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

Supplementary Material

This appendix supplements the main paper with the theoretical analysis of Proposition 3.1, additional details on the experimental configuration and evaluation strategy, and further quantitative and qualitative results that illustrate the token preservation patterns across a range of scenarios. The contents are organized as follows:

- **Appendix A**: Implementation details for KDE experiments in Section 3.2 of the main paper.
- **Appendix B**: Theoretical analysis of the proposed sensitivity estimator.
- **Appendix C**: Experimental setup, implementation details, and evaluation protocols.
- **Appendix D**: Further quantitative results and inference efficiency analysis.
- **Appendix E**: Additional qualitative results, including both successful and failure cases.
- **Appendix F**: Discussion and Future Work.

A. Spearman Correlation Setup

For the correlation analysis in Fig. 2 of the main paper, we examined how closely the token-importance rankings from the vision encoder matched those from the projection layer. Specifically, we selected 50 random samples per dataset from MMMU and POPE. For each sample, token sensitivities were first computed using RGE at both the vision encoder output and the projection layer. To ensure stable ranking comparisons, we applied a 0.5 threshold to filter out low-sensitivity tokens before computing ranks. The Spearman's rank correlation coefficient was then calculated for each sample, and the distribution across 50 samples was visualized using a kernel density estimate (KDE) plot.

The resulting average Spearman correlations were 0.55 for MMMU and 0.49 for POPE, indicating a consistent alignment between token rankings obtained at the projection layer and those from the full vision encoder. This confirms that the projection layer can serve as a reliable proxy for token-level importance estimation while significantly reducing the computational overhead.

B. Theoretical Analysis

B.1. Proof of Proposition

Proposition B.1 (Approximated Mean Sensitivity)**.** *Let $M : \mathbb{R}^n \to \mathbb{R}^m$ be differentiable at $x \in \mathbb{R}^n$ with Jacobian $J(x) = \nabla M(x)$. Let $u \sim \mathcal{N}(0, I_n)$ be an isotropic Gaussian perturbation and $h > 0$ a small step size. Define the finite-difference sensitivity $S(x) = \mathbb{E}_u \left[\left\| \frac{M(x+hu) - M(x-hu)}{2h} \right\|_2 \right]$. Then, for sufficiently small h,*

$$S(x) = \mathbb{E}_u \left[\|J(x)u\|_2 \right] + O(h^2). \tag{1}$$

Proof. Since M is differentiable at x, we apply a first-order Taylor expansion around x for perturbations hu:

$$M(x + hu) = M(x) + hJ(x)u + O(h^2), \tag{2}$$

$$M(x - hu) = M(x) - hJ(x)u + O(h^2). \tag{3}$$

Subtracting and dividing by $2h$ gives the symmetric finite-difference approximation:

$$\frac{M(x + hu) - M(x - hu)}{2h} = J(x)u + O(h^2). \tag{4}$$

Taking the ℓ_2-norm,

$$\left\| \frac{M(x+hu) - M(x-hu)}{2h} \right\|_2 = \|J(x)u + O(h^2)\|_2 \tag{5}$$

$$= \|J(x)u\|_2 + O(h^2). \tag{6}$$

Finally, taking expectation over isotropic Gaussian perturbations $u \sim \mathcal{N}(0, I_n)$ yields

$$S(x) = \mathbb{E}_u \left[\|J(x)u\|_2 \right] + O(h^2). \tag{7}$$

□

This proposition establishes that the finite-difference sensitivity $S(x)$, computed using small isotropic Gaussian perturbations, provides an accurate approximation of the mean local effect of input changes on the output. Specifically, for sufficiently small step size h, the finite-difference estimate is equivalent, up to an $O(h^2)$ error, to the expected ℓ_2-norm of the Jacobian applied to random Gaussian directions. Intuitively, this means that $S(x)$ captures the average magnitude of output variation induced by small, randomly oriented perturbations in the input space. By sampling u from an isotropic Gaussian, all directions are treated equally, ensuring an unbiased and comprehensive measure of token sensitivity without requiring backpropagation.

C. Experimental Setup

C.1. Model Settings

We evaluate the effectiveness of *ZOO-Prune* on widely used VLMs, including LLaVA-v1.5-7B [9], LLaVA-v1.5-13B [9], and LLaVA-1.6-7B [10] (also referred to as LLaVA-NeXT-7B), and Qwen2.5-VL-7B [2]. All LLaVA

Table A. Summary of primary evaluation metrics.

Benchmark	Primary Metric
VQAv2, GQA, ScienceQA, TextVQA	Accuracy (Acc.)
MMBench, MMMU, SeedBench	Accuracy (Acc.)
POPE	F1-score (F1)
MME	Perception + Cognition (P+C)

models adopt the CLIP [13] as the vision encoder and Vicuna [4] as the base language model.

LLaVA-v1.5 models process images at 336×336 resolution, yielding 576 visual tokens, while LLaVA-NeXT-7B supports higher resolutions (up to 672×672), generating up to 2,880 tokens and achieving a 6.0% gain at the cost of $3.5\times$ more computation. Qwen2.5-VL-7B, in contrast, utilizes a dynamic-resolution ViT encoder with window attention and is built upon the Qwen2.5-7B language model, supporting a variable number of visual tokens depending on input resolution. Across all experiments, our pruning is applied in a fully training-free and calibration-free manner.

