Lifting the Veil on Visual Information Flow in MLLMs: Unlocking Pathways to Faster Inference

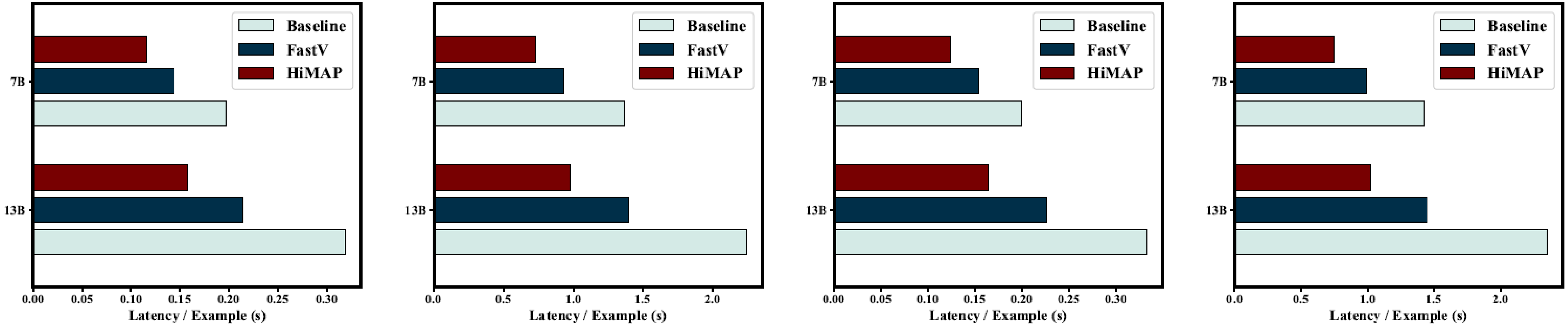
CVPR 2025

Hao Yin, Guangzong Si, Zilei Wang

中国科学技术大学
University of Science and Technology of China

Contributions

- Identifying latent patterns in the interactions between visual and textual modalities within MLLMs
- Introducing HiMAP, a plug-and-play technique that reduces inference latency in MLLMs while maintaining performance

