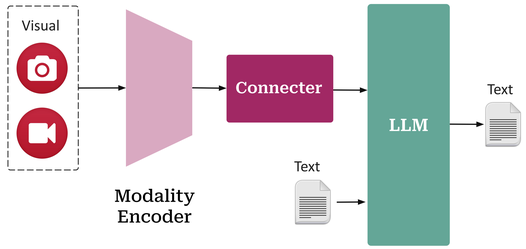
AdaLLaVA: Learning to Inference Adaptively for Multimodal Large Language Models

Zhuoyan Xu[*1], Khoi Duc Nguyen*[*1], Preeti Mukherjee[2], Saurabh Bagchi[2], Somali Chaterji[2], Yingyu Liang[1,3], Yin Li[1]*

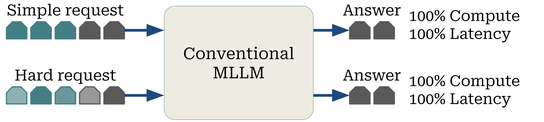
[1]University of Wisconsin-Madison, [2]Purdue University, [3]The University of Hong Kong

ICCV OCT 19-23, 2025 HONOLULU HAWAII

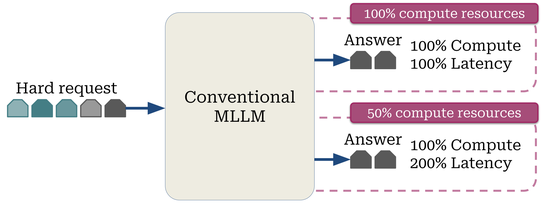
Multimodal Large Language Models



Key Challenge



Fixed model under different input

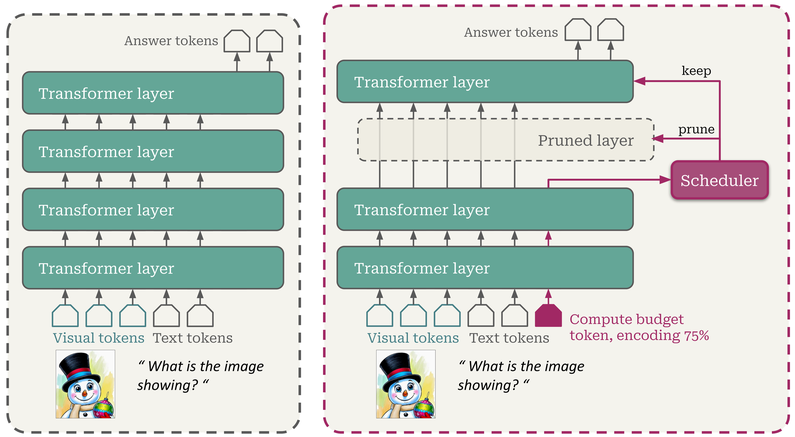


Same content with different compute budget

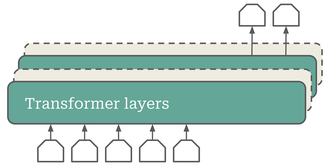
Can we make MLLM adaptive to varying compute resources and input contents?

Goal

Method Outline

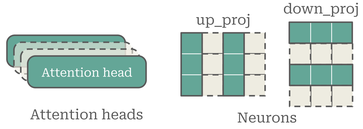


AdaLLaVA-L



Layer-level pruning

AdaLLaVA-H



Attention heads and neuron level skipping

Technical details

AdaLLaVA Probabilistic model:

$$p(x_i^a \mid \mathbf{z}^v, \mathbf{z}^q, \mathbf{z}^a_{<i}, l) = \sum_{\mathbf{s}} p_\theta(x_i^a \mid \mathbf{z}^v, \mathbf{z}^q, \mathbf{z}^a_{<i}, \mathbf{s})\, p_\phi(\mathbf{s} \mid \mathbf{z}^v, \mathbf{z}^q, l)$$

compute budget — execution plan

AdaLLaVA Training:

$$\mathcal{L} = \mathbb{E}_{(\mathbf{z}^v, \mathbf{z}^q, \mathbf{z}^a) \sim \mathcal{D}, l} \left[\mathbb{E}_{\mathbf{s} \sim p_\phi(\mathbf{s} \mid \mathbf{z}^v, \mathbf{z}^q, l)} \left[-\sum_{i=1}^{L} \log p_\theta(x_i^a \mid \mathbf{z}^v, \mathbf{z}^q, \mathbf{z}^a_{<i}, \mathbf{s}) \right] \right]$$

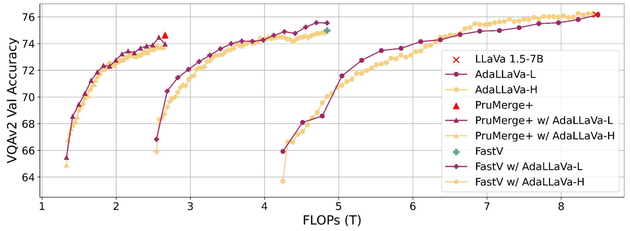
Sample data from dataset. Sample FLOPs budget uniformly in the predefined range.

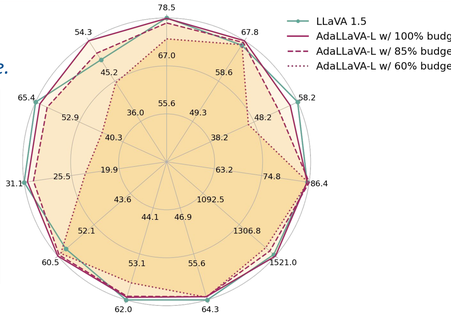
Sample execution plan from the scheduler.

Log likelihood of target next token of the pruned model.

Main results

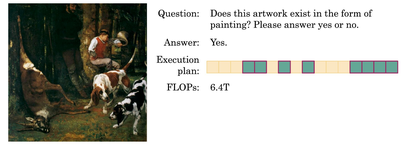
- Models: *LLaVA 1.5, Mipha.*
- Training: *Supervised finetuning with visual instruction data (665K).*
- Metrics: *Accuracy under different latency budget, latency adherence.*



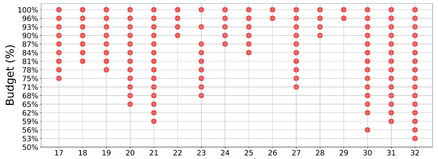


Qualitative results

Latency token attention mapping



Execution plan under different content



Execution plan under different budget

" What is the color of the shirt that the man is wearing?"

60% latency budget

The man is wearing a yellow shirt.

flops	*: 5.1T*
prefill_time	*: 0.05s*
prefill_memory_consumption	*: 8.0G*

80% latency budget

The man is wearing a bright yellow shirt.

flops	*: 6.7T*
prefill_time	*: 0.06s*
prefill_memory_consumption	*: 10.5G*

100% latency budget

The shirt is yellow, matching the taxi.

flops	*: 8.6T*
prefill_time	*: 0.08s*
prefill_memory_consumption	*: 13.4G*

Conclusions

- **Novel adaptive inference** MLLM framework adapt to compute budget and input content.
- **Latency-aware** scheduler and probabilistic modeling approach.
- **Validated across** benchmarks and models, integrates with token selection techniques.

This work was supported by National Science Foundation (CNS-2333487/2333491, CNS-2146449, IIS-2442739), Army Research Lab (W911NF-2020221), Google, and AWS.