

# ShortV: 通过冻结冗余层的视觉标记, 实现高效多模态大语言模型

ShortV: Efficient Multimodal Large Language Models by Freezing Visual Tokens in Ineffective Layers

袁千皓, 张清宇, 刘衍江, 陈嘉伟,  
陆垚杰, 林鸿宇, 郑佳, 韩先培, 孙乐

International Conference on Computer Vision 2025 (ICCV 2025)

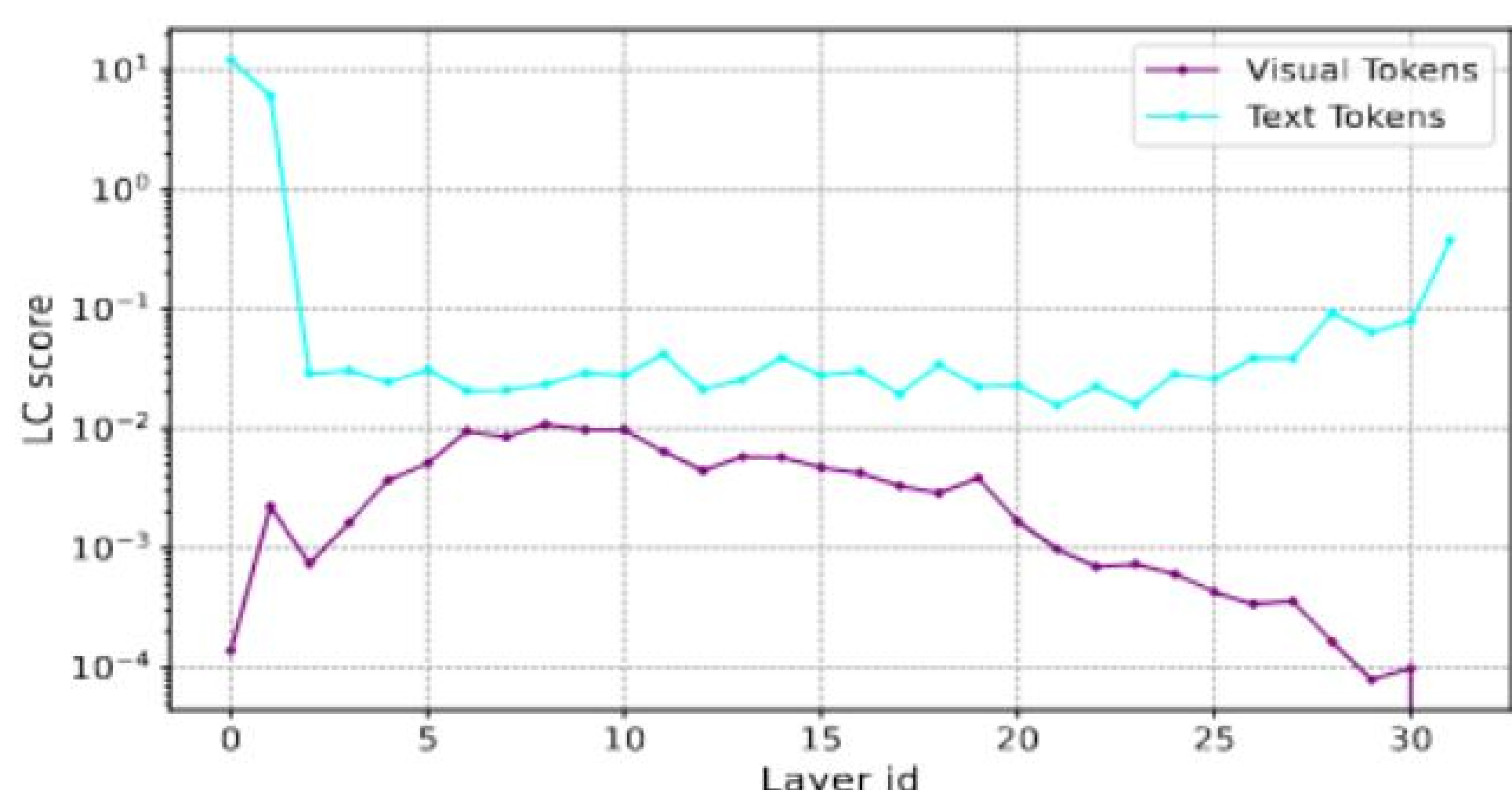
联系人: 袁千皓, yuanqianhao2024@iscas.ac.cn

## 研究背景

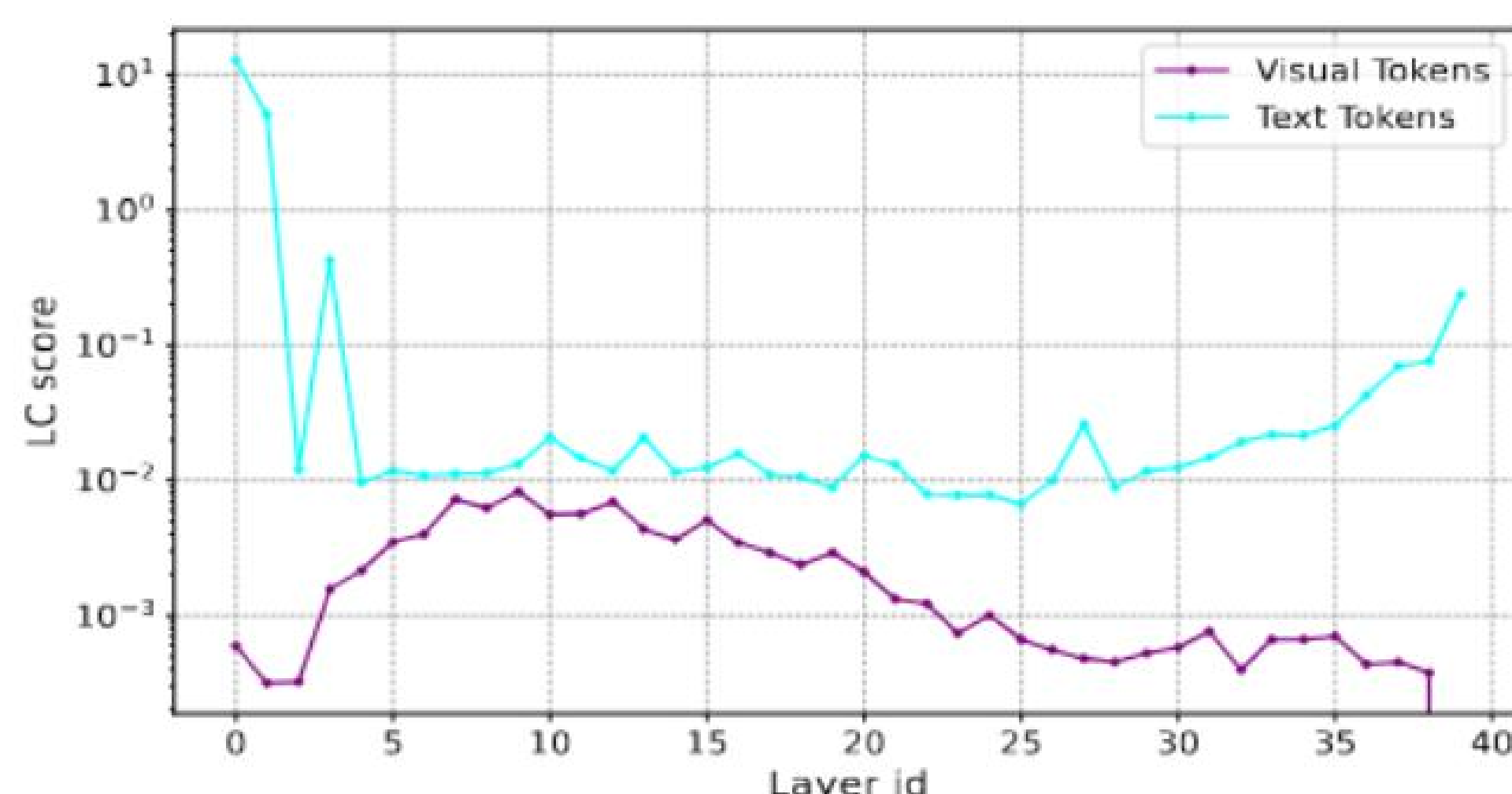
- 多模态大语言模型在处理视觉信息时, 由于模型规模庞大和视觉标记数量多, 计算成本高昂。
- 现有的方法主要关注视觉标记级别的冗余。

