SCOPE: Saliency-Coverage Oriented Token Pruning for Efficient Multimodel LLMs

Jinhong Deng[1,3,4], Wen Li[2,*], Joey Tianyi Zhou[3,4], Yang He[3,4]
[1]University of Electronic Science and Technology of China (UESTC)
[2]Shenzhen Institute for Advanced Study, UESTC
[3]CFAR, Agency for Science, Technology and Research (A*STAR), Singapore
[4]IHPC, Agency for Science, Technology and Research (A*STAR), Singapore

NEURAL INFORMATION PROCESSING SYSTEMS 2025

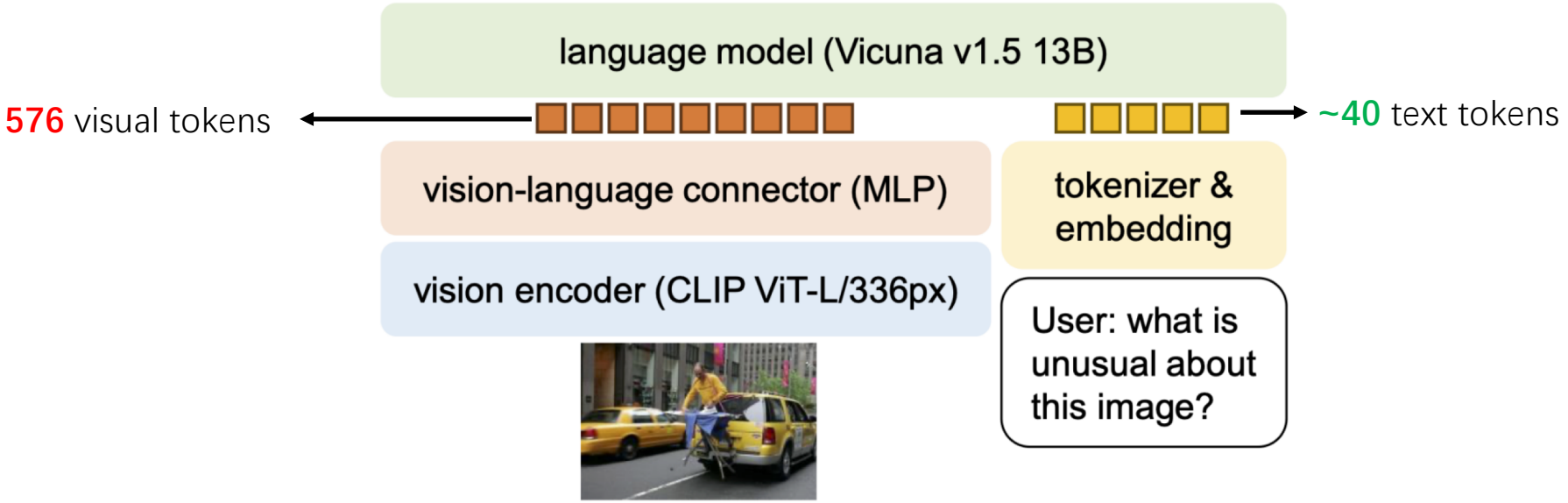
Centre for Frontier AI Research
CFAR

Institute of High Performance Computing
IHPC

Jinhong Deng

CREATING GROWTH, ENHANCING LIVES

Vision Language Model

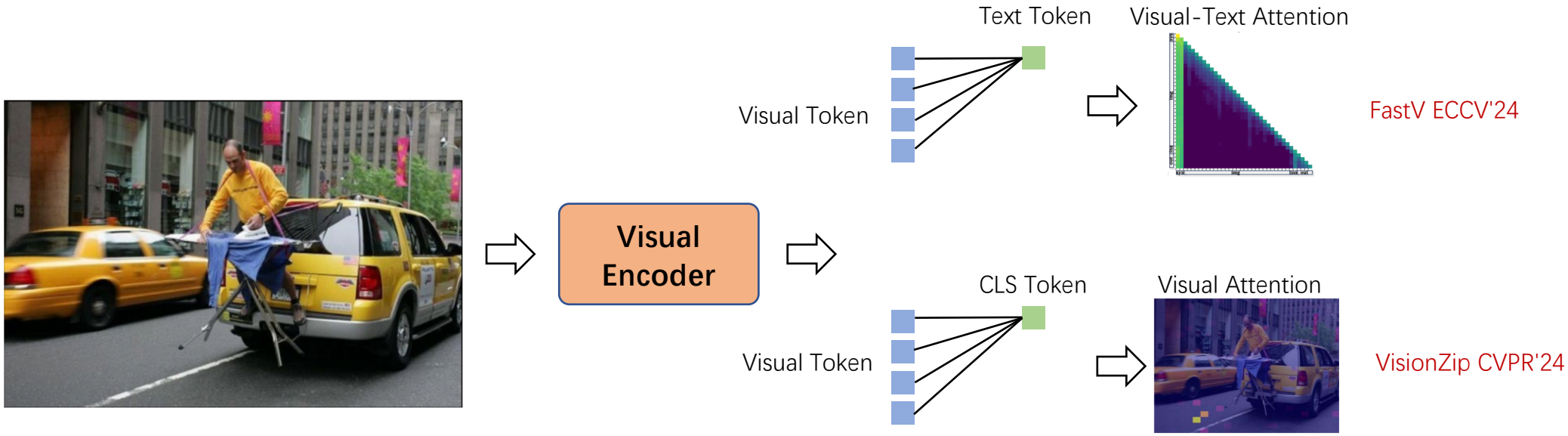


Are all visual tokens necessary?

