Reimplementation and Fine-Grained Analysis of Visual Token Sparsification Methods

Anonymous ACL submission

Abstract

Visual token sparsification represents a critical optimization technique for reducing computational overhead in Vision-Language Models (VLMs) while maintaining performance. This work presents a comprehensive reimplementation and fine-grained analysis of two state-of-the-art methods: VisionZip, a text-agnostic approach utilizing CLIP attention weights, and SparseVLM, a text-aware method employing cross-modal attention. Through systematic evaluation on the POPE dataset with 6,823 samples, we demonstrate that both methods achieve near-equivalent performance (Accuracy: 80.3% vs 80.6%) at 64-token sparsification (88.9% reduction), despite fundamentally different architectural designs. Our analysis reveals that text-aware token selection provides marginal advantages in precision (0.7%) but fails to meaningfully improve recall, suggesting that for object existence tasks, question-adaptive selection may not provide substantial benefits over question-agnostic approaches.

1 Introduction

Modern Vision-Language Models (VLMs) typically process 576 visual tokens per image, representing a significant computational bottleneck. Visual token sparsification methods aim to reduce this number while preserving model performance, with recent approaches demonstrating that 88.9% token reduction (from 576 to 64 tokens) can be achieved with minimal performance degradation. However, the optimal strategy for token selection remains an open question: should tokens be selected based on intrinsic visual importance independent of the query, or should selection adapt to the specific question being asked?

This work addresses this question through a rigorous comparative analysis of two representative approaches: VisionZip (1), which employs text-agnostic token selection based on CLIP attention patterns, and SparseVLM (3), which utilizes text-aware selection through cross-modal attention. By integrating, debugging, and modifying the official implementations of both methods, we successfully configured them for evaluation under identical experimental conditions. This process allowed us to provide insights into the mechanisms, trade-offs, and limitations of each approach.

Our contributions are threefold: (1) a successful replication of the methods, paying careful attention to implementation details and necessary modifications for compatibility, (2) a comprehensive evaluation on the POPE dataset with 6,823 samples, reporting multiple metrics, and (3) a fine-grained analysis of token selection mechanisms, performance characteristics, and failure modes.

2 Method Overview

2.1 VisionZip: Text-Agnostic Token Selection

VisionZip implements a two-stage token selection mechanism that operates entirely within the vision encoder, independent of the language component. The method is grounded in the observation that CLS token attention weights in deep layers of the CLIP encoder capture global visual saliency.

VisionZip operates in two stages as described in the original paper (1). In Stage 1, the method selects 54 dominant tokens by computing CLS token attention weights at layer 22 of the CLIP encoder and selecting the top-k tokens with highest attention scores. The CLS token attention naturally captures global visual saliency, prioritizing prominent objects and high-contrast regions.

In Stage 2, VisionZip selects 10 contextual tokens from the remaining tokens through similarity-based clustering. The method normalizes the remaining tokens and samples target tokens uniformly, then assigns other tokens to the nearest target based on cosine similarity. This ensures diverse spatial coverage, capturing background con-

text and less prominent visual regions that may be missed by the dominant selection.

2.2 SparseVLM: Text-Aware Progressive Token Selection

SparseVLM implements a fundamentally different approach: token selection adapts dynamically to the input question by leveraging cross-modal attention between visual and text tokens. The method applies progressive sparsification at multiple decoder layers, allowing the model to gradually refine token selection based on question-specific requirements.

SparseVLM computes cross-modal attention between visual tokens and text tokens at decoder layers 2, 6, and 15. The method uses standard scaled dot-product attention (2) to compute attention weights between text token queries and visual token keys. The relevance of each visual token to the question is determined by averaging attention scores from all text tokens to that visual token, effectively measuring how much the question "attends" to each visual region.