C.2. Implementation Details

All experiments are conducted on 4×NVIDIA A6000 GPUs with a batch size of 1. *ZOO-Prune* is entirely training-free and attention-free, requiring no manual specification of layers in either the LMM or the vision encoder. Token selection is performed at the lightweight projection layer, which enables seamless integration across different VLM architectures. Sensitivities are also computed at this layer using simple perturbation-based operations, ensuring negligible computational overhead during inference.

For pruning ratios, we adopt 66.7% / 77.8% / 88.9% for LLaVA-v1.5 and 77.8% / 88.9% / 94.4% for LLaVA-NeXT-7B. In the latter case, we follow the implementation of VisionZip [15], where the model dynamically samples up to five image patches, resulting in as many as 2,880 vision tokens. For example, with a pruning budget of 160 tokens, we retain 32 tokens per patch across five patches ($32 \times 5 = 160$). If fewer patches are sampled (*e.g.*, four), the number of retained tokens is adjusted proportionally (*e.g.*, 128 tokens for 160 / 2880). We applied a low-rank factorization ($k = 128$) to the MM-projector layers to further boost efficiency on LLaVA-NeXT, due to the large number of visual tokens. For the dynamic-resolution Qwen2.5-VL-7B, we evaluate at 10% and 20% token retention rates.

Finally, as validated in ablation, our method remains robust across different hyperparameter choices. Unless otherwise noted, we fix the perturbation hyperparameters to $m = 64$ and $h = 0.01$ for all experiments. Evaluation is performed using the `lmms-eval` [17] framework under official protocols and metrics.

C.3. Evaluation Protocol and Benchmark Datasets

We conduct a comprehensive evaluation of *ZOO-Prune* across **nine widely adopted vision-language benchmarks**, spanning four core capabilities: *Visual Question Answering*, *Advanced Multimodal Reasoning*, *Object Hallucination Evaluation*, and *Comprehensive Multimodal Assessment*. All experiments strictly follow the official evaluation protocols, metrics, and data splits of each benchmark to ensure fair and reproducible comparisons.

To facilitate a unified and interpretable comparison, we report both per-benchmark scores and a normalized average performance (Avg.), computed as the mean relative score across benchmarks with respect to the unpruned baseline. Depending on the benchmark, we report Accuracy (Acc), F1-score (F1), or Perception+Cognition (P+C), summarized in Table A. All evaluations of *ZOO-Prune* are performed under a single-model, zero-shot setting, without any task-specific fine-tuning.

Visual Question Answering (VQA). This category evaluates a model's ability to ground language understanding in visual content. Performance across all VQA benchmarks is measured by **Accuracy (Acc)**. We select four representative benchmarks covering diverse scenarios:

- **VQAv2-Test-Dev** [5]: General-purpose VQA with real-world images and open-ended questions.
- **GQA** [6]: Focused on compositional reasoning over scene graphs and structured images.
- **ScienceQA (IMG)** [12]: Multimodal science questions requiring domain knowledge and diagram interpretation.
- **TextVQA** [14]: Requires OCR capabilities to reason over text embedded within images.

Advanced Multimodal Reasoning. To probe deeper reasoning capacities beyond standard VQA, we evaluate on three challenging benchmarks. Performance on these benchmarks is also measured by **Accuracy (Acc)**:

- **MMBench** [11]: Assesses perception and reasoning across 20 fine-grained skill areas.
- **MMMU** [16]: Requires expert-level multimodal reasoning across 30+ subjects grouped into six major disciplines (*e.g.*, Art & Design, Science, Engineering, Medicine), often involving complex diagrams and charts
- **SeedBench** [7]: Designed for evaluating multimodal large language models across diverse visually grounded question types, with an emphasis on perception, reasoning, and knowledge.

Object Hallucination Evaluation. To quantify the critical failure mode of object hallucination, we adopt the **POPE** [8] benchmark, which measures factuality in object recognition through binary existence questions. Performance is evaluated using the **F1-score (F1)** over object existence predictions, balancing precision and recall to reflect grounding reliability.

Table B. Performance Comparison on LLaVA-1.5-7B.

Method	GQA Acc. ↑	MMB Acc. ↑	MME P+C ↑	POPE F1 ↑	SQA Acc. ↑	VQA^{V2} Acc. ↑	VQA^{Text} Acc. ↑	MMMU Acc. ↑	SEED 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%**

Comprehensive Multimodal Assessment. For a holistic evaluation of both perceptual and cognitive abilities across numerous sub-tasks (*e.g.*, OCR, counting, attribute recognition), we employ the **MME** [18]. It reports separate scores for Perception and Cognition tasks, summed to form the combined **Perception and Cognition score (P+C)**.