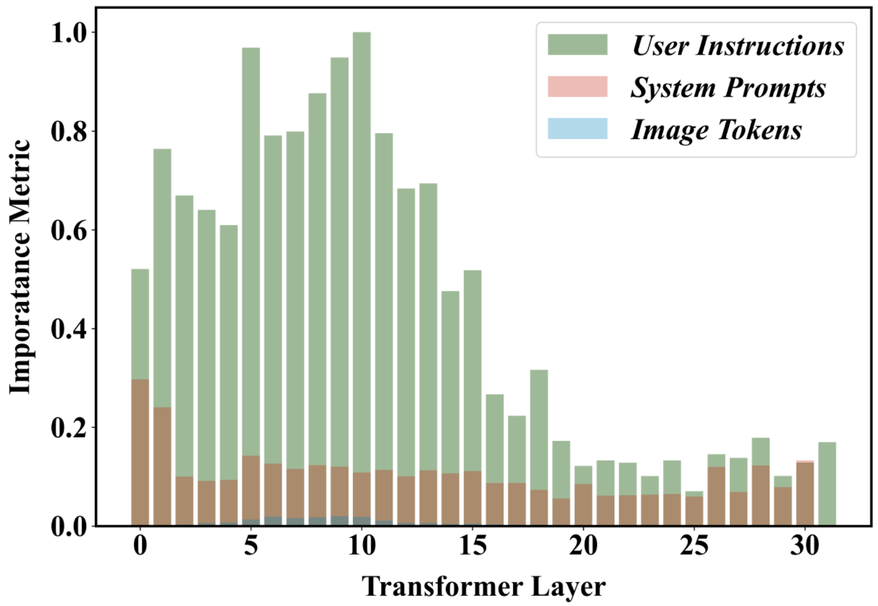


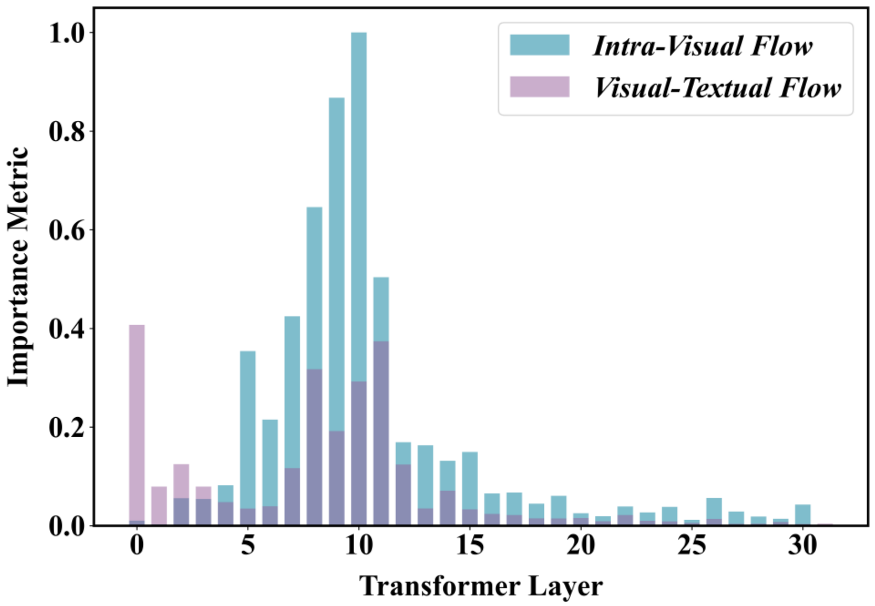
(a) ScienceQA Dataset (b) A-OKVQA Dataset (c) Nocaps Dataset (d) Flickr30k Dataset

Hypothesis Driven by Saliency Scores

$$I_l = \left| \sum_h A_{h,l} \odot \frac{\partial \mathcal{L}(x)}{\partial A_{h,l}} \right|.$$

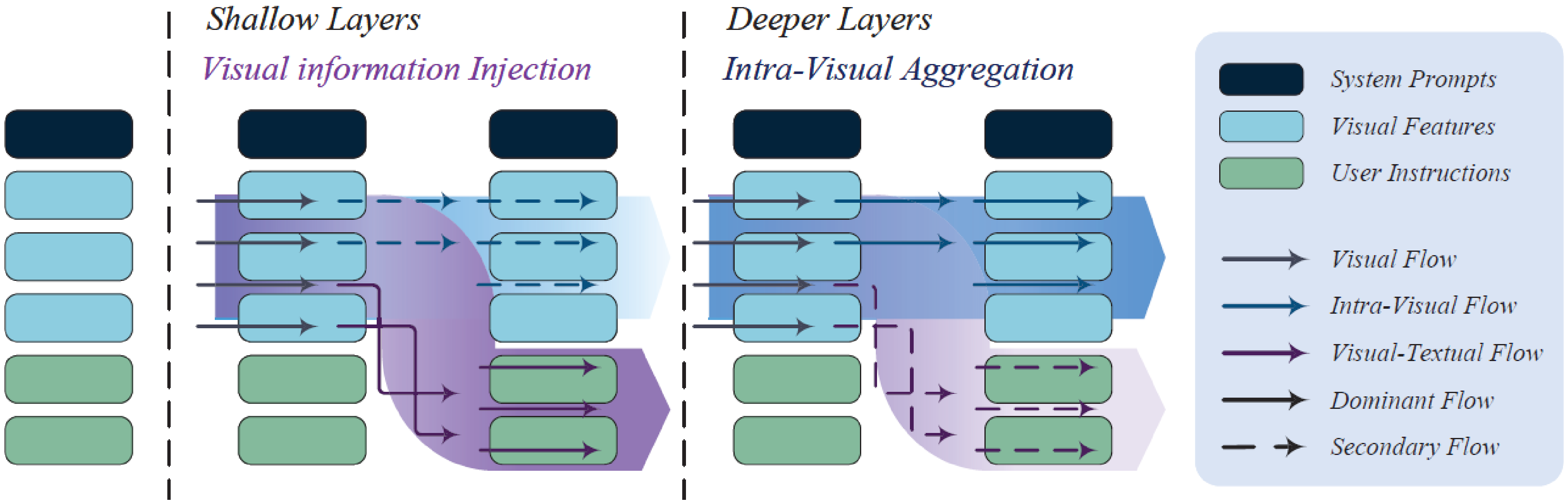
The saliency matrix I_l for the l-th layer is obtained by averaging across all attention heads. The significance of information flow from the j-th token to the i-th token in MLLMs is represented by $I_l(i, j)$.





