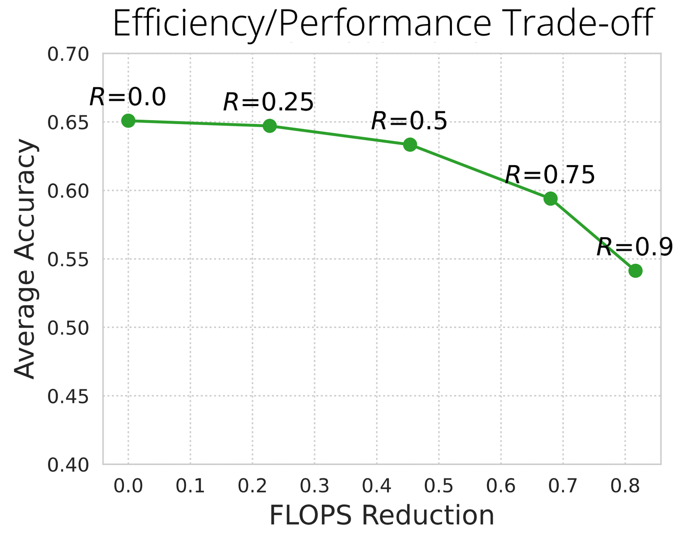
Feather the Throttle: Revisiting Visual Token Pruning for Vision-Language Model Acceleration

Mark Endo, Xiaohan Wang, Serena Yeung-Levy

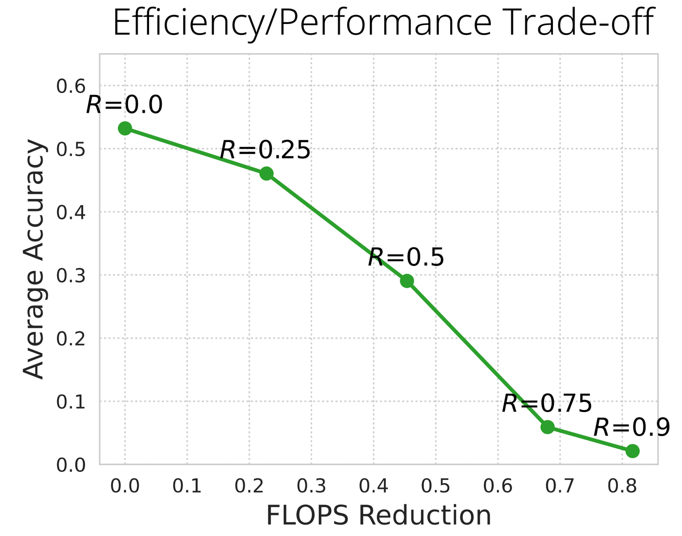
ICCV HONOLULU HAWAII OCT 19-23, 2025

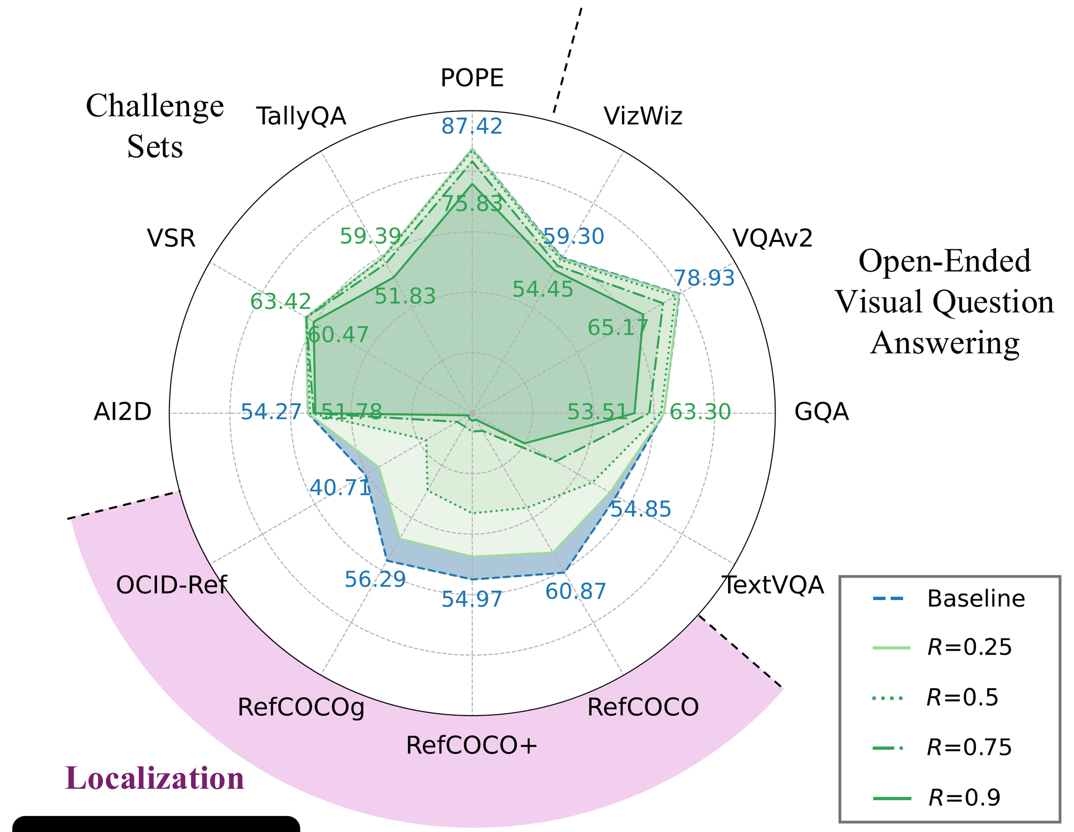
Impact of early token pruning varies drastically by task