Improved Baselines with Visual Instruction Tuning, CVPR 2024. (LLaVA-1.5)

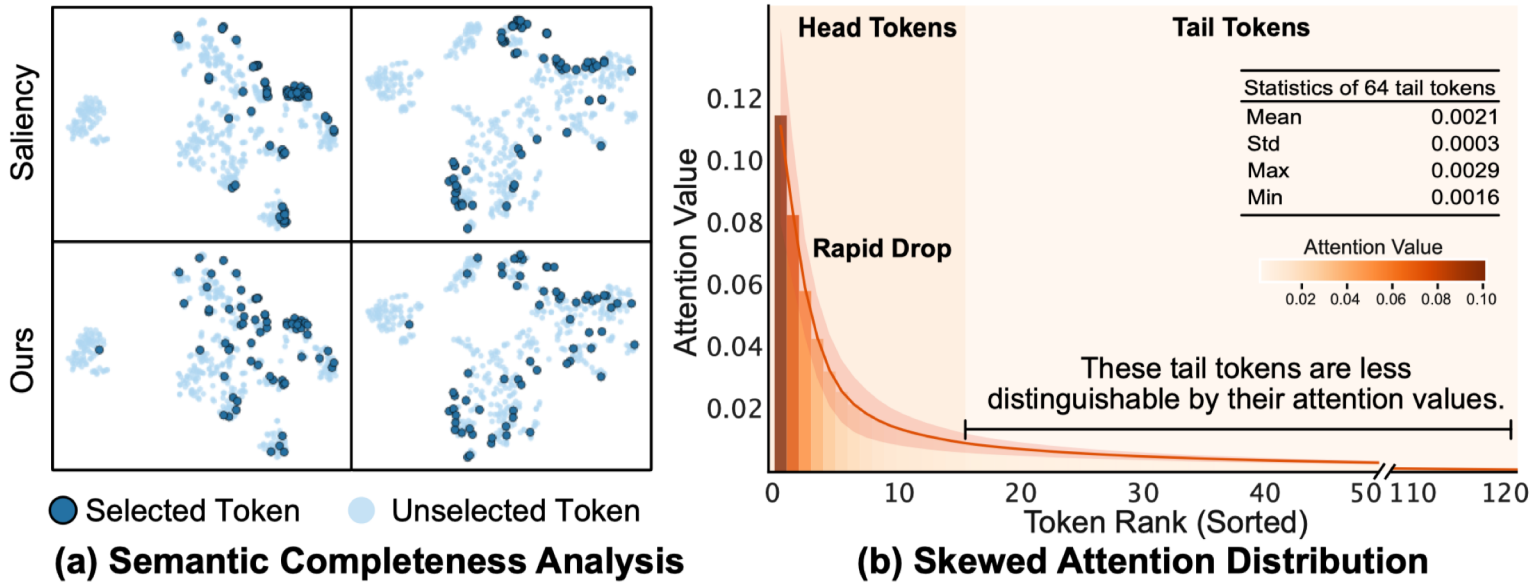
CREATING GROWTH, ENHANCING LIVES

Visual Token Compression



Tokens with high **saliency scores** (text/vision attention) are retained

Limitations of Saliency-Based Methods

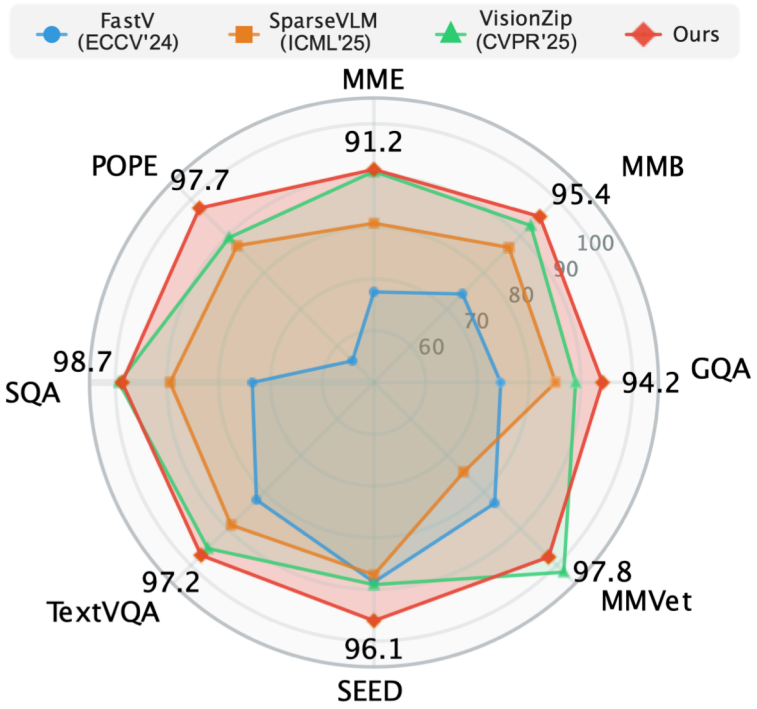


(a) Semantic Completeness Analysis

(b) Skewed Attention Distribution