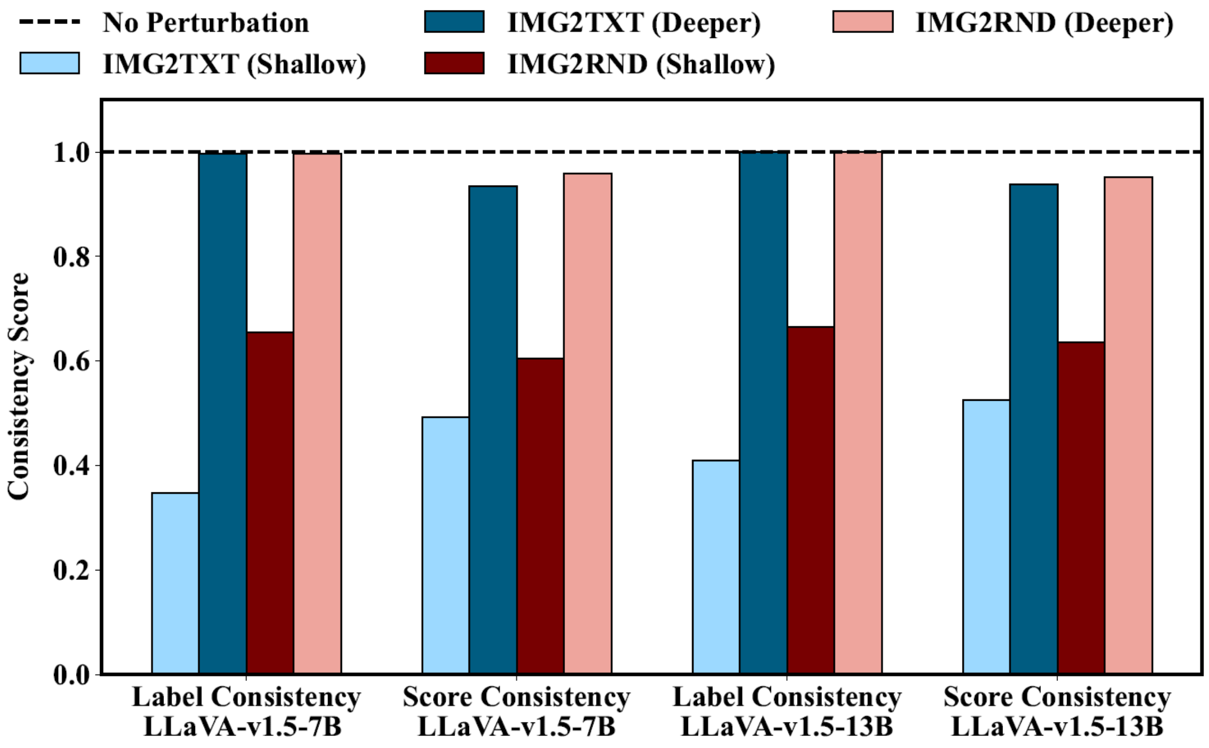
Shift in Dominant Flow of Visual Information