The method applies progressive sparsification: for 64 total tokens, it retains 66 tokens at layer 2, 30 tokens at layer 6, and 17 tokens at layer 15. At each layer, tokens with highest text-visual relevance scores are selected via top-k selection, and remaining tokens are merged into selected tokens using attention-weighted pooling. This progressive approach allows the model to maintain rich visual information in early layers while gradually focusing on question-relevant tokens.

2.3 Key Differences

The fundamental difference between VisionZip and SparseVLM lies in their token selection basis. VisionZip is text-agnostic: it selects tokens based purely on visual features (CLS attention weights) without considering the input question. This means the same set of tokens is selected for an image regardless of what question is asked. SparseVLM is text-aware: it adapts token selection to each specific question by using cross-modal attention between visual and text tokens. This allows the method to focus on different image regions depending on what is being asked, potentially providing advantages for questions requiring attention to less prominent visual regions.

3 Experimental Setup

3.1 Dataset: POPE

The Polling-based Object Probing Evaluation (POPE) dataset (4) is specifically designed to evaluate object hallucination in VLMs. The dataset consists of yes/no questions about object existence, where positive cases refer to objects that exist in the image, and negative cases refer to objects that do not exist.

The dataset contains 6,823 total questions divided into three categories: Adversarial (2,274 samples), Popular (2,274 samples), and Random (2,275 samples). Images are sourced from the COCO val2014 dataset. The LLaVA model zoo reports that `Vanilla LLaVA-1.5-7B` achieves 85.9% accuracy on POPE with full token retention (576 tokens), which we use as a reference baseline (not re-evaluated in this work).

3.2 Model Configuration

We use `LLaVA-1.5-7B` (5) (`liuhaotian/llava-v1.5-7b`) as the base model for both methods. For VisionZip, we configure dominant tokens as 54 and contextual tokens as 10, totaling 64 tokens as specified in the paper (1). Token selection occurs at `CLIP encoder layer 22`. For SparseVLM, we configure retained tokens as 64 total, with progressive sparsification applying token counts of `[66, 30, 17]` at decoder layers `[2, 6, 15]` respectively, matching the method's design.

3.3 Hardware and Software Environment

All experiments were conducted on the CS department's instructional GPU cluster, specifically on the `instgpu-01.cs.wisc.edu` machine. This server is equipped with eight NVIDIA GeForce RTX 2080 Ti GPUs, each with approximately 11 GB of VRAM. Our experiments were primarily executed on GPUs 6 and 7 of this shared machine.

The software environment was built on a Linux operating system with CUDA version 12.4 and NVIDIA driver version 550.67. We used Python 3.10 with `PyTorch` 2.1.2. Key libraries included `transformers`, `bitsandbytes` for 4-bit quantization attempts, and `accelerate` for model handling. Due to the 11 GB memory limitation of the RTX 2080 Ti GPUs when loading the 7B parameter `LLaVA` model (which requires 14 GB in float16), we utilized the `device_map="auto"` feature from the `accelerate` library. This configuration auto-

matically offloaded some model layers to the CPU to prevent out-of-memory errors, a necessary deviation for running these large models on the available hardware.

3.4 Evaluation Metrics

We report four standard classification metrics: Accuracy, Precision, Recall, and F1 Score. Metrics are calculated separately for each category (adversarial, popular, random) and then averaged to obtain overall performance. Since the VisionZip paper reports accuracy as the primary metric for POPE, we use accuracy as our primary comparison metric to ensure fair comparison with published results.

4 Results

4.1 Overall Performance

Table 1 presents comprehensive results on the POPE dataset.

4.2 Comparison with Published Results

The VisionZip paper reports "raw benchmark accuracy" for the POPE dataset at a 64-token configuration (1). Their findings show VisionZip achieving 77.0% accuracy and SparseVLM achieving 75.1% accuracy. This establishes a performance hierarchy where VisionZip outperforms SparseVLM by 1.9 percentage points.