CREATING GROWTH, ENHANCING LIVES

Our Results



- Better Performance at same token budget.
- Token reduction while maintain performance.
- Generalizability on LLaVA-1.5, LLaVA-NeXT and Video-LLaVA.

CREATING GROWTH, ENHANCING LIVES

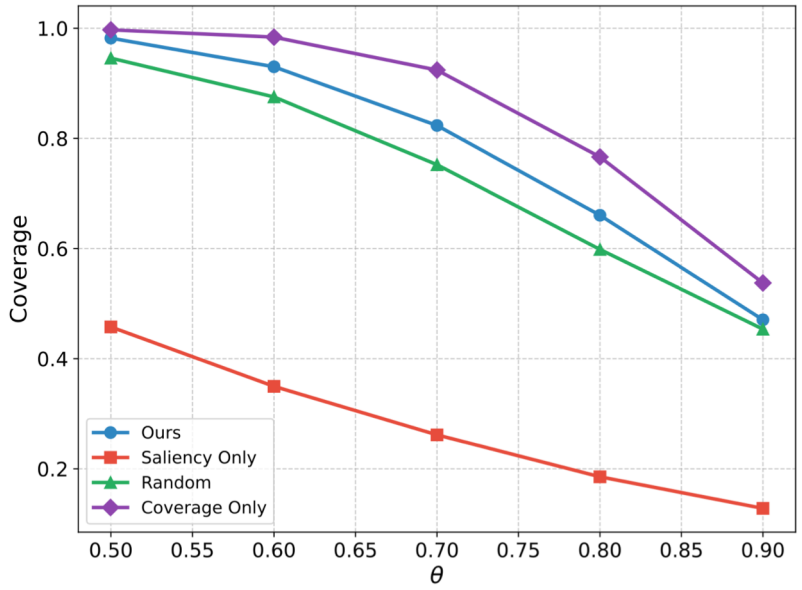
Coverage Analysis

θ-Coverage Definition

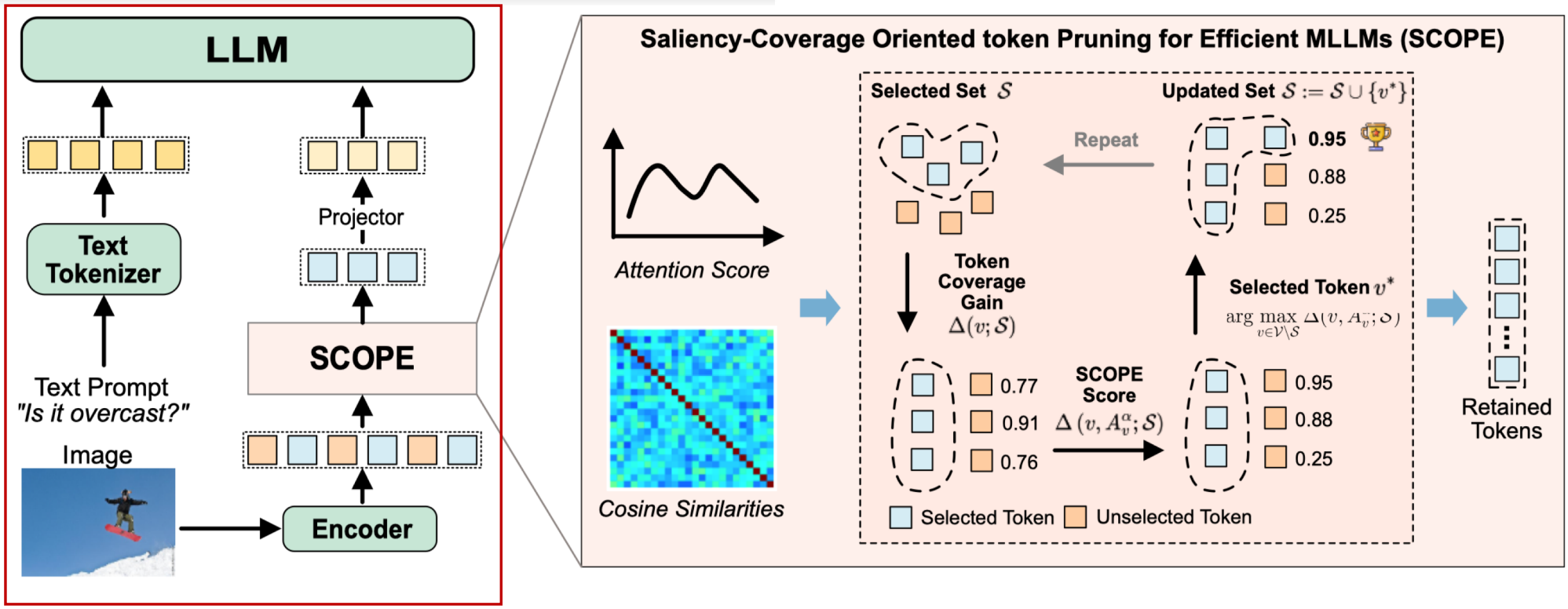
$$\text{sim}(v, v') := \frac{v^\top v'}{\|v\| \cdot \|v'\|} \geq \theta$$