- In shallow layers, image tokens inject visual information into instruction tokens, facilitating cross-modal semantic representations for subsequent computations
- In deeper layers, image tokens consolidate residual visual information, refining the semantic representation within the visual modality



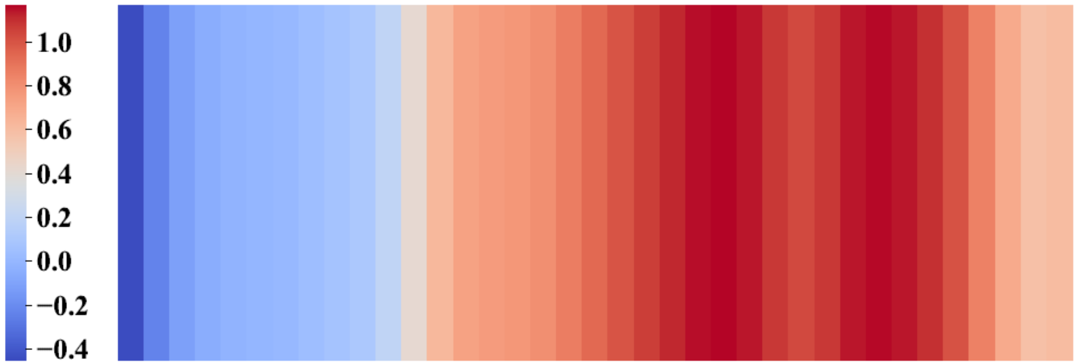
Shallow Layers: Visual Information Injection

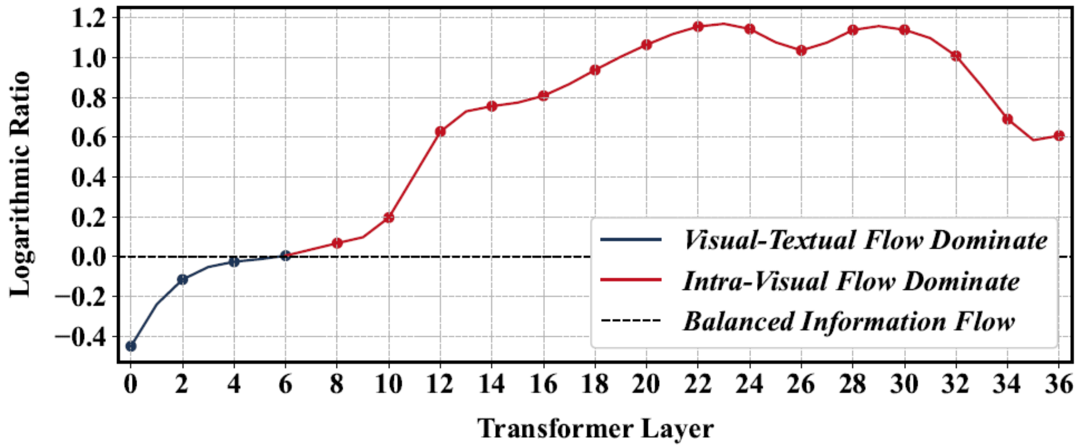
- To examine how visual information is integrated into instruction tokens, we blocked the interaction between image and instruction tokens at specific layers of the model.
- We observed that perturbations in the shallow layers significantly degraded model performance, whereas disruptions in the deeper layers had a much smaller impact.



