

ZOO-Prune: Training-Free Token Pruning via Zeroth-Order Gradient Estimation in Vision-Language Models

Youngeun Kim^{1*} Youjia Zhang^{2*} Hailing Liu² Aecheon Jung² Sunwoo Lee³ Sungeun Hong²
^{*}Equal Contribution ¹Amazon ²Sungkyunkwan University ³Inha University

Introduction

Visual Token Pruning:

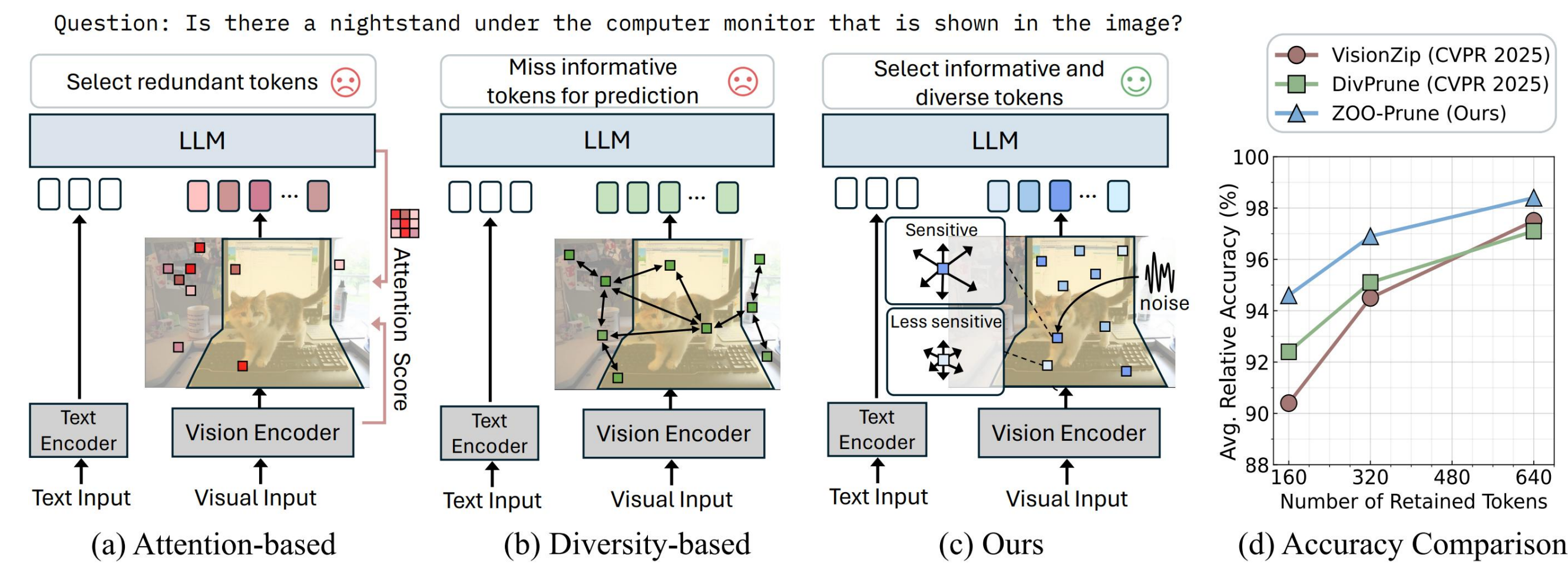
- VLMs suffer from visual token redundancy, leading to high inference cost
- Token pruning accelerates inference by discarding less informative tokens with only modest drops in accuracy

Existing pruning relies on:

- Attention scores, which can be unstable and biased
- Diversity heuristics, which ignore task relevance

! They fail to capture how tokens affect model outputs.

🔍 This motivates a training-free, output-aware pruning strategy.