$$\text{Coverage}_\theta(\mathcal{V}', \mathcal{V}) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{I}\left(\max_{v' \in \mathcal{V}'} \text{sim}(v, v') \geq \theta \right)$$

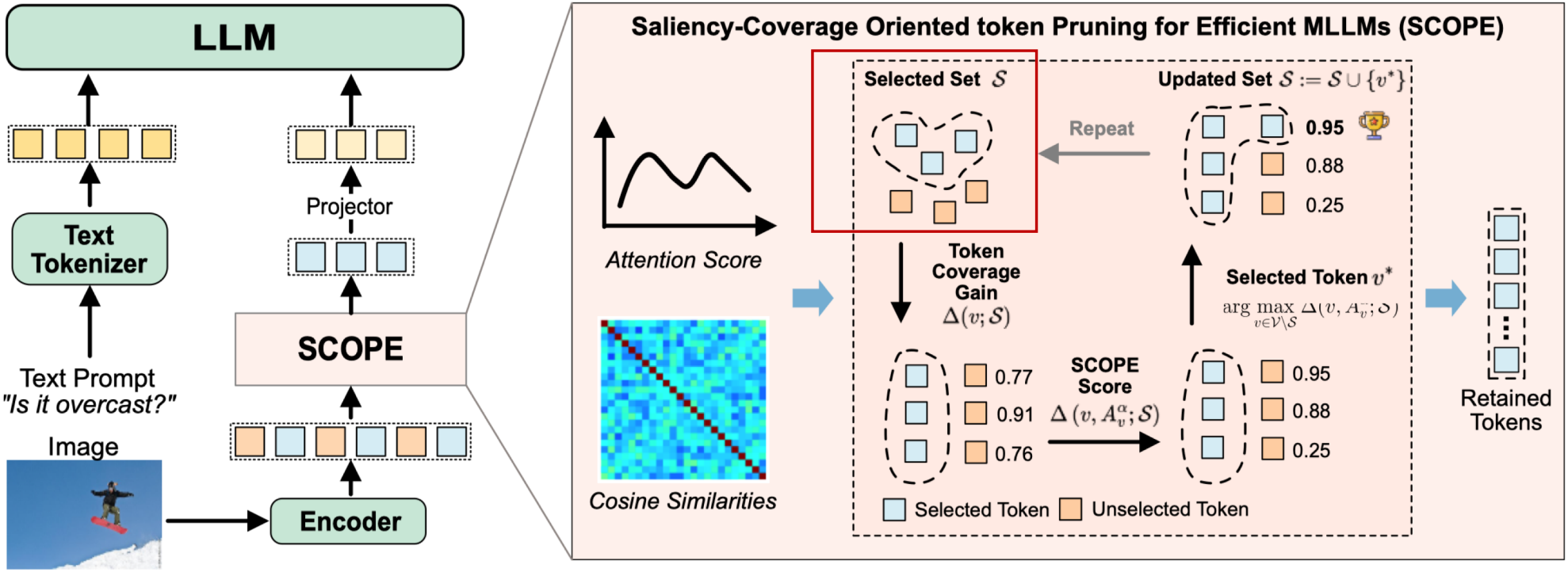
Saliency Only Exhibits Low Coverage



Saliency-based method captures dominant information, it tends to overlook a substantial amount of semantic content.



