

Appendix for "SCOPE: Saliency-Coverage Oriented Token Pruning for Efficient Multimodal LLMs"

A Overview

This appendix provides detailed information on the experimental benchmarks, additional qualitative results, and visualizations that support the main claims of the paper. In Section B, we present comprehensive descriptions of the benchmark datasets. Section C includes supplementary experiments, such as results on larger models (LLaVA 1.5 13B and LLaVA-Next 13B), as well as a hyperparameter analysis. In Section D, we provide additional visualization studies to further illustrate the behavior of our method. Finally, we discuss the broader impact and limitations of our work.

B Benchmarks

We conduct the experiments on several widely used visual understanding benchmarks. In the following, we will give a detailed description of these benchmarks.

GQA. [2]. The GQA benchmark consists of three components: scene graphs, questions, and images. The image component includes raw images, their spatial features, and the features of all objects within the images. The questions in GQA are crafted to evaluate visual scene understanding and reasoning about various aspects of an image. Our method is evaluated on the subset of "testdev_balanced_instructions", which includes 12,578 samples.

MMBench. [5]. MMBench is a comprehensive benchmark designed to evaluate the multi-modal capabilities of large language models, covering a wide range of tasks including visual question answering, image captioning, cross-modal retrieval, and creative generation. It provides a fine-grained assessment from perception to cognition, containing approximately 3,000 multiple-choice questions aggregated from diverse sources. The benchmark aims to measure whether a model is a true "all-around player" in multi-modal understanding and reasoning.

MME. [1]. The MME benchmark is a comprehensive evaluation suite carefully crafted to assess multiple facets of model performance. It comprises 14 distinct subtasks targeting both perceptual and cognitive capabilities of models. By employing manually curated instruction-answer pairs and succinct instruction formats, MME effectively reduces the risks of data leakage and ensures a fairer assessment of model abilities. We evaluate the performance on the dev split including 4,377 samples. The evaluation metric is the accuracy of the model's answer.

POPE. [4] The POPE benchmark focuses on assessing object hallucination in models by presenting them with a set of targeted yes/no questions about object existence within images. This approach reframes the evaluation of hallucination, emphasizing the model's ability to correctly identify whether certain objects are present. To quantitatively analyze performance across three distinct sampling methods, the benchmark utilizes metrics such as accuracy, recall, precision, and F1 score, offering a robust measure of the model's susceptibility to hallucination. We evaluate the model's performance on the test split, including 9,000 samples. The evaluation metric is the F1 score.

ScienceQA (SQA). [6] Encompassing a wide array of fields such as natural sciences, linguistics, and social sciences, SQA structures its questions through a hierarchical framework consisting of 26 topics, 127 categories, and 379 distinct skills. This benchmark is designed to rigorously test a model's proficiency in multimodal comprehension, complex reasoning across multiple steps, and interpretability. By organizing questions first by subject area, then by specific category, and finally by the required skill, SQA ensures a thorough and nuanced assessment of scientific understanding across diverse domains. This layered organization enables a detailed evaluation of a model's ability to handle a broad spectrum of scientific queries. The evaluation metric is the accuracy.

TextVQA. [7] TextVQA is designed to assess a model's capability to interpret and reason over textual content embedded in images. This benchmark challenges models with visual question answering tasks that require both comprehension of image context and accurate reading of the text present within the images. To perform well, models must effectively integrate visual and textual cues, demonstrating robust understanding and reasoning skills related to text in complex visual environments. We evaluate

Table 1: **Performance comparison under different vision token configurations.** The evaluated model is LLaVA 1.5 13B, where the default number of visual tokens is 576. The first row for each method reports the raw accuracy across benchmarks, and the second row indicates the performance relative to the upper bound.