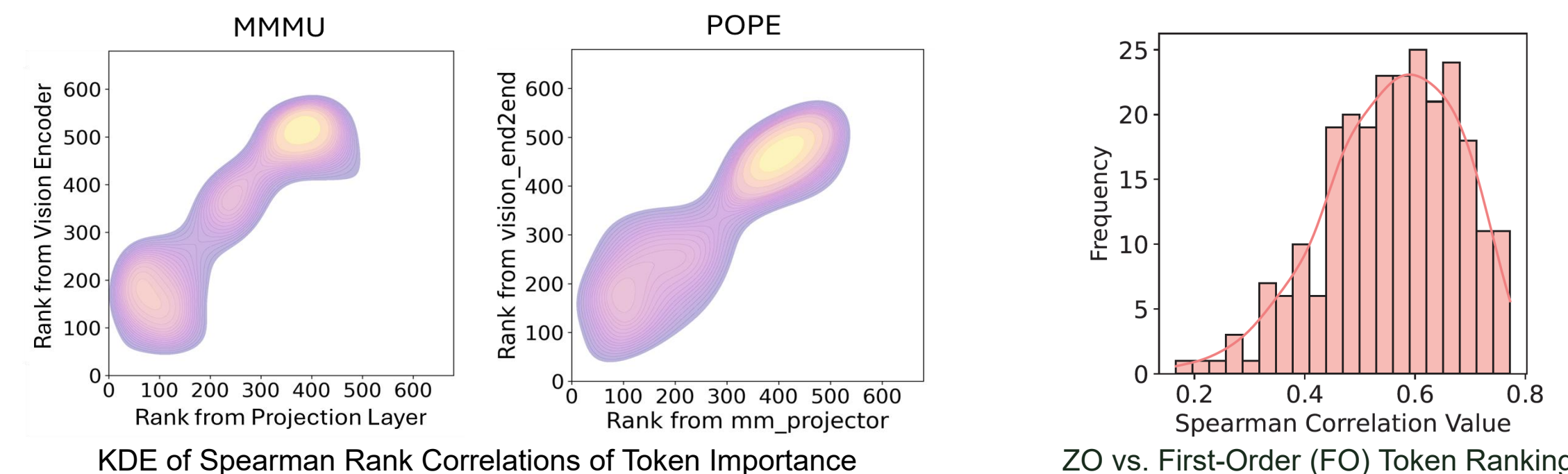
Motivation

Token Sensitivity: directly quantify output changes under token perturbations

Zeroth-Order Estimation: use only forward passes, avoiding backpropagation

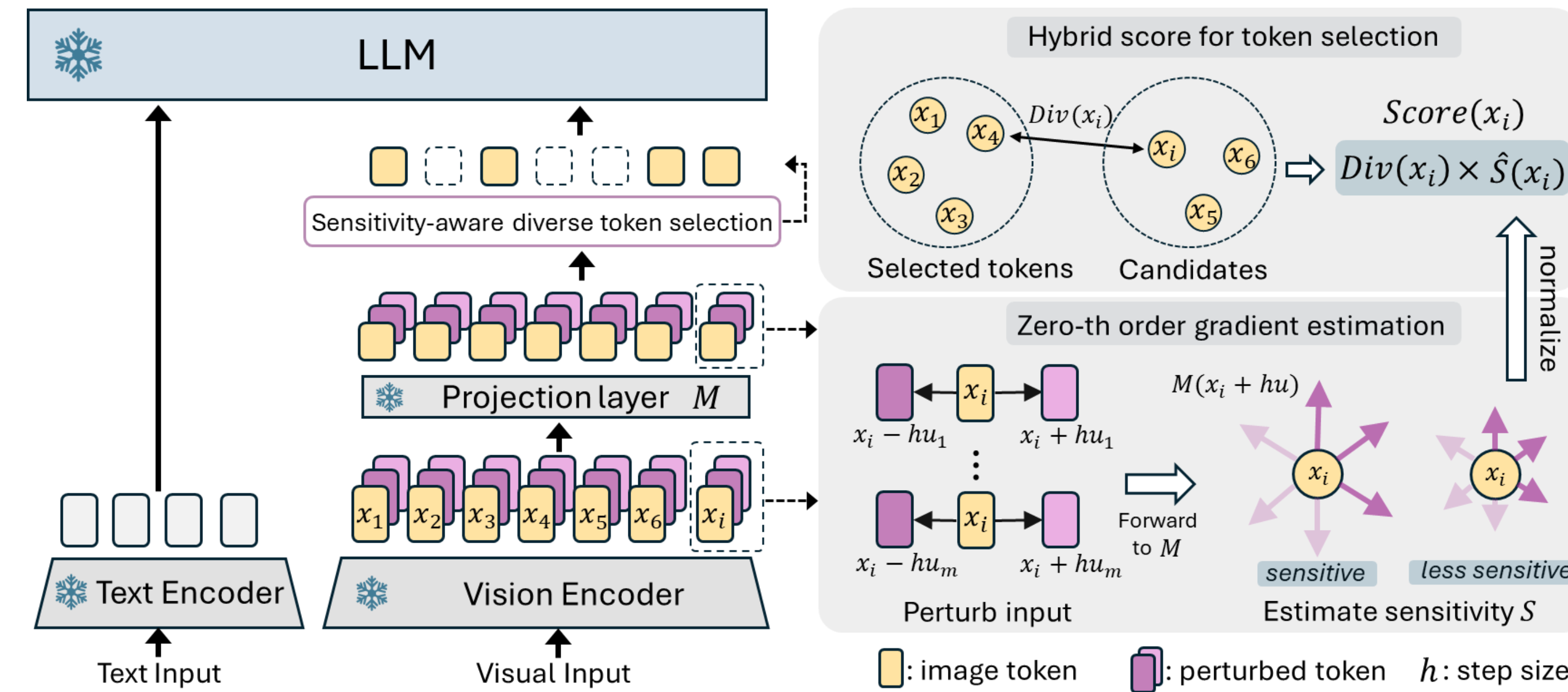
Projector as a Proxy for Visual Sensitivity:

- The projector and vision encoder exhibit consistent token-importance rankings.
- ZO sensitivity is well aligned with first-order (FO) sensitivity.



Method

ZOO-Prune: Zeroth-Order Gradient Estimation for Token Pruning



ZOO-based Sensitivity Estimation:

- Measures token importance via output changes
- Forward-only; no gradient required
- Provides an output-aware importance signal

Sensitivity-Aware Diversity Selection:

- Combines sensitivity and diversity
- Preserves important and non-redundant tokens
- Remains robust under aggressive pruning

Algorithm 1 ZOO-Prune

- Input:** Vision tokens $X \in \mathbb{R}^{N_v \times d_v}$, projection M , number of tokens to select k , step size h , number of perturbations m
- Output:** Selected token indices \mathcal{P}
- % [ZOO-based Sensitivity Estimation]
- Sample m random perturbations $U \in \mathbb{R}^{m \times d_v}$, $u_j \sim \mathcal{N}(0, I)$, normalized to $\|u_j\|_2 = 1$
- Expand X along perturbations: $X^+ = X + hU$, $X^- = X - hU$
- Project perturbed features: $Z^+ = M(X^+)$, $Z^- = M(X^-)$
- Compute finite-difference responses: $\Delta = \frac{Z^+ - Z^-}{2h}$
- Sensitivity: $S(i) = \frac{1}{m} \sum_{j=1}^m \|\Delta_{i,j}\|_2$
- % [Sensitivity-Aware Diversity Selection]
- Normalize sensitivities: $\hat{S}(i) = \frac{S(i) - \min_j S(j)}{\max_j S(j) - \min_j S(j)}$
- Initialize $\mathcal{P} \leftarrow \emptyset$
- while** $|\mathcal{P}| < k$ **do**
- Compute diversity: $\text{Div}(i, \mathcal{P}) = 1 - \max_{j \in \mathcal{P}} \cos(Z_i, Z_j)$ (set to 1 if \mathcal{P} is empty)
- Fusion score: $\text{Score}(i) = \hat{S}(i) \cdot \text{Div}(i, \mathcal{P})$
- Select $i^* = \arg \max_i \text{Score}(i)$
- $\mathcal{P} \leftarrow \mathcal{P} \cup \{i^*\}$
- end while**
- return** \mathcal{P}