D. Further Quantitative Results and Analysis

D.1. Additional Quantitative Results

Results on LLaVA-1.5-7B: Table B reports the results on LLaVA-1.5-7B, the most widely used model in the LLaVA family. Across all token pruning levels, *ZOO-Prune* achieves the strongest overall performance. With a 66.7% pruning (192 tokens), it reaches an Avg. of 98.6%, surpassing the previous best method VisionZip (97.9%) and achieving leading scores on challenging tasks such as MMMU (36.11%) and SEED (58.80%). As pruning becomes more aggressive, the advantage of *ZOO-Prune* becomes even clearer. At 128 tokens, it maintains the best average score of 97.8%, substantially outperforming FastV (84.5%) and SparseVLM (93.0%). *ZOO-Prune* also delivers top results on SQA (59.49%), POPE (87.13%), VQAv2 (76.57%), and TextVQA (57.87%). Even under the extremely compressed 64-token setting, *ZOO-Prune* retains an average of 95.5%, outperforming the attention-based VisionZip by +2.4% and the diversity-based DivPrune by +0.7%. These results demonstrate that our sensitivity-aware diversity selection reliably preserves key visual cues and remains robust even under aggressive pruning.

Results on LLaVA-1.5-13B: The robustness of ZOO-Prune is further validated in Table C, where methods are applied to the larger LLaVA-1.5-13B model. This evaluation tests the ability of pruning strategies to generalize to higher-capacity backbones. With 192 tokens retained, *ZOO-Prune* again obtains the highest average score (98.6%). Under the challenging 64-token setting, *ZOO-Prune* shows a clear performance advantage. It reaches 96.5% average accuracy, outperforming VisionZip by +2.8% and DivPrune by +1.1%. Retaining strong performance while discarding 88.9% of visual tokens highlights the robustness of our approach. Overall, across both 7B and 13B variants, ZOO-Prune consistently provides the best accuracy–efficiency trade-off, demonstrating strong generalization to different model capacities and pruning regimes.

D.2. Further Inference Efficiency Analysis

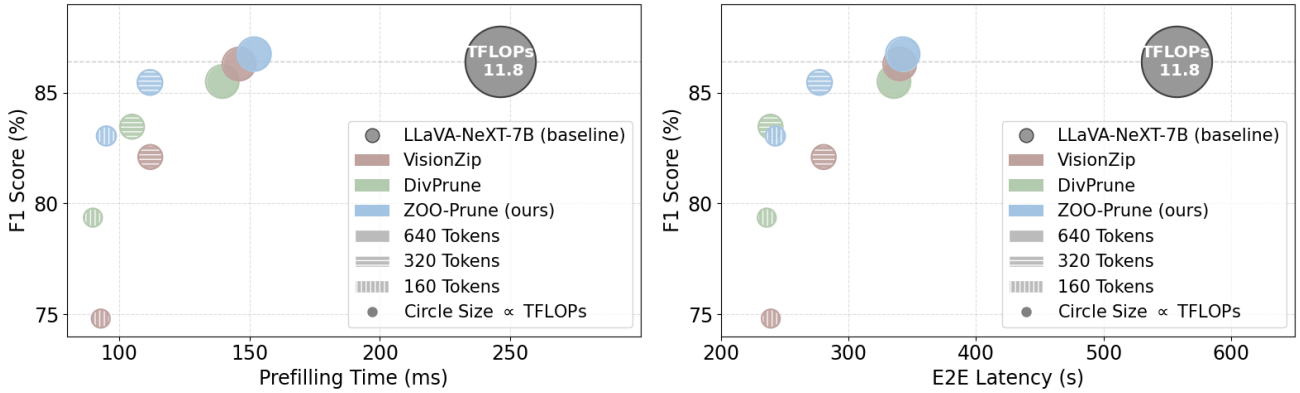
We evaluate the computational benefits of *ZOO-Prune* by measuring prefilling time, end-to-end (E2E) latency, and FLOPs on LLaVA-NeXT-7B using the POPE benchmark. The total computational cost can be expressed as:

$$\text{FLOPs}_{\text{total}} = \text{FLOPs}_{\text{prefill}}(\hat{n}) + \Delta\text{FLOPs}_{\text{ZOO}}, \quad (8)$$

where n and $\hat{n}$ denote the sequence length before and after pruning, respectively. Following prior work [1, 3, 15], the prefilling FLOPs of an L-layer LLM with hidden size d and FFN intermediate size m scale as: $L(4\hat{n}d^2+2\hat{n}^2d+2\hat{n}dm)$. Our pruning overhead comes only from ZOO sensitivity estimation at the lightweight projector. With m_z random perturbation directions and a rank-k factorization, the addi-

Table C. Performance Comparison on LLaVA-1.5-13B.

Method	GQA Acc. ↑	MMB Acc. ↑	MME P+C ↑	POPE F1 ↑	SQA Acc. ↑	VQA^{V2} Acc. ↑	VQA^{Text} 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	66.58	**1781.50**	**86.76**	72.88	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	1780.03	**85.34**	72.09	**76.39**	**58.59**	36.00	**63.02**	**96.5%**



(a) Prefilling Time (ms) vs. Performance (%) vs. TFLOPs

(b) E2E Latency (secs) vs. Performance (%) vs. TFLOPs

Figure A. Inference efficiency on the POPE benchmark relative to the LLaVA-NeXT-7B baseline. Circle size is proportional to TFLOPs. *ZOO-Prune* achieves the strongest accuracy–efficiency trade-off, sustaining higher POPE performance while operating at substantially lower latency and computational cost.

tional cost is:

$$\Delta\text{FLOPs}_{\text{ZOO}} = \text{FLOPs}_{\text{ZOO,LR}} - \text{FLOPs}_{\text{proj}} \quad (9)$$

$$= 2m_z\, n\, \frac{d}{k}(D_v + d) - nD_v d \quad (10)$$

$$= nd^2, \quad \text{when } m_z{=}64, k{=}128. \quad (11)$$

Therefore, under our default configuration, this overhead is about 0.3% of the baseline prefilling FLOPs.