Method	GQA	MMB	MME	POPE	SQA	TextVQA	SEED-I	MMVet	Avg.
Upper Bound, 576 Tokens (100%)									
Vanilla (CVPR’24)	63.2 100%	67.7 100%	1818 100%	85.9 100%	72.8 100%	61.3 100%	66.9 100%	35.3 100%	100%
Retain 192 Tokens (↓ 66.7%)									
VisionZip (CVPR’25)	59.1 93.5%	66.9 98.8%	1754 96.5%	85.1 99.1%	73.5 101.0%	59.5 97.1%	65.2 97.5%	37.5 106.20%	98.7%
Ours	59.7 94.5%	67.6 99.9%	1775 97.6%	86.7 100.9%	73.8 101.4%	60 97.9%	65.5 97.9%	39.4 111.6%	100.2%
Retain 128 Tokens (↓ 77.8%)									
VisionZip (CVPR’25)	57.9 91.6%	66.7 98.5%	1743 95.9%	85.2 99.2%	74 101.6%	58.7 95.8%	63.8 95.4%	37.5 106.2%	97.0%
Ours	59.3 93.8%	67.2 99.3%	1735 95.4%	85.9 100.0%	73.9 101.5%	58.7 95.8%	64.8 96.9%	37.7 106.8%	98.7%
Retain 64 Tokens (↓ 88.9%)									
VisionZip (CVPR’25)	56.2 88.9%	64.9 95.9%	1676 92.2%	76.0 88.5%	74.4 102.2%	57.4 93.3%	60.4 90.3%	33.9 96.0%	93.7%
Ours	58.7 92.9%	65.5 96.8%	1762 96.9%	83.0 96.6%	73.2 100.5%	58.3 95.1%	63.6 95.1%	35.7 101.1%	96.9%

the model’s performance on the test split, including 5,000 samples. The evaluation metric is exact match (EM).

SEEDBench [3] SEEDBench features a collection of 19,000 multiple-choice questions curated by human annotators. Covering 12 different evaluation dimensions, this benchmark examines models’ capabilities in identifying patterns within both images and videos, taking into account spatial as well as temporal characteristics. The evaluation metric is the accuracy.

MMVet [9] The MMVet benchmark is constructed with the understanding that tackling complex tasks typically requires a generalist model to effectively combine multiple fundamental vision-language skills. MMVet identifies six essential vision-language capabilities and systematically evaluates sixteen specific combinations arising from these core abilities, thereby assessing the model’s proficiency in integrating diverse vision-language functions. We evaluate the model’s performance on the test split, including 218 samples. The score is evaluated by the GPT model.

C Additional Experiments

In the main paper, we present experiments on LLaVA 1.5 7B and LLaVA-Next 7B. To further demonstrate the generalizability of our method across model scales, we provide additional results on LLaVA 1.5 13B and LLaVA-Next 13B in this appendix.

C.1 Results on LLaVA 1.5 13B

As shown in Table 1, our method consistently outperforms VisionZip [8] across all token budgets. With 192 tokens, our approach achieves 100.2% of the upper bound’s average performance, slightly higher than VisionZip [8] (98.7%). The advantage becomes more evident as the token count decreases: at 64 tokens, our method retains 96.9% performance, compared to VisionZip’s 93.7%. Notably, on benchmarks like MMVet [9] and POPE [4], our method even surpasses the original model’s performance. These results demonstrate that our joint saliency-coverage strategy better preserves essential information under aggressive token pruning.

Table 2: **Performance comparison under different vision token configurations.** The evaluated model is LLaVA-Next 13B. The vanilla number of vision tokens is 2,880. The first line of each method is the raw accuracy on the benchmarks, and the second line is the proportion relative to the upper bound.

Method	GQA	MMB	MME	POPE	SQA	TextVQA	MMMU	SEED-I	Avg.
Upper Bound, 2880 Tokens (100%)									
Vanilla 13B (CVPR’24)	65.4 100%	70.0 100%	1901 100%	86.2 100%	73.5 100%	64.3 100%	36.2 100%	71.9 100%	100%
Vanilla 7B (CVPR’24)	64.2 98.2%	67.9 97.0%	1842 96.9%	86.4 100.2%	70.2 95.5%	61.3 95.3%	35.1 97.0%	70.2 97.6%	97.2%
Retain 640 Tokens (\downarrow 77.8%)									
VisionZip (CVPR’25)	63.0 96.3%	68.6 98.0%	1871 98.4%	85.7 99.4%	71.2 96.9%	62.2 96.7%	36.4 100.6%	68.8 95.7%	97.8%
Ours	63.7 97.4%	69.2 98.9%	1875 98.6%	86.7 100.6%	72.1 98.1%	62.2 96.7%	36.4 100.6%	70 97.4%	98.5%
Retain 320 Tokens (\downarrow 88.9%)									
VisionZip (CVPR’25)	60.7 92.8%	67.2 96.0%	1805 95.0%	82.0 95.1%	70.3 95.6%	60.9 94.7%	35.6 98.3%	65.2 90.7%	94.8%
Ours	62.7 95.9%	67.9 97.0%	1859 97.8%	85.1 98.7%	70.8 96.3%	61.1 95.0%	36 99.4%	68.2 94.9%	96.9%
Retain 160 Tokens (\downarrow 94.4%)									
VisionZip (CVPR’25)	57.8 88.4%	64.9 92.7%	1739 91.5%	76.6 88.9%	69.3 94.3%	58.4 90.8%	37.0 102.2%	61.1 85.0%	91.7%
Ours	61.4 93.9%	66.8 95.4%	1777 93.5%	82.7 95.9%	71.5 97.3%	59.4 92.4%	35.6 98.3%	65.8 91.5%	94.8%

