Don't Just Chase "Highlighted Tokens" in MLLMs: Revisiting Visual Holistic Context Retention

Xin Zou[1,2], Di Lu[1], Yizhou Wang[1], Yibo Yan[1,2], Yuanhuiyi Lyu[1,2], Xu Zheng[1,3], Linfeng Zhang[4], Xuming Hu[1,2]

[1] HKUST (GZ) [2] HKUST [3] INSAIT, Sofia University "St. Kliment Ohridski" [4] Shanghai Jiao Tong University

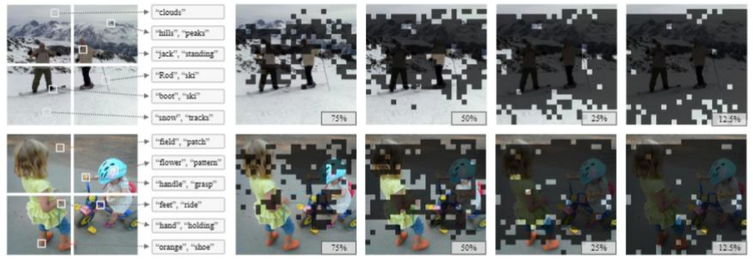
香港科技大學(廣州) THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU) · 香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY · INSAIT Institute for Computer Science, Artificial Intelligence and Technology · 上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

Scan for details !

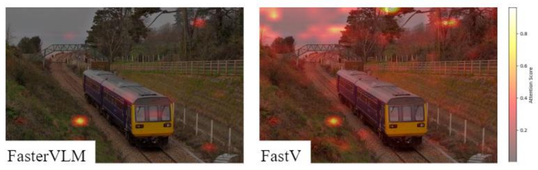
Motivation

The semantic relationships between those tokens from different regions facilitate the overall understanding, e.g., "snow", "ski", "hills" are kind of **self-explanatory**.

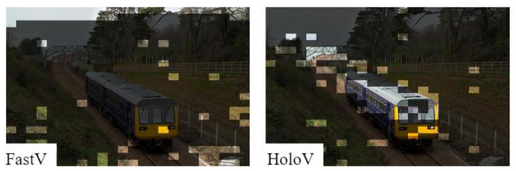


We need a holistic context for understanding!

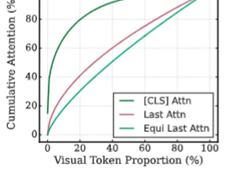
Preliminary



Distribution map of visual token attention

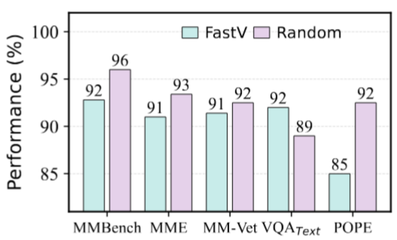


Positional Bias *(last token with higher attention)*

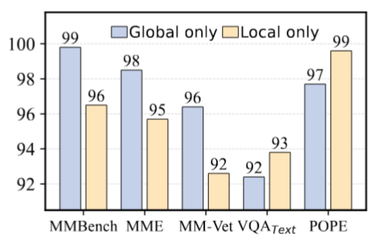


Attention Dispersion

Holistic Context Trumps Local Duplicates



FastV *v.s.* Random strategy.



Thumbnail *v.s.* Local crops

Findings: With only the global thumbnail yields strong results on general visual perception benchmarks, *e.g.,* MME, MMBench, MM-Vet, highlighting the inherent role of holistic context in guiding general visual understanding. On the contrary, using only local crops leads to poor performance in these general perception tasks, *e.g.,* VQA_{Text} and POPE, but excels in fine-grained perception benchmarks.

HoloV Framework

1.Token importance scores:

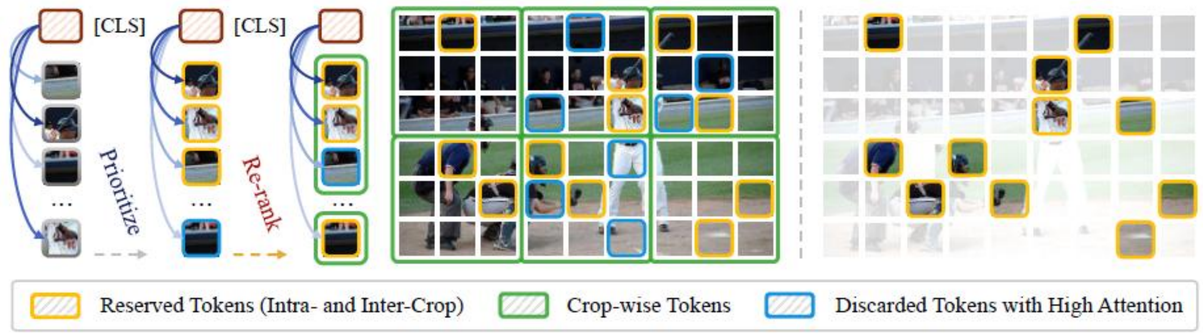
$$\mathcal{H}^c = \gamma_c \mathcal{V}^c + \mathcal{A}^c, \text{ where } \gamma_c = \mathbb{E}[\|\mathcal{A}^c\|]/\mathbb{E}[\|\mathcal{V}^c\|].$$