As shown in Fig. A, all pruning methods provide noticeable efficiency gains across token budgets, but *ZOO-Prune* offers the most favorable balance between speed and accuracy. At the most aggressive setting (160 tokens), prefilling becomes $2.59\times$ faster and E2E latency is reduced by $2.30\times$ relative to the baseline. FLOPs show a similarly substantial reduction: from 11.8 TFLOPs in the baseline to just 0.9 TFLOPs with *ZOO-Prune*. Importantly, despite this nearly $13\times$ computational reduction, *ZOO-Prune* retains 96.12% of the baseline's performance (83.05% F1), indicating only minimal accuracy degradation. These results support that *ZOO-Prune* delivers a consistently superior accuracy–efficiency trade-off, making it a practical and effective pruning strategy for real-world VLM deployment.

D.3. Visual Sensitivity: Projector vs. Encoder

The projection layer acts as a modality-alignment bottleneck, where high-level visual features are consolidated and mapped into the LLM embedding space. Tokens that remain influential at this stage are therefore those most critical for downstream language generation. Motivated by this, *ZOO-Prune* computes ZOO-based sensitivity at the projector rather than across the full vision encoder, avoiding costly end-to-end perturbations.

To evaluate whether the projector serves as a reliable proxy, we compare performance when sensitivity is computed at each encoder layer. As shown in Table D, mid-to-deep encoder layers (*e.g.*, 17–24 layers) provide moderately strong estimates but consistently underperform the

Table D. Performance comparison between the projector and vision encoder layers as sensitivity proxies. The projector consistently provides the strongest accuracy across pruning ratios.

Vision Encoder	**GQA** Acc. ↑	**MMB** Acc. ↑	**MME** P+C ↑	**POPE** F1 ↑	**SQA** Acc. ↑	**VQA**^{Text} Acc. ↑	**MMMU** Acc. ↑	**SEED-I** Acc. ↑	**Avg.** ↑
Total 2880 Tokens									
LLaVA-NeXT-7B	64.20	67.90	1842.00	86.40	70.20	61.30	35.10	70.20	100%
Retain 640 Tokens ↓ 77.8%									
Layer 01	61.23	65.29	1779.27	81.41	67.87	52.49	37.56	66.25	95.8%
Layer 03	61.96	**65.64**	1819.21	86.35	**68.47**	57.80	36.67	67.68	98.1%
Layer 05	61.73	65.55	1815.92	86.74	68.17	57.43	36.78	67.59	97.9%
Layer 07	61.85	65.29	1815.84	86.41	67.18	56.85	37.89	68.07	98.1%
Layer 09	62.18	65.12	1813.60	86.44	67.97	57.39	37.67	67.33	98.1%
Layer 11	61.98	65.21	1806.98	86.41	68.17	57.41	37.22	67.69	98.0%
Layer 13	61.93	65.12	1803.63	86.47	67.72	54.06	37.78	67.62	97.4%
Layer 15	62.03	65.38	1782.60	86.72	67.92	56.41	37.56	67.98	97.8%
Layer 17	61.99	65.38	1803.06	**86.83**	67.87	55.57	**38.00**	**68.28**	98.0%
Layer 19	61.94	65.03	**1840.15**	86.78	67.97	56.07	37.44	68.12	98.1%
Layer 21	62.03	65.29	1820.98	86.78	67.77	57.94	36.89	67.71	98.1%
Layer 23	61.70	65.46	1829.87	84.83	67.23	56.53	37.56	67.32	97.6%
Layer 24	61.99	65.21	1838.67	86.34	67.63	**58.31**	36.56	67.77	98.1%
Projector (Ours)	**62.19**	65.21	1816.45	86.75	68.02	57.98	36.89	67.95	**98.2%**
Retain 320 Tokens ↓ 88.9%									
Layer 01	59.11	63.92	1664.19	76.08	**68.86**	49.38	36.89	63.94	92.4%
Layer 03	60.71	64.09	1796.69	84.08	67.87	56.74	36.44	65.64	96.3%
Layer 05	60.70	64.60	1789.64	84.48	67.82	55.77	**38.44**	65.77	96.9%
Layer 07	61.05	64.60	1746.22	83.23	67.67	55.36	37.11	65.87	96.0%
Layer 09	60.90	63.92	1798.27	83.74	67.53	55.81	37.00	65.62	96.2%
Layer 11	60.73	64.26	1784.26	84.36	65.80	56.13	36.56	65.80	95.9%
Layer 13	60.78	63.83	1712.90	83.76	67.38	52.13	36.67	65.64	94.7%
Layer 15	60.99	63.57	1754.49	84.44	67.53	55.02	37.22	66.39	96.0%
Layer 17	60.83	**64.86**	1799.67	85.00	67.33	53.70	37.11	**66.48**	96.3%
Layer 19	61.05	64.52	**1801.79**	85.44	67.63	54.51	37.67	66.42	96.8%
Layer 21	61.00	64.26	1764.22	84.37	66.93	56.05	37.67	65.72	96.4%
Layer 23	60.43	63.57	1766.28	81.46	68.12	54.95	36.89	65.19	95.3%
Layer 24	**61.15**	64.78	1775.97	84.23	67.03	56.14	37.00	66.29	96.4%
Projector (Ours)	60.97	**64.86**	1787.68	**85.47**	67.77	**57.28**	37.00	66.47	**97.1%**
Retain 160 Tokens ↓ 94.4%									
Layer 01	57.62	61.34	1590.15	70.07	68.12	46.57	36.67	61.45	89.1%
Layer 03	58.88	63.14	1689.52	79.73	67.13	55.26	35.67	62.99	93.2%
Layer 05	59.69	62.54	1705.96	80.85	68.07	54.47	36.11	63.99	93.9%
Layer 07	59.43	63.92	1715.39	78.50	67.38	52.76	**39.00**	63.66	94.3%
Layer 09	59.58	62.46	1712.23	79.35	67.63	53.39	35.67	63.42	93.1%
Layer 11	59.18	63.23	1700.83	80.18	63.59	54.22	37.22	63.59	93.3%
Layer 13	59.33	62.80	1690.78	79.55	67.77	49.61	38.00	63.60	93.1%
Layer 15	59.78	63.14	1685.21	82.00	66.98	52.99	36.44	64.51	93.8%
Layer 17	**60.06**	63.32	1712.95	82.54	66.68	50.78	37.44	**64.78**	94.0%
Layer 19	59.90	62.63	1724.21	82.65	67.63	51.40	37.56	64.43	94.2%
Layer 21	59.53	63.66	1720.90	80.69	67.72	53.94	36.78	63.46	94.1%
Layer 23	58.80	62.71	1665.75	75.96	68.17	52.13	36.67	62.43	92.2%
Layer 24	59.57	63.14	1714.08	81.01	67.67	55.00	37.00	63.87	94.4%
Projector (Ours)	59.93	**64.18**	**1738.64**	**83.05**	**68.42**	**55.42**	37.11	64.05	**95.4%**