74 C.2 Results on LLaVA-Next 13B

75 We present the results on LLaVA-Next 13B in Table 2. We can observe that our method consistently
76 outperforms VisionZip [8] under all token budgets. For example, with 640 tokens, our approach
77 achieves 98.5% of the upper bound’s average performance, compared to VisionZip’s 97.8%. As the
78 token count decreases to 160, our method still retains 94.8% performance, while VisionZip drops to
79 91.7%. These results further confirm the superior robustness of our method under aggressive token
80 pruning.

81 C.3 Hyper-parameter Analysis

82 The hyperparameter α controls the scaling of the attention scores, thereby influencing token selection
83 in our method. As illustrated in Fig. 1, the optimal performance is typically achieved when $\alpha = 1.0$,
suggesting that this setting effectively balances saliency and coverage across most benchmarks.

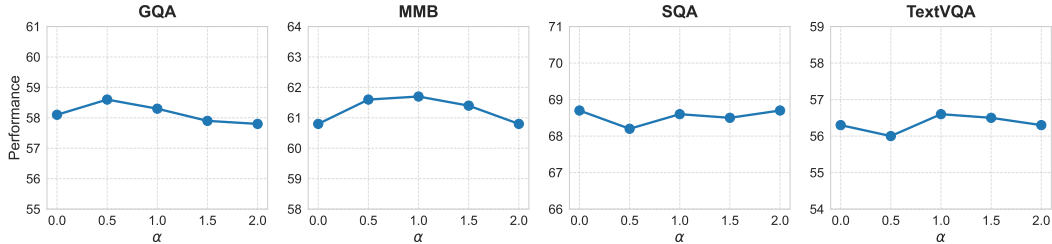


Figure 1: The hyperparameter α analysis on LLaVA 1.5 7B with 64 visual tokens.



Figure 2: The selected token comparison between the saliency-based method and our saliency-coverage oriented method. The total visual token number is 576, and the selected token number is 64.

85 D Visualization Results

86 We present additional results on selected token visualization in Fig. 2. The saliency-based method
 87 selects tokens solely based on attention scores, which may overlook semantically important tokens
 88 that contribute to the overall completeness of the visual representation. In contrast, our saliency-
 89 coverage oriented approach jointly considers both visual saliency and semantic coverage. As a result,
 90 the selected tokens span a broader region in the embedding space.

91 In Fig. 3, we further visualize the attention distribution of selected tokens. Our method preserves the
 92 majority of high-attention tokens, demonstrating its ability to retain both informative and representa-
 93 tive visual content.

94 E Broader Impact

95 Our proposed method aims to improve both the efficiency and effectiveness of multimodal large
 96 language models (MLLMs) by reducing the number of visual tokens while preserving semantic
 97 completeness. This advancement has the potential to significantly reduce the computational cost
 98 and memory footprint of MLLMs, thereby enhancing their feasibility for deployment in resource-
 99 constrained environments such as edge devices, mobile platforms, and real-time applications. By
 100 enabling more efficient inference, our approach can facilitate the broader adoption of vision-language
 101 models across various domains, including education, healthcare, and assistive technologies.

102 However, as with any technology that enhances the scalability and accessibility of AI systems, there
 103 are potential societal risks. For example, more efficient MLLMs could be misused to generate
 104 or disseminate misinformation, enable invasive surveillance, or support other malicious activities,