Experiments

Table 1: Performance Comparison on LLaVA-1.5-7B

Method	GQA Acc. ↑	MMB Acc. ↑	MME P+C ↑	POPE F1 ↑	SQA Acc. ↑	VQA ^{V2} Acc. ↑	VQA ^{T=est} Acc. ↑	MMMU Acc. ↑	SEED-I Acc. ↑	Avg. ↑
Total 576 Tokens										
LLaVA-1.5-7B	61.90	64.70	1862.00	85.90	69.50	78.50	58.20	36.30	58.60	100%
Retain 192 Tokens ↓ 66.7%										
FastV (ECCV 2024)	52.70	61.20	1612.00	64.80	67.30	67.10	52.50	34.30	57.10	89.6%
SparseVLM (ICML 2025)	57.60	62.50	1721.00	83.60	69.10	75.60	56.10	33.80	55.80	95.5%
VisionZip (CVPR 2025)	59.30	63.00	1782.60	85.30	68.90	76.80	57.30	36.60	56.40	97.9%
DivPrune (CVPR 2025)	59.97	62.54	1762.23	87.00	68.91	76.87	56.97	35.44	58.71	98.0%
ZOO-Prune (Ours)	60.03	62.89	1781.66	87.24	69.16	77.34	57.30	36.11	58.80	98.6%
Retain 128 Tokens ↓ 77.8%										
FastV (ECCV 2024)	49.60	56.10	1490.00	59.60	60.20	61.80	50.60	34.90	55.90	84.5%
SparseVLM (ICML 2025)	56.00	60.00	1696.00	80.50	67.10	73.80	54.90	33.80	53.40	93.0%
VisionZip (CVPR 2025)	57.60	62.00	1761.70	83.20	68.90	75.60	56.80	37.90	54.90	96.8%
DivPrune (CVPR 2025)	59.25	62.03	1718.22	86.72	68.96	75.96	56.06	35.56	56.98	96.9%
ZOO-Prune (Ours)	59.49	61.86	1751.60	87.13	68.91	76.57	57.87	35.67	57.53	97.8%
Retain 64 Tokens ↓ 88.9%										
FastV (ECCV 2024)	46.10	48.00	1256.00	48.00	51.10	55.00	47.80	34.00	51.90	75.5%
SparseVLM (ICML 2025)	52.70	56.20	1505.00	75.10	62.20	68.20	51.80	32.70	51.10	87.0%
VisionZip (CVPR 2025)	55.10	60.10	1690.00	77.00	69.00	72.40	55.50	36.20	52.20	93.1%
DivPrune (CVPR 2025)	57.78	59.28	1674.40	85.56	68.17	74.11	54.69	35.56	55.13	94.8%
ZOO-Prune (Ours)	58.47	60.22	1675.59	85.86	68.27	75.02	55.35	35.44	55.84	95.5%

Table 2: Performance Comparison on LLaVA-1.5-13B

Method	GQA Acc. ↑	MMB Acc. ↑	MME P+C ↑	POPE F1 ↑	SQA Acc. ↑	VQA ^{V2} Acc. ↑	VQA ^{T=est} Acc. ↑	MMMU Acc. ↑	SEED-I Acc. ↑	Avg. ↑
Total 576 Tokens										
LLaVA-1.5-13B	63.20	67.70	1818.00	85.90	72.80	80.00	61.30	36.40	66.90	100%
Retain 192 Tokens ↓ 66.7%										
VisionZip (CVPR 2025)	59.10	66.90	1754.00	85.10	73.50	78.10	59.50	36.40	65.20	97.9%
DivPrune (CVPR 2025)	59.42	66.53	1781.50	86.76	72.83	77.98	58.46	36.56	65.72	98.1%
ZOO-Prune (Ours)	59.95	66.67	1762.41	86.73	73.12	78.65	59.11	37.33	65.56	98.6%
Retain 128 Tokens ↓ 77.8%										
VisionZip (CVPR 2025)	57.90	66.70	1743.00	85.20	74.00	76.80	58.70	36.10	63.80	97.0%
DivPrune (CVPR 2025)	58.89	66.07	1748.56	86.53	72.83	77.10	58.17	35.56	64.22	97.0%
ZOO-Prune (Ours)	58.89	67.01	1791.10	86.95	73.38	77.83	58.80	35.56	64.50	97.8%
Retain 64 Tokens ↓ 88.9%										
VisionZip (CVPR 2025)	56.20	64.90	1676.00	76.00	74.40	73.70	57.40	36.40	60.40	93.7%
DivPrune (CVPR 2025)	57.66	64.60	1777.93	84.80	71.34	75.20	57.11	35.22	62.44	95.4%
ZOO-Prune (Ours)	58.58	64.78	1780.83	85.34	72.09	76.39	58.59	36.00	63.02	96.5%