Deeper Layers: Intra-Visual Aggregation

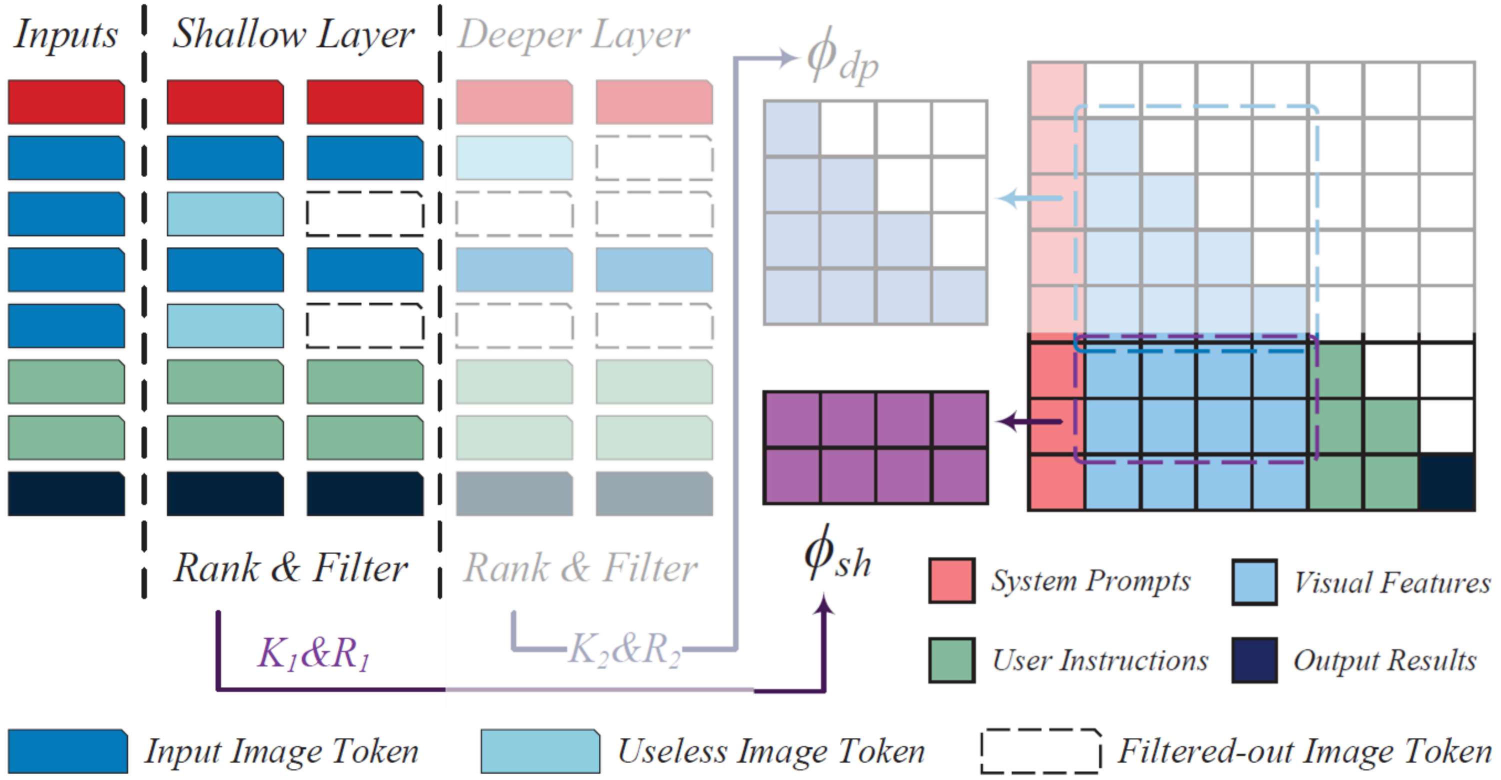
- We compared the relative importance of *visual-textual* and *intra-visual information flows* across different model depths.
- We observed that disrupting *intra-visual information flow* in deeper layers resulted in more substantial deviations in prediction outcomes.