Our reimplementation yields different results in two key aspects. First, our measured accuracies (VisionZip: 80.3%, SparseVLM: 80.6%) are notably higher than those reported in the paper. This discrepancy may stem from differences in the evaluation protocol, minor implementation details, or the specific hardware environment. Second, the performance ranking is reversed: in our experiments, SparseVLM surpasses VisionZip by a margin of 0.3 percentage points. While this margin is minimal and likely within the bounds of experimental variance, it contradicts the paper's conclusion. The key takeaway is that our reimplementation finds the two methods to have virtually equivalent performance on this task.

5 Fine-Grained Analysis

5.1 Token Selection Mechanism Analysis

VisionZip's Design:

VisionZip's two-stage selection mechanism is designed to balance global saliency with spatial diversity. The method uses CLS token attention weights at layer 22 of the CLIP encoder, operating on the principle that deep-layer CLS attention aggregates global visual information. By design, the dominant token selection (54 tokens) should capture visually salient regions—typically objects, high-contrast areas, and text—while the contextual token selection (10 tokens) through similarity-based clustering should provide spatial coverage of less prominent regions.

A key characteristic of this approach is that token selection is question-independent: the same 64 tokens are selected for an image regardless of what is being asked. This design choice prioritizes efficiency but theoretically limits the method's ability to adapt when questions focus on background or non-salient objects.

SparseVLM's Design:

SparseVLM employs a fundamentally different strategy: progressive, question-aware token selection across multiple decoder layers. The method computes cross-modal attention between text and visual tokens, allowing token selection to adapt dynamically based on the specific question. The progressive schedule (66 tokens at layer 2, 30 at layer 6, 17 at layer 15) is designed to maintain rich visual information early in processing while gradually focusing on question-relevant regions.

Theoretically, this text-aware approach should provide advantages when questions require attention to specific regions that differ from global visual saliency patterns. However, our empirical results show that both methods achieve nearly identical performance (80.3% vs 80.6% accuracy), suggesting that for POPE's object existence questions—which primarily query about prominent, visually salient objects—the adaptability of text-aware selection does not translate to measurable performance gains.

5.2 Precision-Recall Trade-off Analysis

Both methods exhibit a **high precision, moderate recall** profile, indicating conservative prediction strategies. This is particularly important for hallucination detection, where false positives are costly.

Precision Analysis: VisionZip achieves 94.5% precision, while SparseVLM achieves 95.2% (0.7% higher). The text-aware selection provides more relevant visual tokens, leading to slightly higher confidence in predictions and reducing false positives.

Recall Analysis: Both methods achieve nearly identical recall (64.9% vs 65.0%), indicating that at

Method	Accuracy	Precision	Recall	F1 Score
VisionZip	0.803 (80.3%)	0.945 (94.5%)	0.649 (64.9%)	0.769 (76.9%)
SparseVLM	0.806 (80.6%)	0.952 (95.2%)	0.650 (65.0%)	0.772 (77.2%)
Difference	-0.003 (-0.3%)	-0.007 (-0.7%)	-0.001 (-0.1%)	-0.003 (-0.3%)

Table 1: Performance comparison on POPE dataset. Both methods achieve near-equivalent performance with minimal differences across all metrics.

64 tokens (88.9% reduction), both struggle equally with detecting all relevant objects. This suggests that 64 tokens may be insufficient to capture all object information needed for high recall, and neither selection strategy can overcome the fundamental information loss from aggressive sparsification.

5.3 Theoretical Limitations and Failure Modes

Based on the architectural designs and our quantitative results, we can identify theoretical limitations for each method.

VisionZip's Potential Limitations: The question-independent selection means that when POPE asks about objects in background regions, the CLS attention-based selection may prioritize foreground tokens over the queried object. Questions about small or non-salient objects may fail if those objects are not among the 64 selected tokens. The high precision (94.5%) but moderate recall (64.9%) we measured is consistent with this conservative selection strategy.