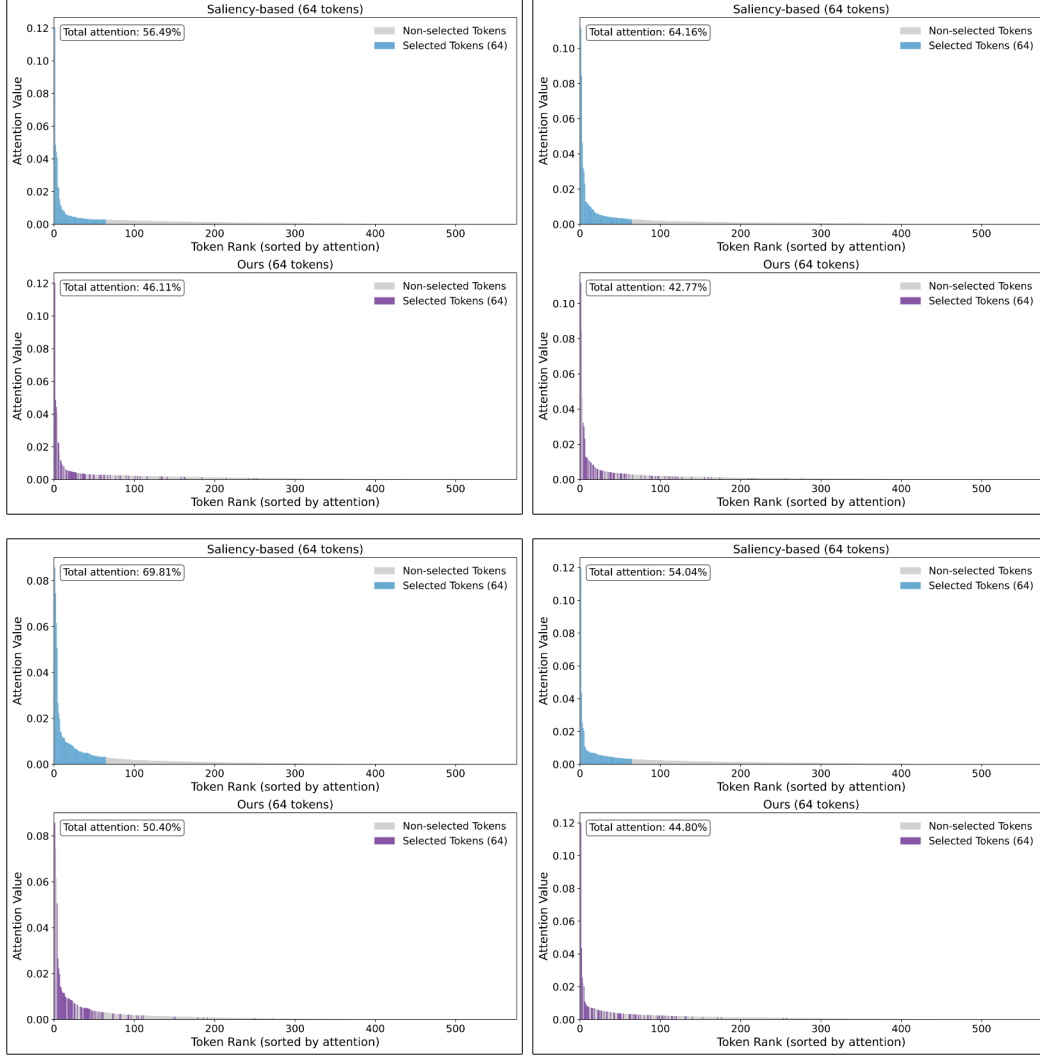


Figure 3: Attention distribution visualization for selected token. The total visual token number is 576, and the selected token number is 64. Our method retained most of the high attention tokens and some low attention tokens to maximize the coverage.

105 particularly when deployed at scale. It is therefore essential to consider these ethical implications and
 106 implement appropriate safeguards when deploying such models in practice.

107 F Limitations

108 While SCOPE demonstrates strong performance and efficiency gains across multiple benchmarks
 109 and model architectures, several limitations remain. (1) Despite our efforts to balance saliency
 110 and coverage, aggressive token pruning may still result in the loss of fine-grained or rare semantic
 111 information, potentially affecting tasks that require detailed visual understanding. (2) Our experiments
 112 are primarily based on widely used vision-language benchmarks and two representative MLLMs,
 113 LLaVA 1.5 and LLaVA-Next. Therefore, the generalizability of SCOPE to other tasks or model
 114 architectures has yet to be fully validated.

References

- [1] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv:2306.13394*, 2023.
- [2] Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [3] 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.
- [4] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv:2305.10355*, 2023.
- [5] 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? *arXiv:2307.06281*, 2023.
- [6] 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. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.
- [7] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.
- [8] 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. *arXiv preprint arXiv:2412.04467*, 2024.
- [9] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *International conference on machine learning*. PMLR, 2024.