2.Crop importance weights:

$$w_c = \left(\frac{1}{M}\sum_{t=1}^{M}\mathcal{H}_t^c\right)^\tau \Big/ \sum_{c'=1}^{\mathcal{C}}\left(\frac{1}{M}\sum_{t=1}^{M}\mathcal{H}_t^{c'}\right)^\tau,$$

3.Top-k visual token selection:

$$\text{argmax}_{\Omega_c \subset \{1,\dots,M\}} \sum \mathcal{H}^c, \text{ subject to } |\Omega_c| = q_c$$



HoloV. ***We just re-rank highlighted visual tokens for holistic context retention!***

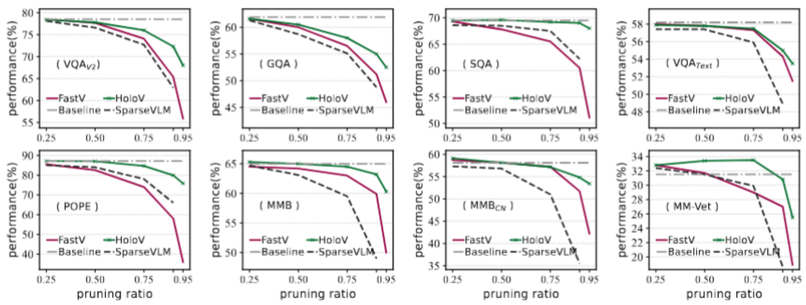
Fast Visual Context Refetching *(A supplementary strategy for holistic context retention)*

$$\text{FFN}^{(l)}(\boldsymbol{x} \propto \boldsymbol{z}_v) = \alpha\underline{\Delta} + (1-\alpha)\,\text{FFN}^{(l)}(\boldsymbol{x}), \quad \underline{\Delta}(\boldsymbol{z}_v \mid \boldsymbol{x}) = \sum_{i=1}^{N_v} \phi(\langle \boldsymbol{x}, \boldsymbol{z}_{v,i}\rangle) \cdot \boldsymbol{z}_{v,i}.$$

Main Results & Efficiency Analysis & Visualization

Table 1: Performance comparison of various methods across different benchmarks. Results are shown for different pruning ratios, with accuracy and average performance highlighted. Best results in **blue**.