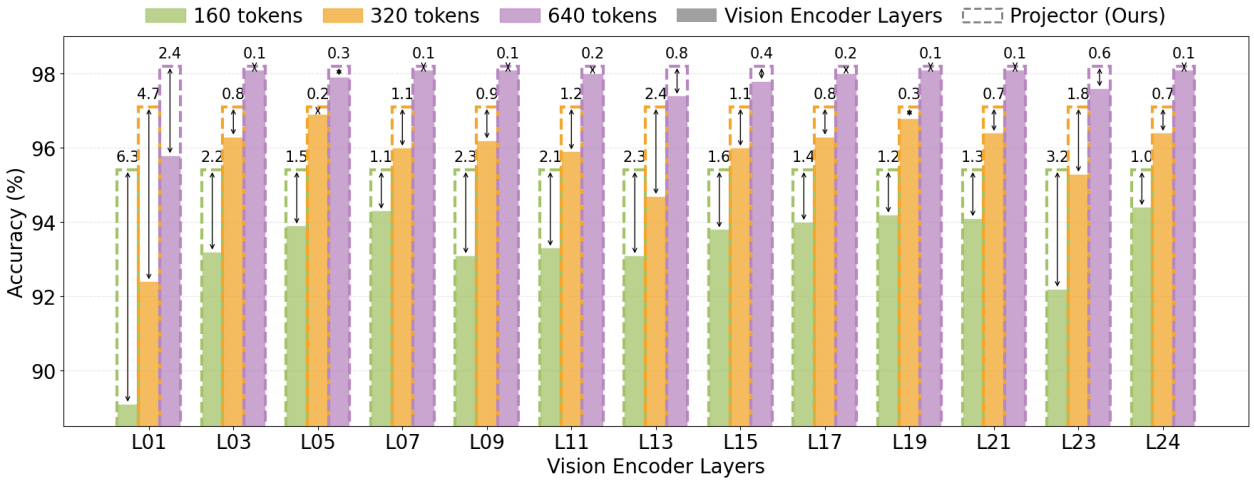


Figure B. Revisiting Token Sensitivity Proxy: Projector vs. Encoder. Compared to the visual encoder, the projector is a reliable proxy for zoo-based token sensitivity estimation.

projector, while early layers degrade sharply under aggressive pruning (*e.g.*, Layer 1 falls to 89.1% at a 160-token budget). Fig. B further highlights this trend: the projector achieves the highest and most stable accuracy across all pruning levels, with its advantage growing more pronounced as token budgets shrink. Overall, these findings show that the projection layer delivers the most stable, accurate, and efficient signal for token sensitivity, motivating its use as the default estimation strategy in *ZOO-Prune*.

E. Additional Qualitative Results

Sensitivity vs. Attention vs. Diversity. Fig. C compares token selection patterns driven by different token importance strategies. Attention-based methods exhibit clear limitations: text-visual (T2V) attention often suffers from positional bias (*e.g.*, disproportionately focusing on lower image regions, leading to the omission of key contextual cues), while visual-visual (V2V) attention tends to form redundant clusters. In contrast, diversity-based pruning treats all tokens equally, lacking semantic focus, whereas sensitivity-based selection captures key semantics but may result in spatial clustering. *ZOO-Prune* addresses these shortcomings by unifying sensitivity and diversity. This ensures that selected tokens are both semantically informative and spatially distributed, maintaining robust coverage of task-relevant regions even under an aggressive 64-token budget.