Non-localization
e.g., what animal is in front of the trees?



Localization
e.g., where is the player in white shirt and black shorts?

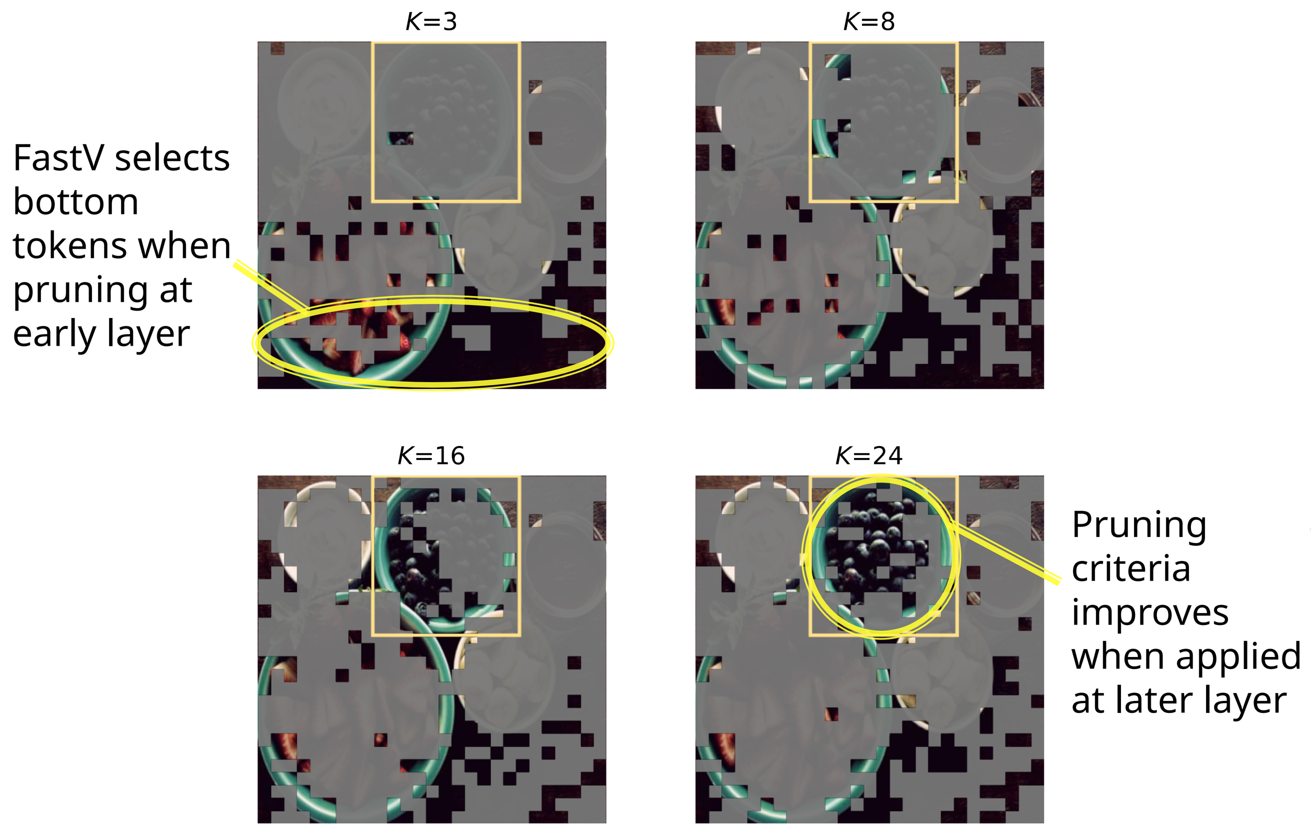




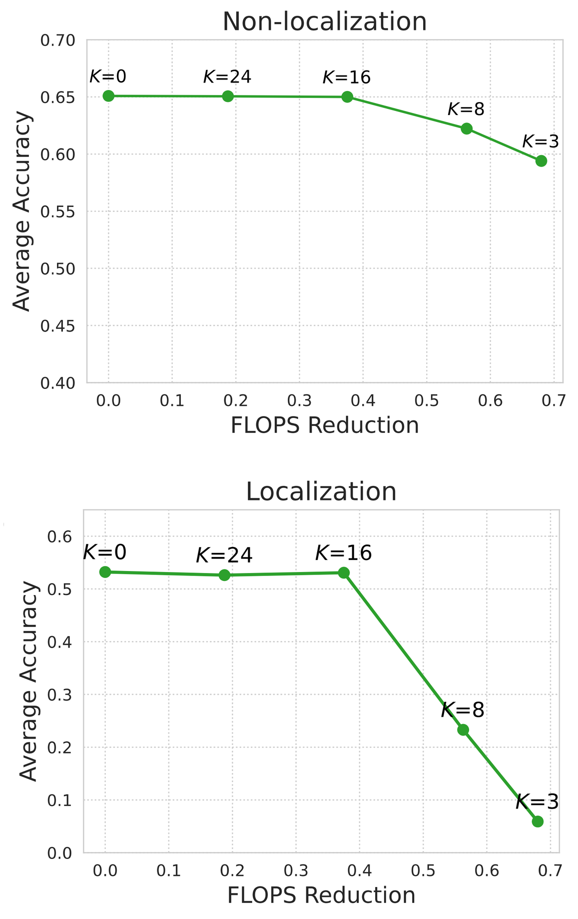
Finding 1

Pruning visual tokens after shallow LLM layers results in a jarring performance decline for localization benchmarks and a moderate decrease for TextVQA, whereas performance remains relatively unchanged for other evaluated tasks.