SparseVLM's Potential Limitations: The progressive reduction schedule (66→30→17 tokens) means that if critical visual information is not captured at layer 2, later layers cannot recover it. The method's reliance on cross-modal attention means that poor question representations could lead to suboptimal token selection. Despite the theoretical advantage of question-aware selection, SparseVLM achieves only marginally higher precision (95.2%) and nearly identical recall (65.0%) compared to VisionZip, suggesting these theoretical advantages may not materialize for simple object existence queries.

5.4 Theoretical Implications

Our experimental findings suggest that for object existence tasks like POPE, which primarily focus on prominent objects, text-aware token selection does not provide substantial advantages over text-agnostic selection. This is because POPE questions typically query about visually salient objects that are naturally captured by both selection strategies. However, we hypothesize that for more complex tasks requiring spatial reasoning, attribute understanding, or multi-object relationships, text-aware selection may provide advantages by allowing the model to focus on different image regions depending on the question.

At 64 tokens (88.9% reduction from 576), both methods achieve approximately 80% accuracy, compared to the baseline performance. The absolute performance drop suggests that 64 tokens capture most but not all critical information needed for accurate object existence detection. Further token reduction would likely degrade performance more significantly, suggesting that 64 tokens may be near the lower bound for maintaining reasonable performance on this task.

6 Limitations

This study has several limitations: (1) evaluation on POPE only—results may not generalize to other benchmarks, (2) single sparsification level—only 64 tokens evaluated, (3) accuracy values higher than paper—our accuracy (80.3%, 80.6%) exceeds paper's (77.0, 75.1%), suggesting potential differences in evaluation protocol, (4) implementation variance—subtle differences may affect results, and (5) hardware constraints—the 7B LLaVA model requires approximately 14GB in float16, exceeding the 11GB capacity of our RTX 2080 Ti GPUs. Both methods used `device_map="auto"` which likely offloaded some model layers to CPU. This memory constraint may have affected absolute performance, though both methods were evaluated under identical conditions.

7 Conclusion

This work presents a comprehensive reimplementation and fine-grained analysis of two visual token sparsification methods: VisionZip (text-agnostic) and SparseVLM (text-aware). Through rigorous evaluation on the POPE dataset with 6,823 samples, we demonstrate that both methods achieve near-equivalent performance (Accuracy: 80.3% vs 80.6%) at 64-token sparsification, despite fundamentally different architectural designs.

Our analysis reveals that for object existence tasks, text-agnostic and text-aware selection achieve essentially identical performance, suggesting that question-adaptive selection may not provide substantial benefits for simple binary questions about prominent objects. However, SparseVLM's text-aware selection provides marginal precision advantage (0.7%), indicating better alignment between visual tokens and question requirements. Both methods struggle equally with recall at high sparsification levels, suggesting that 64 tokens may be insufficient regardless of selection strategy.

These findings suggest that for object existence tasks, the choice between text-agnostic and text-aware selection may not significantly impact performance. However, for more diverse question types requiring attention to different image regions, text-aware selection may provide advantages. Future work should evaluate both methods on more diverse benchmarks to test this hypothesis.

Acknowledgments

We thank the authors of VisionZip and SparseVLM for making their code publicly available, enabling this reimplementation study.

References

[1] Yang, S., Chen, Y., Tian, Z., Wang, C., Li, J., Yu, B., & Jia, J. (2024). VisionZip: Longer is Better but Not Necessary in Vision Language Models. *arXiv preprint arXiv:2412.04467*.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*.

[3] Zhang, Y., et al. (2024). SparseVLM: Visual Token Sparsification for Efficient Vision-Language Model Inference. *arXiv preprint arXiv:2410.04417*.

[4] Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J. R. (2023). Evaluating Object Hallucination in Large Vision-Language Models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[5] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual Instruction Tuning. *Advances in Neural Information Processing Systems (NeurIPS)*.