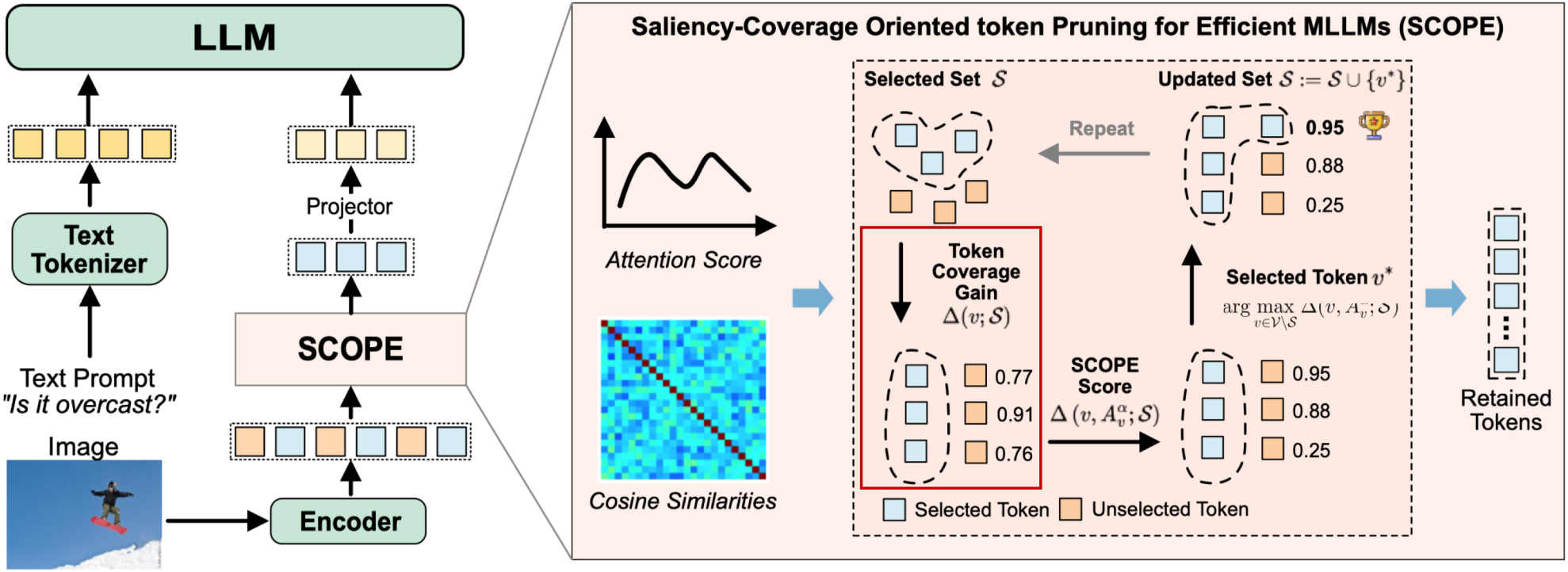
Reduce tokens before LLM



Set-coverage for selected tokens

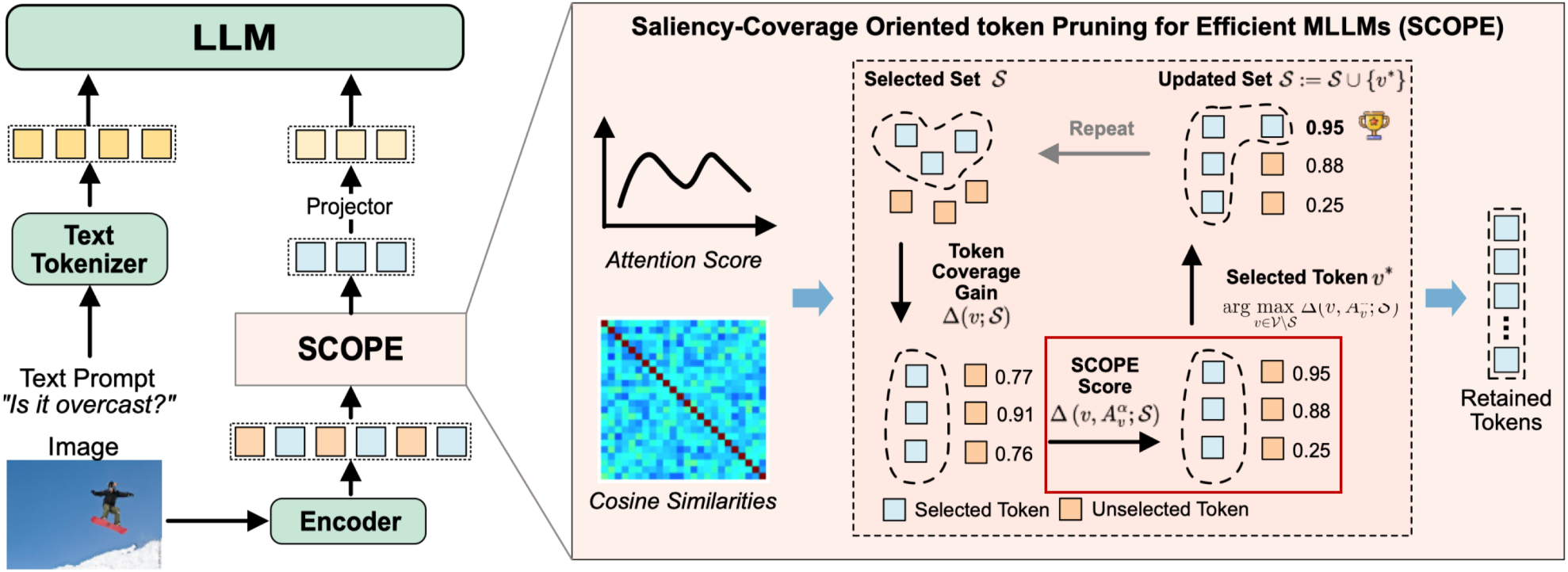
$$f(\mathcal{S}) = \sum_{u \in \mathcal{V}} C(u, \mathcal{S}) = \sum_{u \in \mathcal{V}} \max_{s \in \mathcal{S}} \text{sim}(u, s)$$