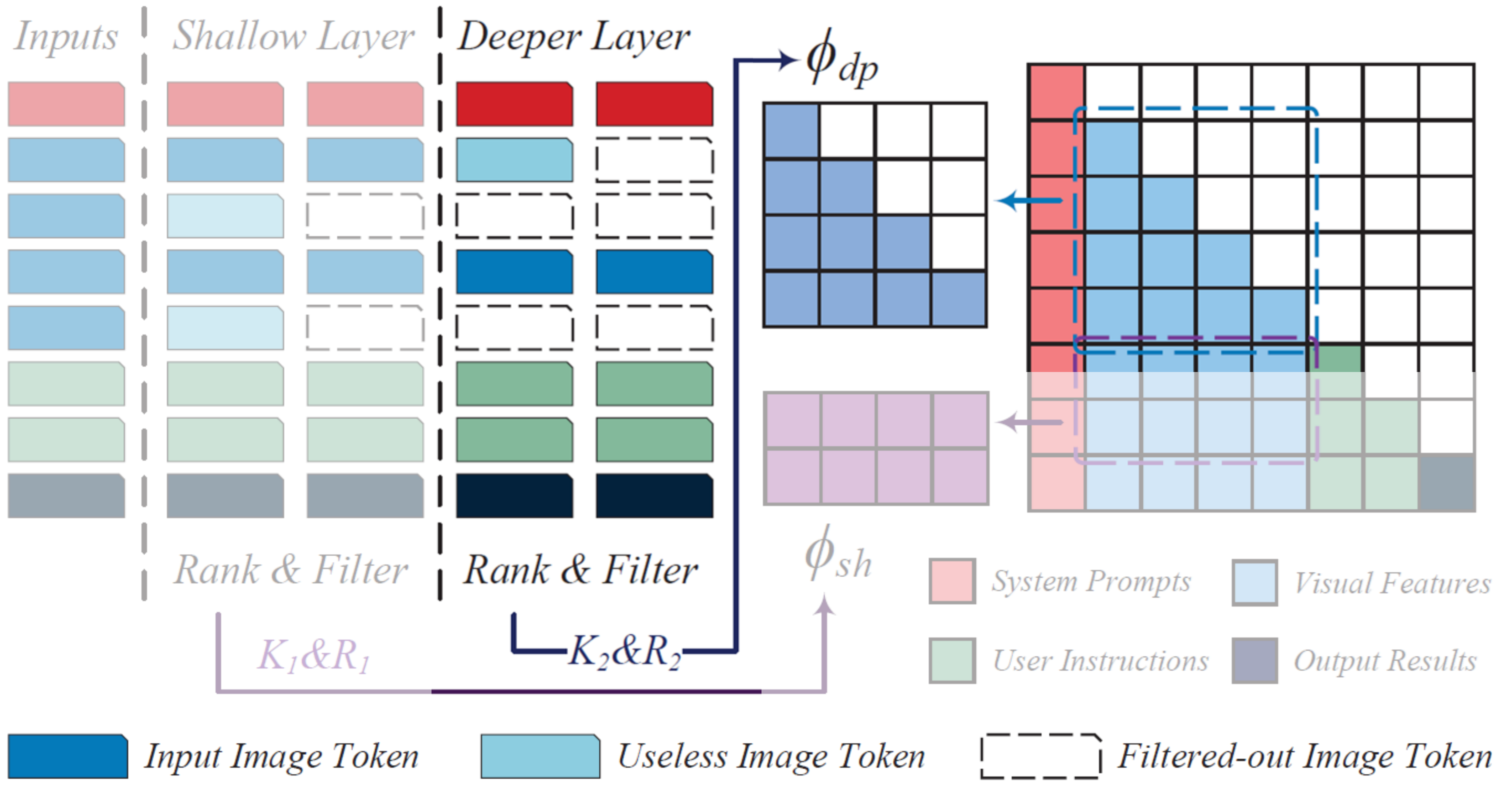
Hierarchical Modality-Aware Pruning

In shallow layers, HiMAP ranks image tokens at the K_1-th layer based on the importance criterion ϕ_{sh}, removing the image tokens in the bottom $R_1\%$.



Hierarchical Modality-Aware Pruning

In deeper layers, HiMAP ranks the remaining image tokens at the K_2-layer based on the importance criterion ϕ_{dp}, filtering out those in the bottom $R_2\%$.