More Visualization Examples. Fig. D visualizes the pruning masks and QA results. VisionZip prioritizes visual saliency, often focusing on the high-contrast kitchen background while missing the foreground "Bench" across all ratios. DivPrune enforces diversity but compromises object integrity (*e.g.*, it sparsifies "Dog"-related tokens, leading to a Bear" hallucination). ZOO-Prune overcomes these limitations by using sensitivity to anchor dense token clusters on semantic targets. It effectively preserves the visual cues of key objects, ensuring correct predictions even when 88% of tokens are removed. This localization capability is further corroborated by Fig. F, demonstrating that *ZOO-Prune* consistently retains semantically critical tokens.

Failure Case Examples. Fig. E illustrates representative failure cases under aggressive pruning, where *ZOO-Prune* produces semantically close but inexact predictions. These errors reveal two limitations: difficulty in capturing fine-grained attributes for distinguishing closely related concepts, and vulnerability to visual clutter with multiple interfering objects. While our method preserves high-level semantic integrity at 64 tokens, extreme reduction inevitably compromises nuanced visual cues required for precise recognition in complex scenes.

F. Discussion and Future Work

Although *ZOO-Prune* achieves strong gains across vision–language reasoning tasks, several avenues remain open. First, our evaluation focuses on encoder–decoder VLMs such as LLaVA-NeXT. Extending *ZOO-Prune* to broader architectures, including emerging Omni-style unified models, may reveal different sensitivity patterns and grounding behaviors. Second, the method is currently image-centric. Applying *ZOO-Prune* to video, 3D scenes, or egocentric data will require reconsidering token selection under temporal and geometric structure. Finally, integrating *ZOO-Prune* into Vision-Language-Action agents is a promising direction, as adaptive preservation of task-critical visual cues may improve long-horizon stability and reduce cascading errors. In sum, extending *ZOO-Prune* toward Omni-style models, richer modalities, and interactive agents presents a promising direction for broadening the scope and impact of test-time visual refocusing.