CREATING GROWTH, ENHANCING LIVES



Token-coverage Gain $\Delta(v; \mathcal{S}) = f(\mathcal{S} \cup \{v\}) - f(\mathcal{S})$

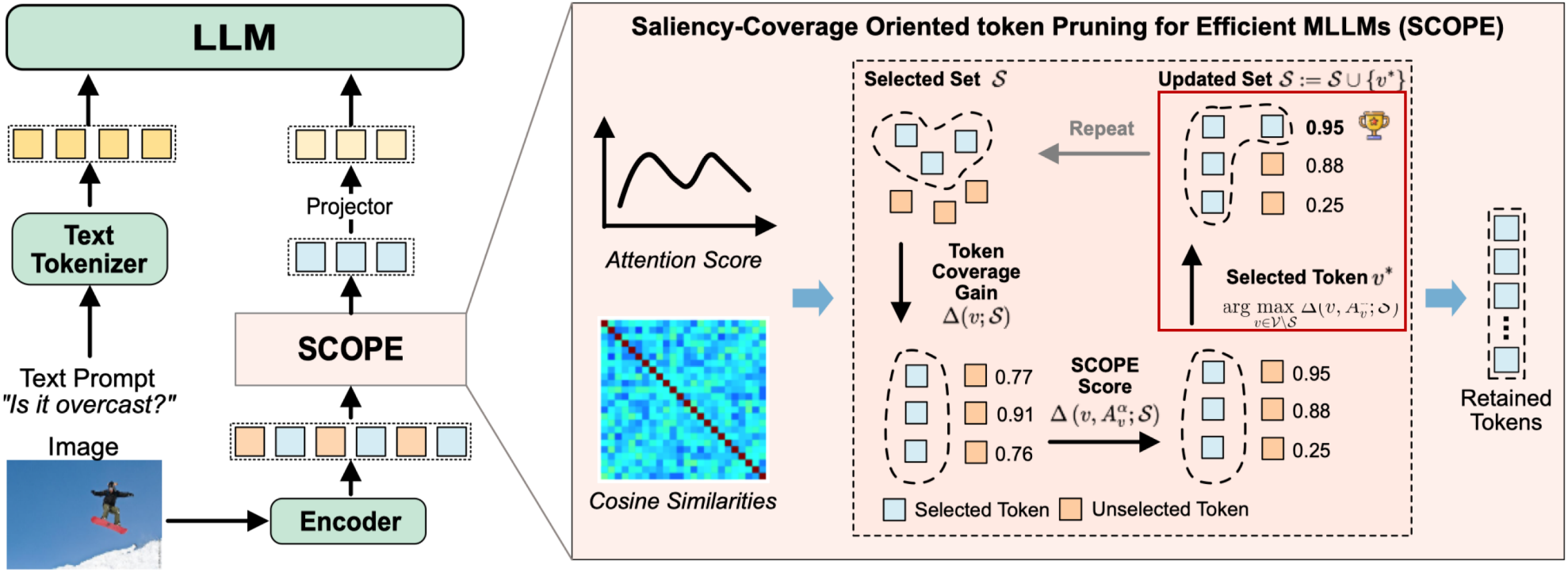
how much additional coverage is achieved by selecting token v.



Coverage

SCOPE score $\Delta(v, A_v^{\alpha}; \mathcal{S}) = \boxed{\Delta(v; \mathcal{S})} \cdot \boxed{A_v^{\alpha}}$ Saliency

Integration of saliency and coverage gain.



Selected token $v^* \in \arg\max_{v \in \mathcal{V} \setminus \mathcal{S}} \boxed{\Delta(v, A_v^{\alpha}; \mathcal{S})}$ SCOPE score