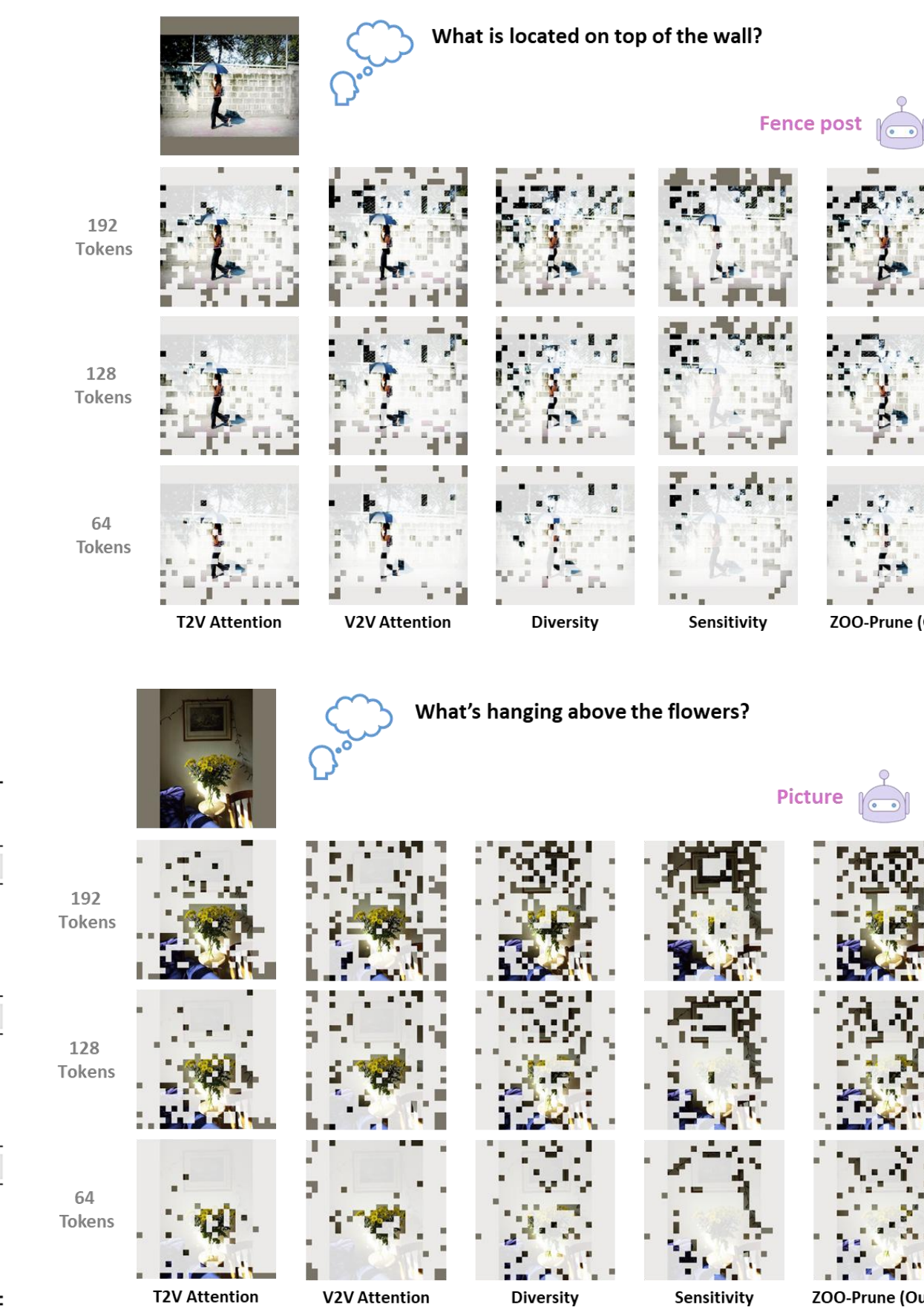
Table 3: Performance Comparison on LLaVA-NeXT-7B

Method	GQA Acc. ↑	MMB Acc. ↑	MME P+C ↑	POPE F1 ↑	SQA Acc. ↑	VQA ^{V2} Acc. ↑	VQA ^{T=est} Acc. ↑	MMMU Acc. ↑	SEED-I Acc. ↑	Avg. ↑
Total 2880 Tokens										
LLaVA-NeXT-7B	64.20	67.90	1842.00	86.40	70.20	80.10	61.30	35.10	70.20	100%
Retain 640 Tokens ↓ 77.8%										
SparseVLM (ICML 2025)	60.30	65.70	1772.00	—	67.70	77.10	57.80	34.60	—	—
VisionZip (CVPR 2025)	61.30	66.30	1787.00	86.30	68.10	79.10	60.20	34.70	66.70	97.5%
DivPrune (CVPR 2025)	61.58	65.38	1773.04	85.51	67.82	78.94	55.41	36.89	67.56	97.1%
ZOO-Prune (Ours)	62.19	65.21	1816.45	86.75	68.02	79.64	57.98	36.89	67.95	98.3%
Retain 320 Tokens ↓ 88.9%										
SparseVLM (ICML 2025)	57.70	64.30	1694.00	—	67.30	73.40	55.90	34.40	—	—
VisionZip (CVPR 2025)	59.30	63.10	1702.00	82.10	67.30	76.20	58.90	35.30	63.40	94.5%
DivPrune (CVPR 2025)	59.63	63.66	1731.04	83.47	67.82	76.64	53.84	37.11	65.35	95.1%
ZOO-Prune (Ours)	60.97	64.86	1787.68	85.47	67.77	78.08	57.28	37.00	66.47	97.1%
Retain 160 Tokens ↓ 94.4%										
SparseVLM (ICML 2025)	51.20	63.10	1542.00	—	67.50	66.30	46.40	32.80	—	—
VisionZip (CVPR 2025)	55.50	60.10	1630.00	74.80	68.30	71.40	56.20	36.10	58.30	90.4%
DivPrune (CVPR 2025)	57.79	62.29	1658.25	79.36	68.02	73.92	52.42	36.44	62.54	92.4%
ZOO-Prune (Ours)	59.93	64.18	1738.64	83.05	68.42	76.12	55.42	37.11	64.05	95.4%