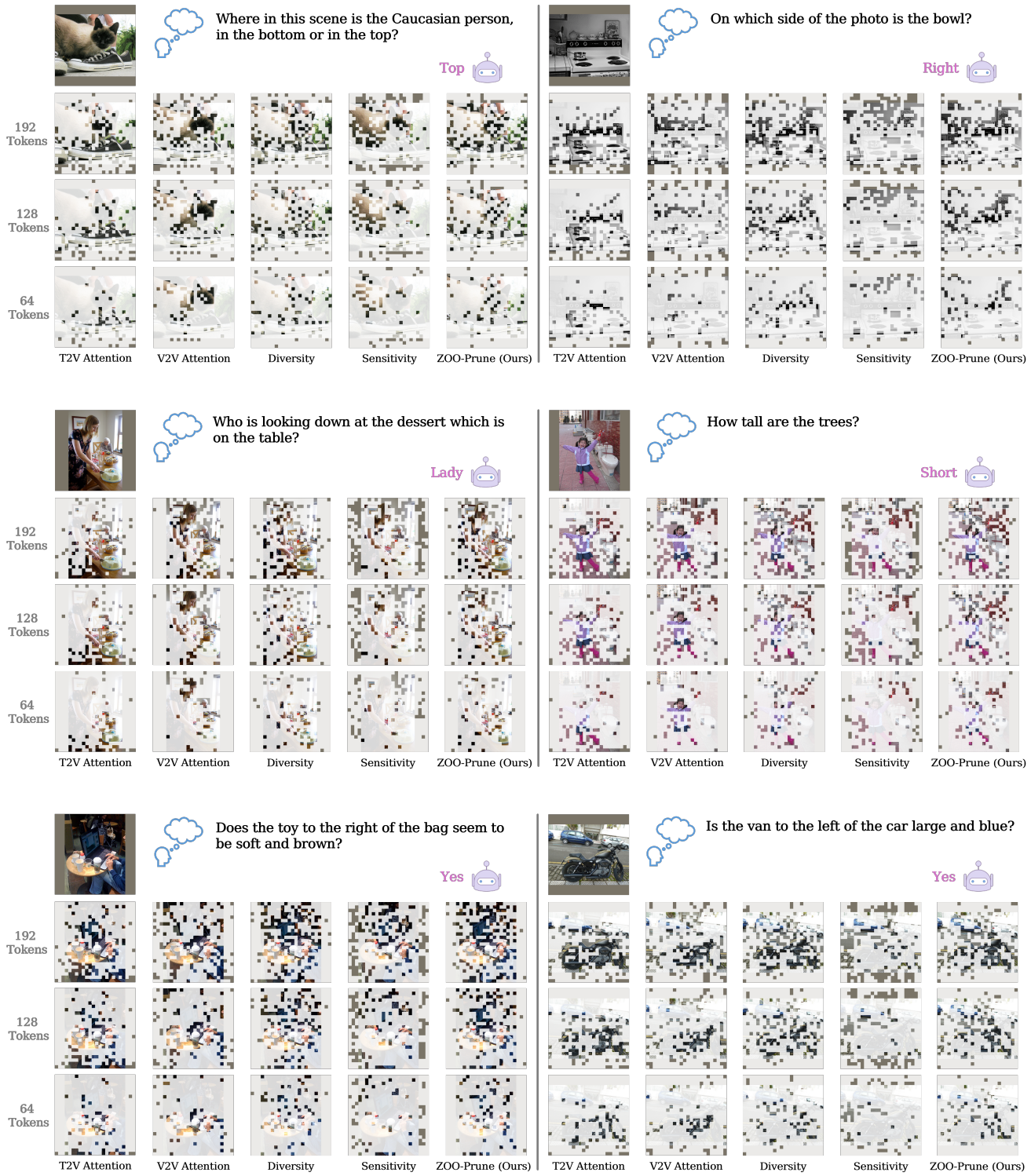


Figure C. Selected visual tokens comparison on the GQA benchmark. Token pruning driven by text-visual (T2V) attention often suffers from positional bias, while visual-visual (V2V) attention tends to retain redundant token clusters. Diversity-based pruning spreads tokens broadly but lacks semantic focus. ZOO-based sensitivity can capture output-related tokens but overlooks spatial coverage. Our ZOO-Prune jointly optimizes sensitivity and diversity for balanced selection across compression ratios.

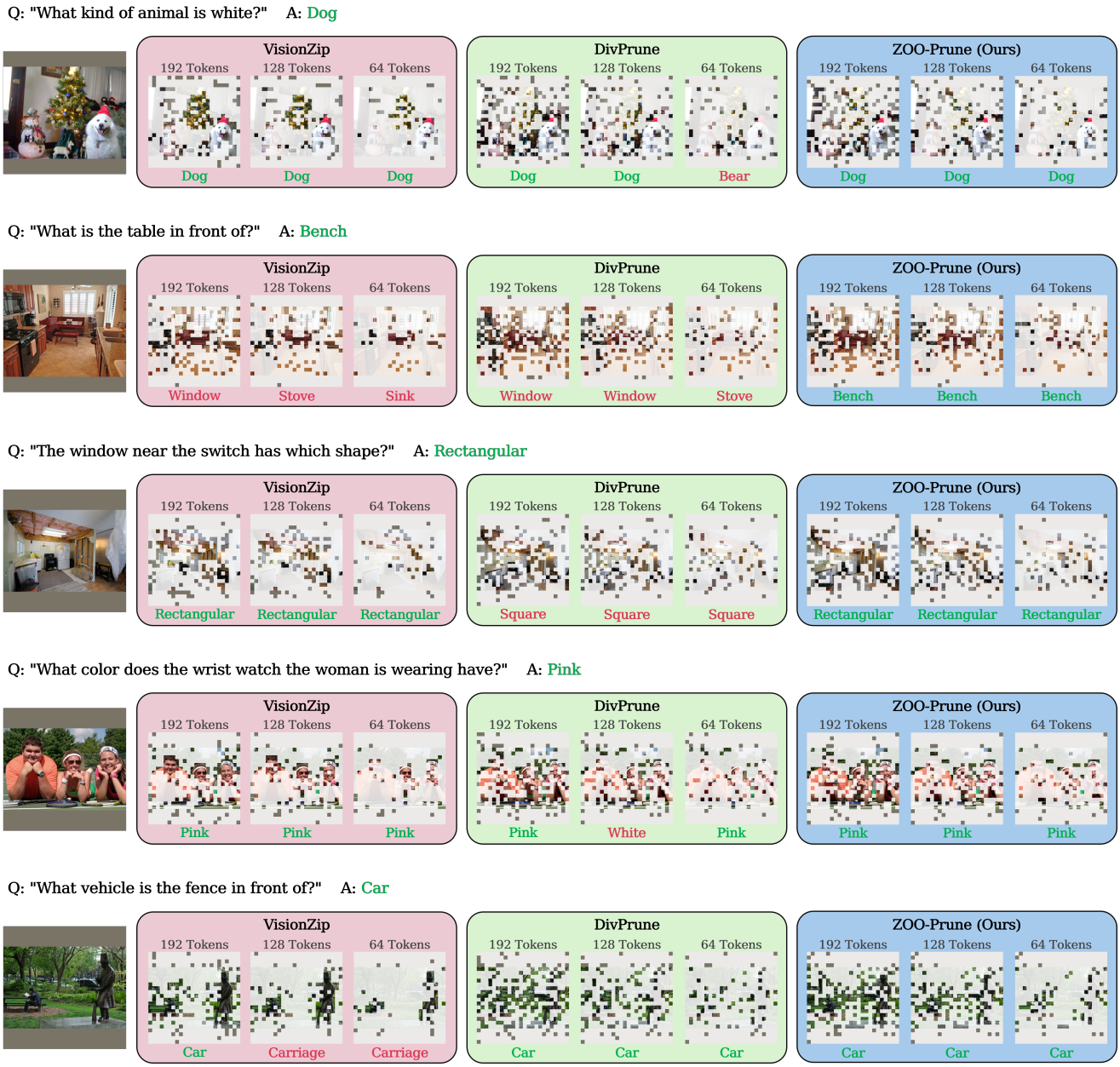


Figure D. Qualitative comparison of pruned token masks from VisionZip, DivPrune and *ZOO-Prune* on the GQA benchmark.