Interpreting poor vision-centric task performance



Reference expression: a bowl of blueberries

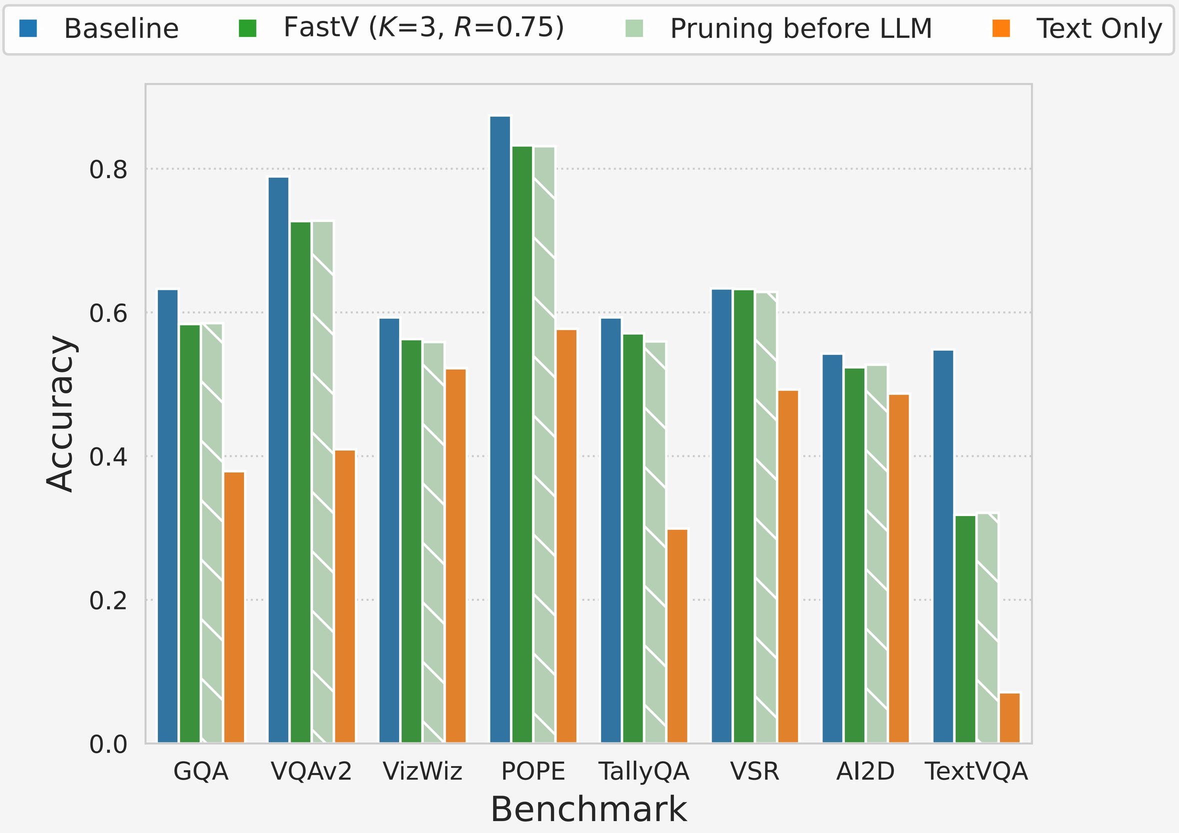


Finding 2

The variable performance across pruning layers is linked to the effectiveness of selecting important tokens, as early token pruning leads to suboptimal token selection biased toward bottom tokens due to the long-term decay property of RoPE.