CREATING GROWTH, ENHANCING LIVES

Results on **LLaVA-1.5 7B**

Method	GQA	MMB	MME	POPE	SQA	TextVQA	SEED	MMVet	Avg.
Upper Bound, 576 Tokens (100%)									
Vanilla (CVPR'24)	61.9 100%	64.7 100%	1862 100%	85.9 100%	69.5 100%	58.2 100%	58.6 100%	31.1 100%	100%
Retain 192 Tokens (↓ **66.7%**)									
FastV (ECCV'24)	52.7 85.1%	61.2 94.6%	1612 86.6%	64.8 75.4%	67.3 96.8%	52.5 90.2%	57.1 97.4%	27.7 89.7%	89.5%
SparseVLM (ICML'25)	57.6 93.1%	62.5 96.6%	1721 92.4%	83.6 97.3%	69.1 99.4%	56.1 96.4%	55.8 95.2%	31.5 101.3%	96.5%
VisionZip (CVPR'25)	59.3 95.8%	63.0 97.4%	1783 95.7%	85.3 99.3%	68.9 99.1%	57.3 98.5%	56.4 96.2%	31.7 101.9%	98.0%
PDrop (CVPR'25)†	57.1 92.2%	63.2 97.7%	1766 94.8%	82.3 95.8%	70.2 101.0%	56.1 96.4%	54.7 93.3%	30.5 98.1%	96.2%
Ours	60.1 97.1%	63.6 98.3%	1804 96.9%	86.4 100.6%	68.8 99.0%	57.7 99.1%	58.7 100.2%	32.5 104.5%	99.5% (↓ **0.5%**)
Retain 128 Tokens (↓ **77.8%**)									
FastV (ECCV'24)	49.6 80.1%	56.1 86.7%	1490 80.0%	59.6 69.4%	60.2 86.6%	50.6 86.9%	55.9 95.4%	28.1 90.4%	84.4%
SparseVLM (ICML'25)	56.0 90.5%	60.0 92.7%	1696 91.1%	80.5 93.7%	67.1 96.5%	54.9 94.3%	53.4 91.1%	30.0 96.5%	93.3%
VisionZip (CVPR'25)	57.6 93.1%	62.0 95.8%	1761.7 94.6%	83.2 96.9%	68.9 99.1%	56.8 97.6%	54.9 93.7%	32.6 104.8%	96.9%
PDrop (CVPR'25)†	56 90.5%	61.1 94.4%	1664 89.4%	82.3 95.8%	69.9 100.6%	55.1 94.7%	53.3 91.0%	30.8 99.0%	94.4%
Ours	59.7 96.4%	62.5 96.6%	1776 95.4%	86.1 100.2%	68.4 98.4%	57.2 98.3%	57.8 98.6%	31.4 101.0%	98.1% (↓ **1.9%**)
Retain 64 Tokens (↓ **88.9%**)									
FastV (ECCV'24)	46.1 74.5%	48.0 74.2%	1256 67.5%	48 55.9%	51.1 73.5%	47.8 82.1%	51.9 88.6%	25.8 83.0%	74.9%
SparseVLM (ICML'25)	52.7 85.1%	56.2 86.9%	1505 80.8%	75.1 87.4%	62.2 89.5%	51.8 89.0%	51.1 87.2%	23.3 74.9%	85.1%
VisionZip (CVPR'25)	55.1 89.0%	60.1 92.9%	1690 90.8%	77.0 89.6%	69.0 99.3%	55.5 95.4%	52.2 89.1%	31.7 101.9%	93.5%
PDrop (CVPR'25)†	41.9 67.7%	33.3 51.5%	1092 58.6%	55.9 65.1%	69.2 99.6%	45.9 78.9%	40.0 68.3%	30.7 98.7%	73.5%
Ours	58.3 94.2%	61.7 95.4%	1698 91.2%	83.9 97.7%	68.6 98.7%	56.6 97.3%	56.3 96.1%	30.4 97.7%	96.0% (↓ **4.0%**)

Results on **LLaVA-Next 7B**

Method	GQA	MMB	MME	SQA	TextVQA	MMMU	Avg.
Upper Bound, 2880 Tokens (100%)							
Vanilla (CVPR'24)	64.2 100%	67.9 100%	1842 100%	70.2 100%	61.3 100%	35.1 100%	100%
Retain 640 Tokens (↓ **77.8%**)							
SparseVLM (ICML'25)	60.3 93.9%	65.7 96.8%	1772 96.2%	67.7 96.4%	57.8 94.3%	34.6 98.6%	96.0%
VisionZip (CVPR'25)	61.3 95.5%	66.3 97.6%	1787 97.0%	68.1 97.0%	60.2 98.2%	34.7 98.9%	97.4%
Ours	61.9 96.4%	66.2 97.5%	1842 100.0%	67.8 96.6%	60.1 98.0%	36.9 105.1%	98.9% (↓ **1.1%**)
Retain 320 Tokens (↓ **88.9%**)							
SparseVLM (ICML'25)	57.7 89.9%	64.3 94.7%	1694 92.0%	67.3 95.9%	55.9 91.2%	34.4 98.0%	93.6%
VisionZip (CVPR'25)	59.3 92.4%	63.1 92.9%	1702 92.4%	67.3 95.9%	58.9 96.1%	35.3 100.6%	95.0%
Ours	61.0 95.0%	65.9 97.1%	1789 97.1%	67.7 96.4%	58.4 95.3%	35.6 101.4%	97.1% (↓ **2.9%**)
Retain 160 Tokens (↓ **94.4%**)							
SparseVLM (ICML'25)	51.2 79.8%	63.1 92.9%	1542 83.7%	67.5 96.2%	46.4 75.7%	32.8 93.4%	86.9%
VisionZip (CVPR'25)	55.5 86.4%	60.1 88.5%	1630 88.5%	68.3 97.3%	56.2 91.7%	36.1 102.8%	92.5%
Ours	60.0 93.5%	64.3 94.7%	1700 92.3%	67.4 96.0%	56.8 92.7%	35.6 101.4%	95.1% (↓ **4.9%**)

CREATING GROWTH, ENHANCING LIVES

Table 3: Performance comparison on Video-LLaVA. The original Video-LLaVa's video token number is 2048, while our method only retains the 136 tokens.

