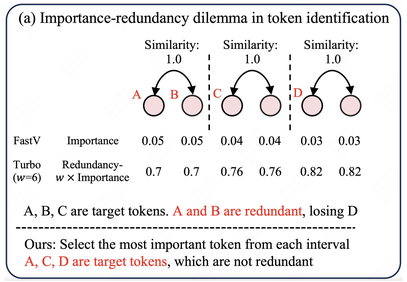
Libra-Merging: Importance-redundancy and Pruning-merging Trade-off for Acceleration Plug-in in Large Vision-Language Model

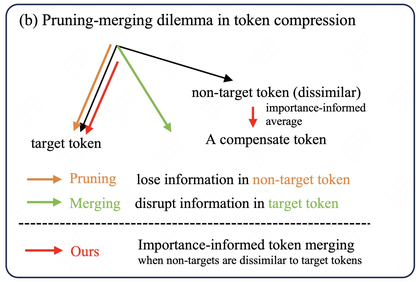
Longrong Yang[1*], Dong Shen[2*], Chaoxiang Cai[1], Kaibing Chen[2], Fan Yang[2], Tingting Gao[2], Di Zhang[2], Xi Li[1]

[1]Zhejiang University, [2]Kuaishou Technology

Motivation

- Visual token compression can be modeled as identifying important non-redundant tokens (named target tokens) and compressing the remaining tokens (named non-target tokens).
- Target token identification faces an importance-redundancy trade-off:
  - Select the most important tokens → Selected tokens may be redundant.
  - Balance importance and redundancy with a hyper-parameter → Hyper-parameter is not universal.
- Token compression faces a pruning-merging trade-off:
  - Pruning non-target tokens loses information in non-target tokens.
  - Merging non-target into target tokens may disrupt information in target tokens.





Adjacent tokens are usually redundant.

- We select non-adjacent important tokens as target tokens.

Token merging is harmful when dissimilar tokens merge.

- Tokens merge only when they are similar.

Feel free to communicate if you have any questions.

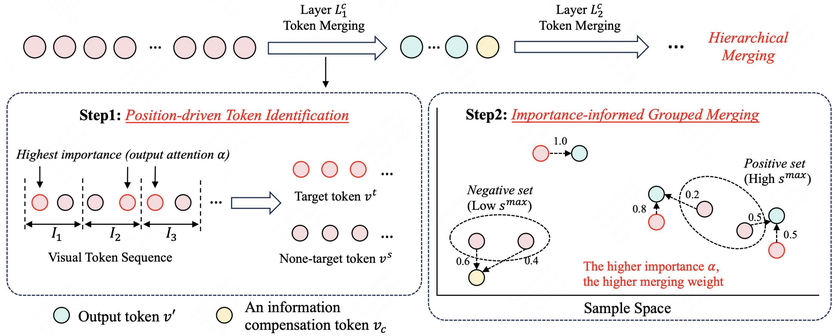
Libra-Merging

1. Position-driven token identification:

- First, we evenly divide the visual token sequence into several intervals.
  - We only select one token from each interval.
- Second, we extract the token with the highest importance from each interval as target token. Other tokens are labeled as non-target tokens.

2. Importance-informed grouped merging:

- First, we compute the max-similarity metric s^{max} of each token and group tokens in *positive set* and *negative set*.
  - s^{max}: the highest similarity between each token and target tokens.
  - when tokens satisfy $s^{max} > \tau$, they are grouped in *positive set*; the remaining tokens are grouped in *negative set*.
- Second, importance-informed grouped merging operates on two dimensions:
  - We directly merge tokens of the *positive set* into target tokens.
  - Tokens of the *negative set* are merged into an information compensation token, to further prevent losing information in non-target tokens
- **Flash-attn compatibility**: We propose the hybrid attention mechanism, with the naive attention for token importance computation.



The merging weight is generated from token importance.

Experiments

➢ LVLMs (image-text models) with token compression methods

Model	Layer	R	Nocaps	Flickr30k	GQA	POPE	MME	Avg
Qwen2-VL-7B	-	-	102.6	77.4	62.4	87.8	1683.6	82.5
+FastV	3	50%	102.8	76.7	60.5	86.8	1654.9	81.7
+Libra-Merging	3	50%	102.9	76.4	61.9	86.9	1690.3	81.9
+FastV	3	80%	98.8	69.0	55.0	81.8	1549.9	76.1
+Libra-Merging	3	80%	102.4	71.0	58.7	85.2	1650.1	79.3

Model		Flops (T)	Ratio	GQA	SQA^I	MME	MMB	MMB^{CN}	TextVQA	Avg
LLaVA-1.5-7B	*vanilla*	3.82	100%	62.0	69.5	1512.0	64.7	58.2	58.2	62.5
	FastV	2.13	56%	60.4	68.8	1511.7	64.2	58.0	57.6	61.8
	Turbo	2.13	56%	61.6	68.7	1471.7	63.7	57.5	57.4	61.8
	Libra-Merging	1.78	47%	61.3	68.9	1502.5	64.3	58.5	57.4	62.1
	Libra-Merging	1.41	37%	60.7	69.2	1480.1	63.9	58.2	57.4	61.9
LLaVA-1.5-13B	*vanilla*	7.44	100%	63.2	72.8	1531.3	68.5	63.6	61.2	65.9
	FastV	4.06	55%	62.7	73.0	1549.8	68.3	63.5	60.8	65.7
	Turbo	4.06	55%	62.8	72.7	1561.0	68.1	63.2	60.7	65.5
	Libra-Merging	3.47	47%	63.3	73.1	1531.1	68.4	63.7	61.1	65.9
	Libra-Merging	2.74	37%	62.5	72.4	1512.3	68.4	63.1	60.4	65.4
LLaVA-NeXT-8B	*vanilla*	17.17	100%	65.9	77.3	1552.1	74.4	70.4	69.8	71.6
	FastV	9.36	55%	65.5	77.2	1572.6	74.5	70.6	69.5	71.5
	Turbo	9.36	55%	64.7	77.7	1505.3	73.4	69.1	65.0	70.0
	Libra-Merging	7.86	47%	65.7	77.6	1565.8	74.7	70.8	69.7	71.7
	Libra-Merging	6.24	37%	65.6	77.2	1565.7	73.9	70.2	69.4	71.3

➢ Actual runtime latency and memory usage

	Model	R	Layer	Time (one A800)	Memory	Score	Latency/Example
GQA	LLaVA-1.5-7B	-	-	21:45	16.0G	61.95	0.104s
	+FastV	50%	3	19:36	15.6G	60.35	0.093s
	+Libra-Merging	50%	3	19:48	15.6G	61.38	0.094s
	+FastV	80%	3	17:54	15.4G	56.57	0.085s
	+Libra-Merging	80%	3	17:58	15.4G	58.81	0.086s
MME	LLaVA-1.5-7B	-	-	03:59	16.0G	1512.0	0.101s
	+FastV	50%	3	03:27	15.6G	1511.7	0.087s
	+Libra-Merging	50%	3	03:30	15.6G	1513.1	0.088s
	+FastV	80%	3	03:14	15.4G	1427.6	0.082s
	+Libra-Merging	80%	3	03:16	15.4G	1440.0	0.083s

	Model	R	Layer	Time (one A800)	Memory	Score	Latency/Example
MME	Qwen2-VL-7B	-	-	13:35	27.7G	1693.6	0.343s
	w flash-attention 2	-	-	08:48	18.5G	1683.6	0.222s
	+FastV	50%	3	12:25	28.6G	1673.9	0.314s
	+FastV *w* hybrid	50%	3	08:12	17.6G	1654.9	0.207s
	+Libra-Merging *w* hybrid	50%	3	08:19	17.6G	1690.3	0.210s
MME	LLaVA-1.5-7B	-	-	03:59	16.0G	1512.0	0.101s
	w flash-attention 2	-	-	03:58	16.0G	1507.5	0.100s
	+FastV	50%	3	03:27	15.6G	1511.7	0.087s
	+FastV *w* hybrid	50%	3	03:25	15.5G	1495.5	0.086s
	+Libra-Merging *w* hybrid	50%	3	03:26	15.5G	1502.1	0.087s