How does the pruning approach still maintain high performance across wide range of tasks aimed to evaluate visual capabilities?

- **Explanation 1:** Important visual information from pruned image tokens is transferred to maintained tokens before pruning
- **Explanation 2:** Many questions can still be inferred with access to only the suboptimal set of maintained visual tokens



Finding 3

High performance of early visual token pruning on many tasks does not stem from the effectiveness of visual information transfer in early layers but rather signifies that **many benchmarks do not demand a detailed understanding of visual information**.

Improving visual token pruning

Pruning Layer	Criteria	FLOPS Red	Localization Avg	OCID-Ref	RefCOCOg	RefCOCO+	RefCOCO	Open-Ended VQA Avg	TextVQA	GQA	VQAv2	VizWiz	Challenge Sets Avg	POPE	TallyQA	VSR	AI2D
$K=3$	*Attention-based*																
	ϕ_{original}	68%	5.9	5.7	5.1	6.1	6.7	54.8	31.8	58.4	72.7	56.3	64.0	83.2	57.1	**63.3**	52.4
	$\phi_{\text{-}R}$ (Ours)	68%	16.7	22.9	15.1	13.3	15.3	59.0	41.6	61.2	76.0	57.3	64.7	85.2	58.2	62.2	53.2
	Δ		+10.7	+17.2	+10.0	+7.3	+8.6	+4.2	+9.7	+2.8	+3.2	+1.1	+0.7	+1.9	+1.1	-1.1	+0.8
	Non-attention-based																
	ϕ_{KNN}	66%	23.9	15.1	24.9	26.0	29.6	58.4	39.9	60.9	74.4	**58.4**	62.8	81.2	55.9	61.5	52.8
	ϕ_{uniform}	66%	**28.0**	20.6	**28.6**	**29.7**	**33.3**	59.0	41.4	61.8	75.9	57.1	64.6	85.2	58.1	61.9	53.0
	Ensemble																
	$\phi_{\text{-}R}+\phi_{\text{uniform}}$ (Ours)	61%	27.2	**29.1**	27.2	24.7	27.7	**61.2**	**46.6**	**62.3**	**77.4**	**58.4**	**65.4**	**86.0**	**58.9**	62.7	**54.0**
$K=8$	*Attention-based*																
	ϕ_{original}	56%	23.3	19.4	23.5	24.0	26.3	59.8	45.0	60.3	76.1	57.8	64.6	85.4	57.5	62.6	53.0
	$\phi_{\text{-}R}$ (Ours)	56%	27.3	27.1	26.7	26.4	29.2	61.4	49.0	61.5	77.4	57.8	65.5	86.7	58.6	63.0	53.7
	Δ		+4.0	+7.6	+3.2	+2.5	+2.9	+1.6	+4.0	+1.2	+1.3	+0.0	+0.8	+1.3	+1.1	+0.4	+0.6
	Non-attention-based																
	ϕ_{KNN}	55%	23.6	15.4	24.4	25.2	29.4	58.6	40.2	61.1	74.5	58.5	62.9	81.4	56.2	60.9	53.0
	ϕ_{uniform}	55%	30.3	24.6	31.0	30.9	34.8	59.3	42.2	61.8	76.0	57.4	64.4	85.3	57.9	61.0	53.2
	Ensemble																
	$\phi_{\text{-}R}+\phi_{\text{uniform}}$ (Ours)	50%	**35.6**	**32.0**	**35.9**	**35.4**	**38.8**	**62.7**	**51.7**	**62.4**	**78.1**	**58.6**	**66.0**	**87.4**	**59.1**	**63.6**	**54.0**

- Removing RoPE from token selection criteria improves performance
- Token selection improves when pruning later
- Integrating uniform sampling with attention-based criteria is beneficial in early layers

Distilling Insights

Guided by our insights, we propose **FEATHER** (***F**ast and **E**ffective **A**cceleration wi**TH** **E**nsemble c**R**iteria*). Specifically, we prune after an early layer with our proposed RoPE-free attention + uniform criteria. Furthermore, we apply more extensive pruning at a later layer using attention-based criteria, when the effectiveness of the attention-based criteria has improved.

Over 5x improvement over FastV