Method	TGIF	MSVD	MSRVTT	ActivityNet	Avg
Video-LLaVA	47.1	69.8	56.7	43.1	100.0%
FastV	23.1 49.0%	38.0 54.4%	19.3 34.0%	30.6 71.0%	52.1%
SparseVLM	44.7 94.9%	68.2 97.7%	31.0 54.7%	42.6 98.8%	86.5%
VisionZip	42.4 90.0%	63.5 91.0%	52.1 91.9%	43.0 99.8%	93.2%
Ours	47.1 100.0%	69.2 99.1%	55.9 98.6%	44.9 104.2%	**100.5%**

Results on **Video-LLaVA**

Table 4: Ablation studies of the proposed method.

	GQA	MMB	MME	POPE	TextVQA
Random	55.5	54.0	1556	75.2	48.4
Saliency-only	55.0	60.8	1665	76.8	55.4
Coverage-only	58.1	60.8	1687	82.1	56.3
Ours	**58.3**	**61.7**	**1698**	**83.9**	**56.6**

Ablation Studies

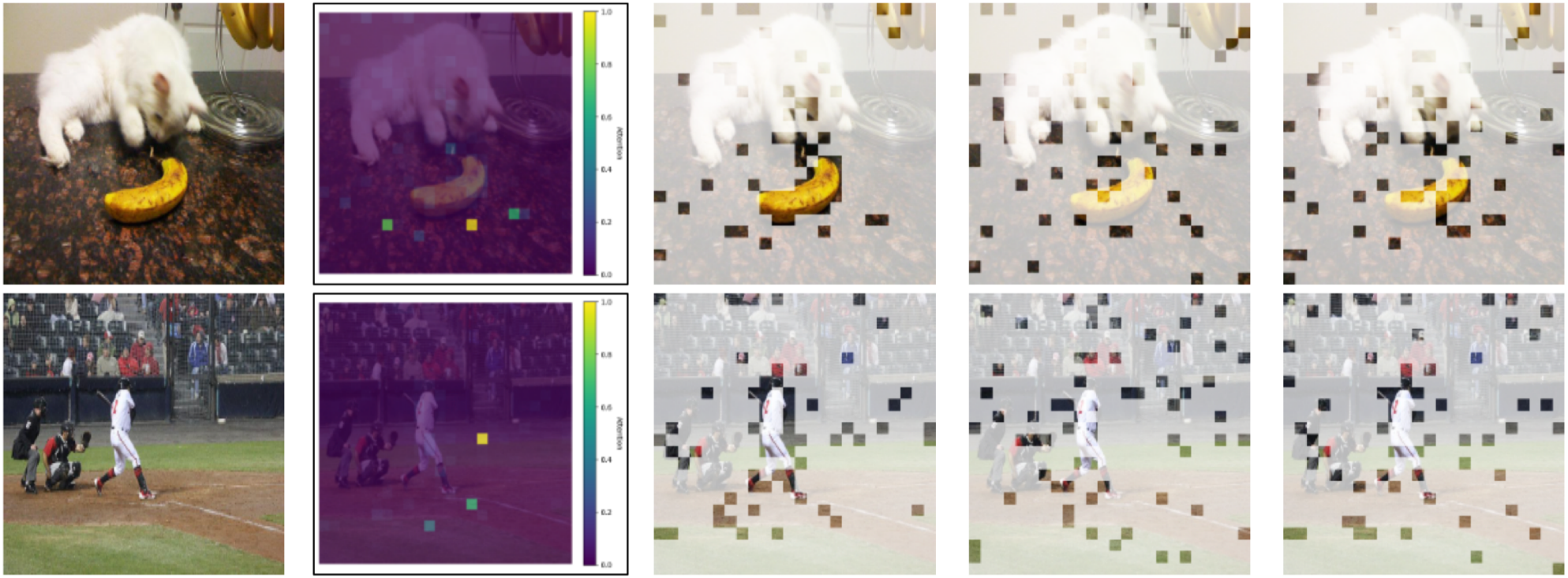
Table 5: Efficiency analysis of our method on LLaVA-NeXT 7B. The experiments are conducted on a system equipped with 4×A100. Δ denotes the reduction ratio.

	Token Number	POPE	Latency (s)	Δ
Vanilla	2880	86.4	601.9	-
PDrop	160	53.2	184.0	3.3×
Ours	160	81.3	188.8	3.2×

Efficiency Analysis

CREATING GROWTH, ENHANCING LIVES

Visualization of token pruning among different pruning strategies



CREATING GROWTH, ENHANCING LIVES

CREATING GROWTH, ENHANCING LIVES

Centre for Frontier AI Research
CFAR

Institute of High Performance Computing
IHPC

code

arXiv Paper

THANK YOU

www.a-star.edu.sg