Methods	GQA	MMB	MMB_{CN}	MME	POPE	SQA	VQA_{V2}	VQA_{Text}	VizWiz	Average
Upper Bound, 576 Tokens	61.9	64.7	58.1	1862	85.9	69.5	78.4	58.2	50.0	100%
LLaVA-1.5 7B	*Retain 192 Tokens* (↓ 66.7%)									
ToMe (ICLR23)	54.3	60.5	-	1563	72.4	65.2	68.0	52.1	-	88.5%
FastV (ECCV24)	52.7	61.2	57.0	1612	64.8	67.3	67.1	52.5	50.8	90.5%
MustDrop (2024.11)	58.2	62.3	55.8	1787	82.6	69.2	76.0	56.5	51.4	97.2%
LLaVA-PruMerge (ICCV25)	54.3	59.6	52.9	1632	71.3	67.9	70.6	54.3	50.1	91.4%
PDrop (CVPR25)	57.1	63.2	56.8	1766	82.3	68.8	75.1	56.1	51.1	96.7%
FiCoCo-V (2025.03)	58.5	62.3	55.3	1732	82.5	67.8	74.4	55.7	51.0	96.1%
HiRED (AAAI25)	58.7	62.8	54.7	1737	82.8	68.4	74.9	47.4	50.1	94.6%
VisionZip (CVPR25)	**59.3**	64.5	57.3	1767	86.4	68.9	**76.8**	57.3	**51.6**	98.1%
SparseVLM (ICML25)	57.6	62.5	53.7	1721	83.6	69.1	75.6	56.1	50.5	96.1%
DART (EMNLP25)	58.9	63.6	57.0	**1856**	82.8	69.8	76.7	57.4	51.1	98.5%
HoloV (Ours)	59.0	**65.4**	**58.0**	1820	**85.6**	**69.8**	76.7	**57.4**	50.9	**99.2%**
LLaVA-1.5 7B	*Retain 128 Tokens* (↓ 77.8%)									
ToMe (ICLR23)	52.4	53.3	-	1343	62.8	59.6	63.0	49.1	-	80.4%
FastV (ECCV24)	49.6	56.1	56.4	1490	59.6	60.2	61.8	50.6	51.3	85.4%
MustDrop (2024.11)	56.9	61.1	55.2	1745	78.7	68.5	74.6	56.3	**52.1**	95.7%
LLaVA-PruMerge (ICCV25)	53.3	58.1	51.7	1554	67.2	67.1	68.8	54.3	50.3	89.4%
PDrop (CVPR25)	56.0	61.1	56.6	1644	82.3	68.3	72.9	55.1	51.0	94.9%
FiCoCo-V (2025.03)	57.6	61.1	54.3	1711	82.2	68.3	73.1	55.6	49.4	94.9%
HiRED (AAAI25)	57.2	61.5	53.6	1710	79.8	68.1	73.4	46.1	51.3	93.1%
VisionZip (CVPR25)	57.6	63.4	56.7	1768	84.7	68.8	75.6	56.8	52.0	97.2%
SparseVLM (ICML25)	56.0	60.0	51.1	1696	80.5	67.1	73.8	54.9	51.4	93.8%
DART (EMNLP25)	**57.9**	63.2	**57.0**	**1845**	80.1	69.1	**75.9**	56.4	51.7	97.5%
HoloV (Ours)	57.7	**63.9**	56.5	1802	**84.0**	**69.8**	75.5	**56.8**	51.5	**98.0%**
LLaVA-1.5 7B	*Retain 64 Tokens* (↓ 88.9%)									
ToMe (ICLR23)	48.6	43.7	-	1138	52.5	50.0	57.1	45.3	-	70.1%
FastV (ECCV24)	46.1	48.0	52.7	1256	48.0	51.1	55.0	47.8	50.8	76.7%
MustDrop (2024.11)	53.1	60.0	53.1	1612	68.0	63.4	69.3	54.2	51.2	90.1%
LLaVA-PruMerge (ICCV25)	51.9	55.3	49.1	1549	65.3	68.1	67.4	54.0	50.1	87.7%
PDrop (CVPR25)	41.9	33.3	50.5	1092	55.9	68.6	69.2	45.9	50.7	77.5%
FiCoCo-V (2025.03)	52.4	60.3	53.0	1591	76.0	68.1	71.3	53.6	49.8	91.5%
HiRED (AAAI25)	54.6	60.2	51.4	1599	73.6	68.2	69.7	44.2	50.2	89.4%
VisionZip (CVPR25)	55.1	60.1	**55.4**	1690	77.0	69.0	72.4	**55.5**	**52.9**	94.5%
SparseVLM (ICML25)	52.7	56.2	46.1	1505	75.1	62.2	68.2	51.8	50.1	87.3%
DART (EMNLP25)	**55.9**	60.6	53.2	**1765**	73.9	**69.8**	72.4	54.4	51.6	93.9%
HoloV (Ours)	55.3	**63.3**	55.1	1715	**80.3**	69.5	**72.8**	55.4	52.8	**95.8%**



Methods	Time	Prefill	Latency	Mem.	Acc.	Time	Prefill	Latency	Mem.	Acc.
Upper Bound, 576 Tokens	49:41	0.5ms	0.334s	19.0G	100.%	49:41	0.5ms	0.334s	19.0G	100.%
LLaVA-1.5-7B	*Retain 192 Tokens* (↓ 66.7%)					*Retain 58 Tokens* (↓ 90%)				
FastV (ECCV24)	35:34	0.5ms	0.239s	16.0G	75.4%	30:41	0.5ms	0.206s	15.6G	66.8%
MustDrop (2024.11)	32:30	0.5ms	0.273s	15.6G	96.2%	29:40	0.6ms	0.199s	14.5G	87.1%
FasterVLM (ICCV25)	**30:09**	0.5ms	**0.202s**	15.6G	100.%	25:08	0.5ms	**0.168s**	14.5G	92.5%
HiRED (AAAI25)	30:08	0.6ms	0.210s	15.7G	96.4%	**25:03**	0.6ms	0.168s	14.5G	92.7%
SparseVLM (ICML25)	40:51	0.6ms	0.251s	15.8G	97.3%	31:28	0.6ms	0.212s	14.6G	92.3%
HoloV (Ours)	31:02	0.5ms	0.208s	**15.6G**	**99.7%**	27:36	0.5ms	0.176s	**14.5G**	**95.7%**