Figure E. Failure cases of *ZOO-Prune* illustrating two typical error patterns. Left: Predictions remain semantically close to references but differ in fine-grained age distinctions (e.g., "boy" vs. "baby"). Right: In visually cluttered scenes with multiple objects, the model makes incorrect predictions due to confusion among different items present in the scene (e.g., "mug" vs. "glass").

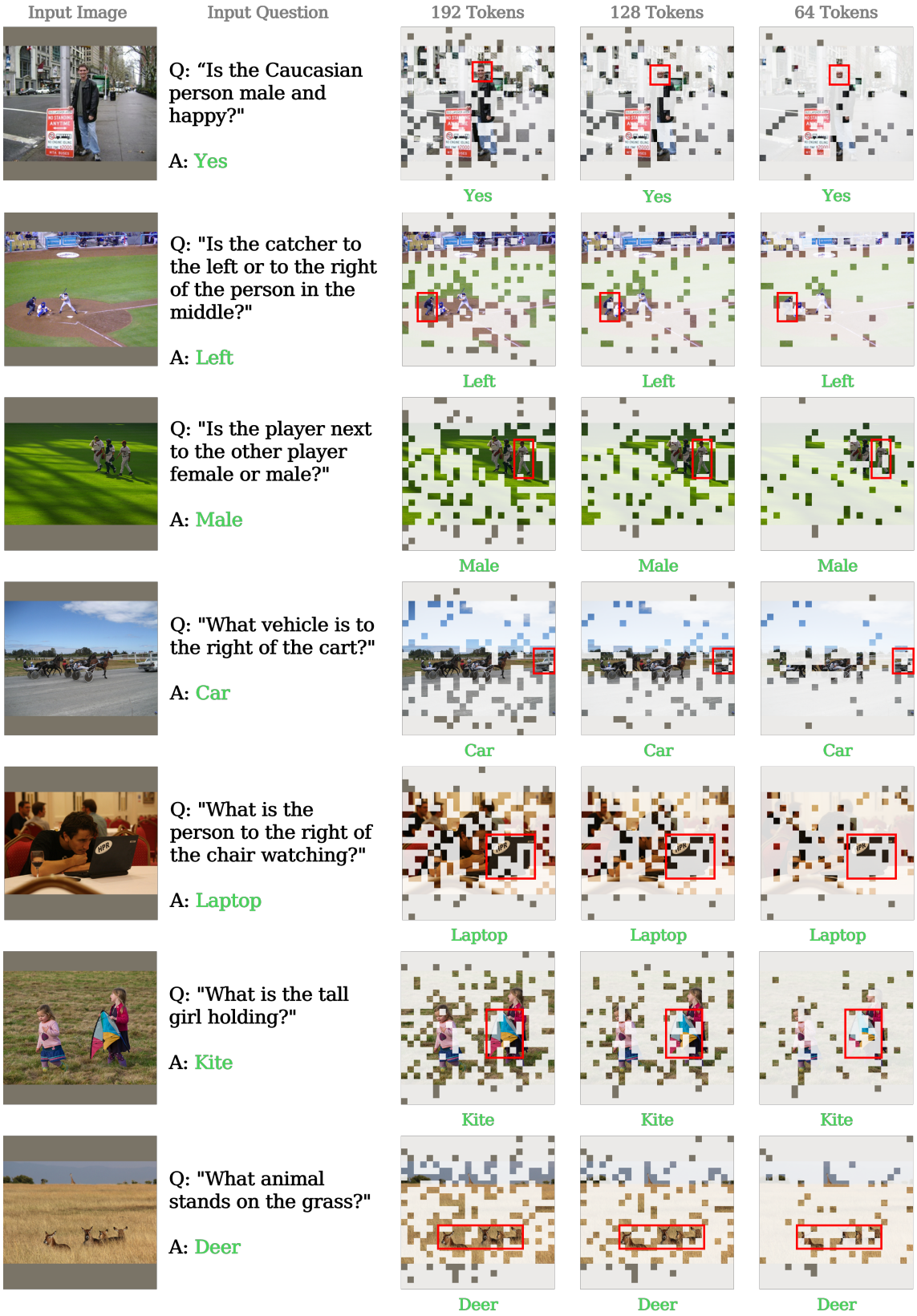


Figure F. Additional success examples demonstrating that *ZOO-Prune* preserves key visual cues across attributes, spatial relations, and object identification, enabling accurate predictions even under aggressive pruning.

References

[1] Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[3] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2024. 3

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 2

[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[6] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[7] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 2

[8] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proc. of Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. 2

[9] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[10] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024. 1

[11] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *Proc. of European Conf. on Computer Vision (ECCV)*, 2024. 2

[12] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2022. 2

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. of Int'l Conf. on Machine Learning (ICML)*, 2021. 2

[14] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[15] Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 19792–19802, 2025. 2, 3

[16] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[17] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916, 2025. 2

[18] Yunhang Shen Yulei Qin Mengdan Zhang, Xu Lin Jinrui Yang Xiawu Zheng, Ke Li Xing Sun Yunsheng Wu, Rongrong Ji Chaoyou Fu, and Peixian Chen. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2021. 3