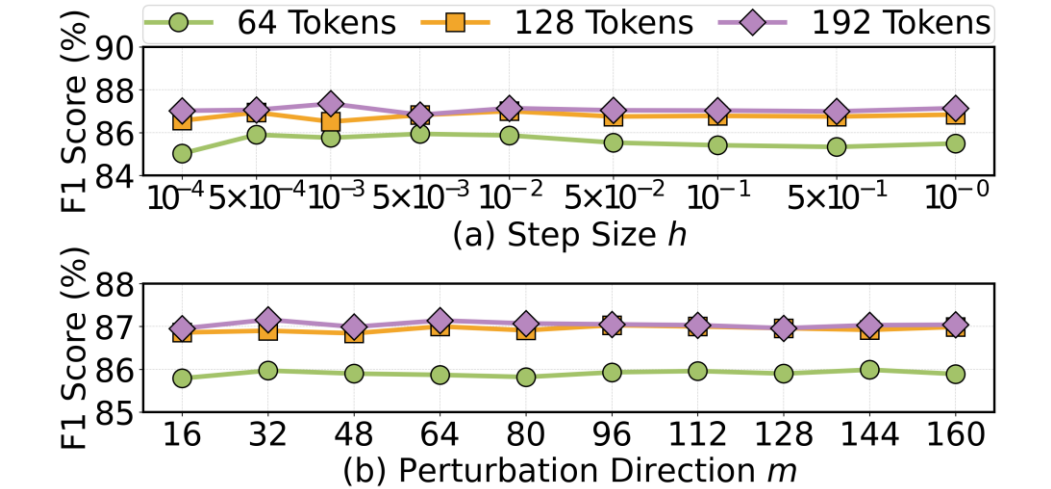
Table 4: Ablation on Token Selection Metrics with LLaVA-NeXT-7B

Sensitivity	Diversity	Fusion	GQA Acc. ↑	MMB Acc. ↑	MME P+C ↑	POPE F1 ↑	SQA Acc. ↑	VQA ^{V2} Acc. ↑	VQA ^{T=est} Acc. ↑	MMMU Acc. ↑	SEED-I Acc. ↑	Avg. ↑
Retain 640 Tokens ↓ 77.8%												
✓	✓	-	61.23	65.21	1818.62	86.54	68.07	78.47	54.12	35.78	66.29	96.7%
✓	✓	Sum	61.58	65.38	1773.04	85.51	67.82	78.94	55.41	36.89	67.56	97.1%
✓	✓	Multiply	62.19	65.21	1816.45	86.75	68.02	79.64	57.98	36.89	67.95	98.3%
Retain 320 Tokens ↓ 88.9%												
✓	✓	-	59.22	64.69	1744.42	83.15	67.63	75.69	47.25	34.78	63.69	92.9%
✓	✓	Sum	59.63	63.66	1731.04	83.47	67.82	76.64	53.84	37.11	65.35	95.1%
✓	✓	Multiply	60.97	64.83	1787.68	85.47	67.77	78.08	57.28	37.00	66.47	97.1%
Retain 160 Tokens ↓ 94.4%												
✓	✓	-	57.23	61.86	1674.35	77.27	68.82	72.58	50.96	35.56	61.04	91.2%
✓	✓	Sum	57.79	62.29	1658.25	79.36	68.02	73.92	52.42	36.44	62.54	92.4%
✓	✓	Multiply	59.93	64.18	1738.64	83.05	68.42	76.12	55.42	37.11	64.05	95.4%

Qualitative Comparison:



Hyperparameter Analysis:



Inference Efficiency:

