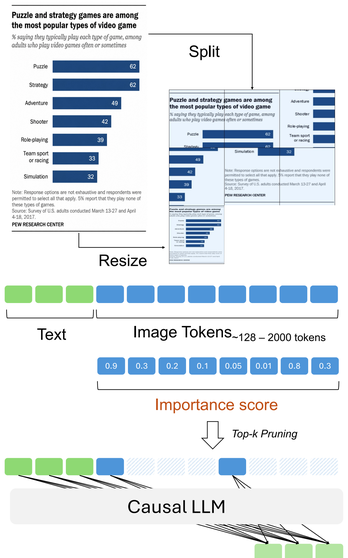
IF-Prune: Information-Flow Guided Token Pruning for Efficient Vision-Language Models

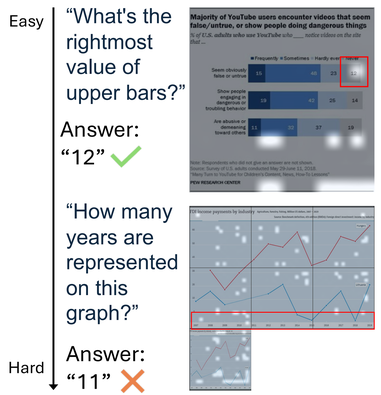
Guohao Sun[1,2]; Yufei Wang[1]; Sizhuo Ma[1]; Yuege Xie[1]; Yuting Cheng[1]; Zhiqiang Tao[2]; Jian Wang[1]
[1]Snap Inc., [2]Rochester Institute of Technology

CVPR JUNE 3-7, 2026 DENVER COLORADO

How To Provide An Optimal Pruning Guidance?



Attention-based importance score fails.



Summary: Attention scores are mostly answer driven and rely heavily on model's prior knowledge, leading to unstable and unreliable pruning guidance.

We use information flow to estimate how much each visual token contributes to the model's answer.

- Acquire pruning guidance in **one** forward pass.
- Supports **Flash-Attention**.

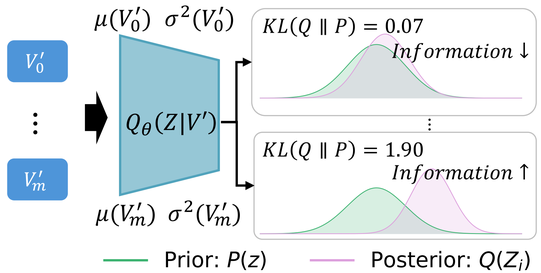
GitHub

LinkedIn

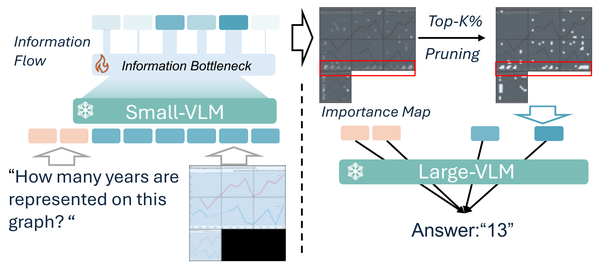
Why We Use Information Theory In Pruning?

- **Token importance beyond attention magnitude:** A visual token should be preserved if it carries information useful for predicting the answer, not merely because it receives high attention.
- **Pruning as compression:** Information theory provides a principled way to decide which tokens can be removed with minimal loss.
- **Probabilistic modeling:** IF-Prune estimates how much each visual token contributes to the model's output distribution.

Quantify Visual Token Information in VLM



Inference Pipeline



Experiments

We use InternVL2.5-1B to provide pruning guidance.

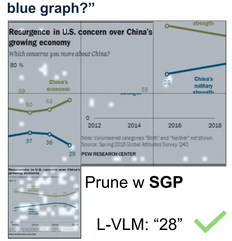
- Improve interpretability.
- One can serve many.

Method	K	L	TextVQA val	ChartQA All	GQA test-dev	MMStar test	MMBench en-dev	MM-Vet test	MME test	RealWorldQA test	Score ratio ↑
InternVL2-26B	100%	-	82.45	84.92	64.89	60.08	83.46	64.00	2270	67.58	100.00%
ToMe	5%	2	51.69	28.60	57.52	-	73.09	37.70	1933	-	82.33%
FastV†	5%	2	43.84	26.10	44.90	32.65	62.33	31.60	1799	44.05	75.05%
SGP†	5%	2	78.70	71.08	62.04	50.92	73.71	49.82	2007	64.84	88.50%
IF-Prune (ours)	5%	2	**79.24**	**71.12**	**63.52**	**53.10**	**77.58**	**50.83**	**2189**	**65.62**	**95.41%**
FastV†	5%	0	20.06	24.64	43.41	32.65	36.94	21.74	1418	44.05	59.10%
SGP†	5%	0	78.77	70.68	62.08	50.62	73.28	50.23	2028	65.10	89.25%
IF-Prune (ours)	5%	0	**79.04**	**70.96**	**63.53**	**52.49**	**77.23**	**51.42**	**2190**	**66.01**	**95.44%**

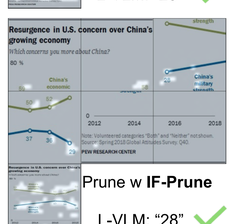
Method	K	Prefill (ms) ↓	Decode (ms) ↓	Throughput (token/s) ↑
InternVL2-8B	100%	229.1	55.8	16.9
SGP	5%	524.5	51.3	16.4
IF-Prune	5%	238.5	47.6	19.5

Simple Question → *Complex Question*

1. "What's the least value of blue graph?"

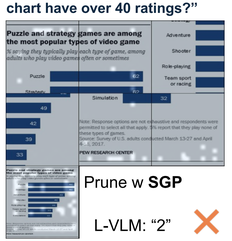


Prune w **SGP** L-VLM: "28" ✓

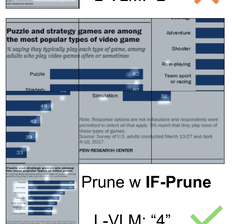


Prune w **IF-Prune** L-VLM: "28" ✓

2. "How many games in the chart have over 40 ratings?"

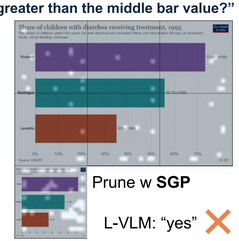


Prune w **SGP** L-VLM: "2" ✗

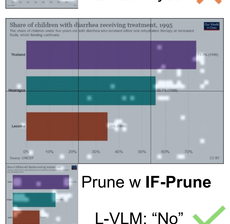


Prune w **IF-Prune** L-VLM: "4" ✓

3. "Is the average of two extreme values greater than the middle bar value?"



Prune w **SGP** L-VLM: "yes" ✗



Prune w **IF-Prune** L-VLM: "No" ✓