## 多模态大语言模型的层冗余性

提出了一种新的度量标准, 层贡献度 (LC), 用于量化某一层中视觉和文本标记更新对模型输出的贡献。



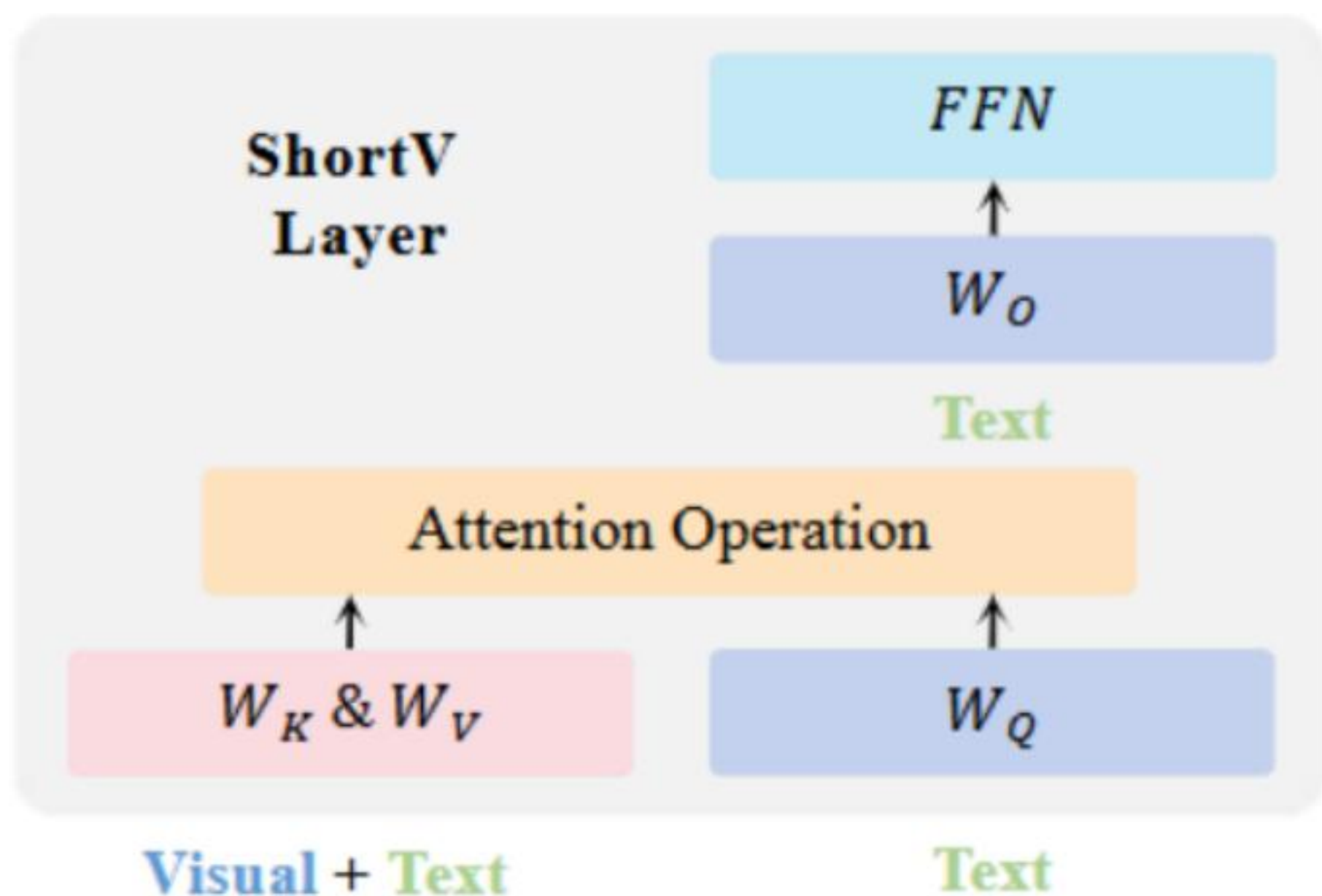
(a) LLaVA-1.5-7B



(b) LLaVA-1.5-13B

多模态大语言模型在处理视觉标记时存在显著的层冗余。

## 方法: ShortV



用 LC 指标识别模型中处理视觉标记的  $N$  个无效层, 即 LC 指标较小的层, 并将这些层替换为 ShortV 层 (左图)。在 ShortV 层中, 视觉标记不更新, 不作为注意力模块的查询, 也不参与前馈网络的计算, 从而实现高效的多模态大模型。

## 实验结果与分析

在多个多模态大模型中, ShortV 均在保持模型性能的同时, 减少了超过 45% 的计算量 (TFLOPs)。

Method	TFLOPs	FLOPs Ratio	VQAv2	GQA	SEED-Bench	MMMU (val)	MME	MMBench EN	MMStar
<i>LLaVA-1.5-7B</i>									
Vanilla	8.5	100%	76.5	61.9	66.1	36.3	1510.7	64.1	33.7
FastV ( $K=2, R=50\%$ )	4.9	58%	73.5	60.2	65.4	35.8	1475.6	64.3	32.4
VTW ( $K=16$ )	4.7	55%	66.3	55.1	66.2	36.1	1497.0	64.0	32.8
ShortV (Ours, $N=19$ )	4.7	55%	75.7	60.9	66.2	36.2	1503.1	64.8	33.3
<i>LLaVA-1.5-13B</i>									
Vanilla	16.6	100%	78.0	63.3	68.2	35.4	1531.3	68.9	36.1
FastV ( $K=2, R=50\%$ )	9.4	57%	76.7	59.4	67.8	34.6	1506.6	68.3	35.9
VTW ( $K=20$ )	9.1	55%	75.3	60.6	68.2	34.9	1533.0	68.5	36.1
ShortV (Ours, $N=24$ )	9.1	55%	77.2	62.0	68.0	35.8	1535.9	68.6	37.1
<i>LLaVA-NeXT-7B</i>									
Vanilla	42.7	100%	80.0	64.1	70.2	36.4	1519.0	67.1	37.1
FastV ( $K=2, R=50\%$ )	22.0	52%	79.5	63.0	69.6	35.1	1482.0	66.3	36.5
VTW ( $K=16$ )	21.8	51%	75.6	55.8	70.2	35.7	1518.2	67.1	37.6
ShortV (Ours, $N=19$ )	21.6	51%	78.8	63.4	70.4	36.0	1525.1	67.2	37.8
<i>LLaVA-NeXT-13B</i>									
Vanilla	81.8	100%	80.9	65.7	71.9	35.9	1570.0	69.3	39.9
FastV ( $K=2, R=50\%$ )	42.1	51%	76.8	62.9	71.5	35.9	1546.4	68.5	39.6
VTW ( $K=20$ )	41.7	51%	77.0	61.5	71.8	34.8	1569.4	69.1	39.8
ShortV (Ours, $N=24$ )	41.0	50%	79.7	63.6	71.8	36.2	1553.0	70.2	39.9