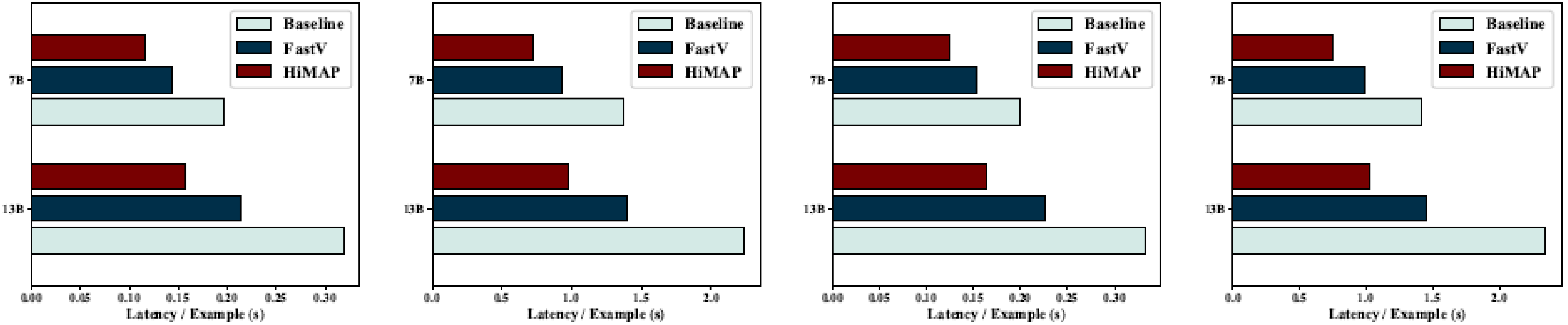
Quantitative Results: Prediction Performance

- Improved performance is observed in nearly all short-answer tasks
- Effectively showcases the precision of redundant token elimination

Model	Method	TFLOPs	FLOPs Ratio	VQAv2	T-VQA	POPE	MME	S-VQA	A-OKVQA
LLaVA-7B	Baseline	2.98	100%	**78.3**	58.2	**86.4**	**1749.9**	67.9	76.6
	FastV	1.56	54%	78.1	**58.3**	84.9	1742.6	**68.1**	**77**
	HiMAP	**0.73**	**24%**	**78.6**	**58.4**	**86.2**	**1785.1**	**68.3**	**77.2**
LLaVA-13B	Baseline	5.81	100%	79.8	61.4	**87.2**	1794.4	**71.6**	**82**
	FastV	3.09	53%	**79.9**	**61.4**	84.8	**1796.3**	71.3	81.3
	HiMAP	**1.36**	**23%**	**80.2**	**61.7**	**86.5**	**1809.4**	**72.1**	**81.4**
QwenVL-7B	Baseline	3.6	100%	78.4	**60.8**	**84.5**	**1782.6**	68	75.7
	FastV	1.9	53%	**78.5**	58.3	82.7	1767.2	**68.2**	**75.3**
	HiMAP	**0.89**	**25%**	**78.8**	**61.3**	**83.7**	**1798.3**	**68.5**	**75.9**

Quantitative Results: Inference Speed

- Compared to the baseline, inference is approximately 50% faster
- Compared to FastV, inference is around 25% faster



(a) ScienceQA Dataset (b) A-OKVQA Dataset (c) Nocaps Dataset (d) Flickr30k Dataset

Figure 8. **Comparison of real-world inference speeds between HiMAP and FastV.** The experiment was conducted using LLaVA-v1.5 model family on a server equipped with a single 80GB A800 GPU.

Conclusions

Phased Processing of Visual Information

- In shallow layers, strong interactions are observed between image tokens and instruction tokens, where most visual information is injected into instruction
- In deeper layers, image tokens primarily interact with each other, aggregating the remaining visual information

Hierarchical Modality-Aware Pruning

- Inference speed is approximately 50% faster than the baseline, and about 25% faster than FastV