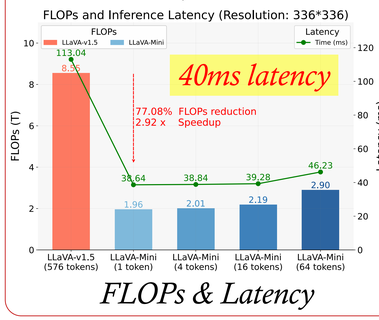
Experiments

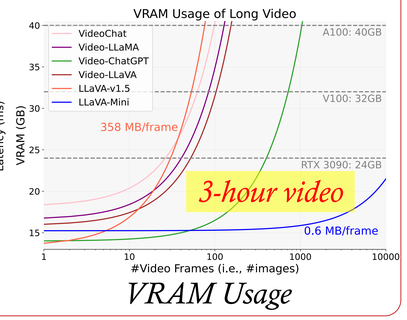
- Image/Video Understanding (1fps for video)





- Efficiency





Homepage Paper Code