SparseZip: Text-Aware Visual Token Selection and Compression for Efficient Vision-Language Model Inference

Alexander Geppert
University of Wisconsin-Madison
ageppert@wisc.edu

Qinxinghao Chen
University of Wisconsin-Madison
qchen463@wisc.edu

Shivam Mittal
University of Wisconsin-Madison
smittal39@wisc.edu

Najib Akram Maheen Aboobacker
University of Wisconsin-Madison
maheenabooba@wisc.edu

Abstract

We propse SparseZip, a Text-Specific Visual Token Selection and Compression framework that (i) predicts a visual token retention budget based on inherent image complexity, (ii) ranks tokens using a score based on both visual token self-attention and text token cross-attention, and (iii) condenses the top tokens chosen by the attention scores and budget, turning high-resolution visual context into a small set of semantically dense tokens.

1 Introduction

Visual tokens often suffer large computational overhead due to a sparsity of information when compared with text tokens, which causes signficant latnecy during inference time for Vision Language Models (VLMs). Existing methods, including FastV, SparseVLM, and VisionZip, attempt to mitigate by (i) compressing visual tokens, (ii) selecting the most relevant parts of an image via, e.g., pruning, or (iii) restructuring cross-attention to reduce redundant vision-text interactions. However, what remains underexplored is text-specific visual token selection at inference time in addition to compression.

To explore this, our work situates itself at this intersection: leveraging CLIP-style aligned encoders to obtain reliable token-level saliency signals while incorporating both SparseVLM's learned selective retention policies and embracing VisionZip's technique of compressing visual evidence into succinct, information-rich tokens. We propse SparseZip, a Text-Specific Visual Token Selection and Compression framework that (i) predicts a visual token retention budget based on inherent image complexity, (ii) ranks tokens using a score based on both visual token self-attention and text token cross-attention, and (iii) condenses the top tokens chosen by the attention scores and budget, turning high-resolution visual context into a small set of semantically dense tokens. By utilizing these techniques, our approach aims to close the gap between efficiency-first pruning and accuracy-first long-context encoding, delivering consistent latency savings where possible and preserving fidelity where necessary.

2 Related Works

CLIP ViT/ResNet Image Encoder. Early vision-language systems established the now standard recipe of aligning images and text through contrastive pretraining on (image, text) pairs collected from the internet. CLIP paired a ViT/ResNet image encoder with a Transformer text encoder and optimized a bidirectional InfoNCE objective to bring matched (image, text) embedding pairs together and push mismatches apart (Radford et al., 2021). CLIP's key contribution was demonstrating that large, weakly supervised corpora suffice to learn universal visual concepts that transfer to downstream tasks with zero-shot prompts, effectively reframing recognition as prompt engineering rather than supervised fine-tuning. This setup provided the modern VLM stack: frozen or lightly tuned vision encoders that feed compact visual tokens or pooled embeddings into language models.

Designing Training-Free Methods for Reducing VLM Inference Latency As VLMs evolved from pure alignment to instruction-following pipelines, a central bottleneck emerged: the quantity and placement of visual tokens that enter the Language Model. Naively feeding all patch tokens from a ViT to the LLM incurs high attention costs and latency mismatches between the vision encoder and the autoregressive decoder. This sparked a wave of efficiency-first methods that retain accuracy while reducing token count, cross-modal attention, or both. Although designs differ, these systems typically (i) compress visual tokens, (ii) select the most relevant parts of an image via, e.g., pruning, or (iii) restructure cross-attention to re-

duce redundant vision-text interactions. The shared goal is to shorten or restructure the visual evidence passed downstream without sacrificing task performance.

Pruning Vision Tokens with Small Attention Scores during Forward Pass of LLM. FastV dynamically prunes less-important visual tokens in deeper layers of the LLM (based on attention scores) to accelerate inference with minimal accuracy loss (Chen et al., 2024). While FastV utilizes both self-attention between visual tokens and cross-attention with text tokens, it does so after some number of layers, meaning there are a significant number of attention calculations computed before redundant visual tokens are pruned.

Pruning Vision Tokens with Small Cross-Attention Prior to LLM. A complementary line of research focuses on visual token sparsity as a means of improving efficiency before feeding image representations into the language model. SparseVLM introduces a training-free mechanism that prunes redundant visual tokens by leveraging text-guided attention signals rather than dense all-to-all interactions among visual tokens. This approach selectively routes computation toward informative, text-relevant regions while maintaining global pathways for long-range dependencies (Zhang et al., 2025). SparseVLM's text-conditioned sparsification offers a lightweight, plug-and-play alternative that preserves task performance without additional training, complementing architectural sparsity methods that jointly optimize selection policies with task objectives to enable efficient compute during inference.

Pruning Vision Tokens with Small Self-Attention Prior to LLM. VisionZip targets redundancy in visual representations before they are passed to the language model. Unlike FastV, which prunes tokens within the LLM, VisionZip operates immediately after the vision encoder by analyzing self-attention among visual tokens to estimate their importance (Yang et al., 2025). Tokens with low self-attention significance are discarded or merged into a smaller set of contextual tokens, producing a compact yet information-preserving representation. The compression techniques used in VisionZip achieve substantial speedups (e.g., up to 8x faster prefilling) with minimal performance degradation, demonstrating that longer visual sequences are beneficial but not strictly necessary for strong VLM performance. However, while VisionZip outperforms FastV and SparseVLM, cross-attention is not used. Therefore, the selected visual tokens are text-agnostic.

3 Proposed Methodology

To address the text-agnostic issue while maintaining the same performance, our method utilizes two main innovations:

Hybrid Attention VisionZip determines the dominant visual tokens based only on self-attention. SparseVLM uses cross-attention, but prunes tokens while proceeding with a forward pass through the LLM, leading to many unnecessary FLOPs. To mitigate the text-agnostic problem while maintaining performance, we adapt the VisionZip dominant visual token selection technique with benefits of the SparseVLM technique. We propose calculating the sum of the self-attention and cross-attention scores for determining the dominant visual tokens before compression. Since some information will be lost during compression, this ensures that the user-provided task, which is inheriently embedded in the text tokens, influences our choice of dominant visual tokens.

Dynamic K Selection A dynamic K selection algorithm will adaptively determine how many tokens to retain based on image complexity:

$$K = log(Var(S_d)) + c \tag{1}$$

Where S_d is the scoring funtion used (e.g., attention) for all visual tokens. This ensures flexible compression that preserves key information for both simple and complex scenes.

3.1 Data Sets

We plan to use the following data sets:

GQA A large scale visual question answering (VQA) dataset designed to test scene understanding. Features 22 millions question & answer pairs associated with an image depicting a day-to-day scene.

MMBench A bilingual dataset designed to evaluate the multi-modal understanding of VLMs. Features 3000 Multiple-Choice Questions (MCQs).

MME A dataset designed to assess Multimodal Large Language Models (MLLMs) on tasks involving perception and cognition. Features 29000 question & answer pairs.

POPE A dataset used to evaluate object hallucinations by assessing accuracy in identifying object in image. Features image & question pairs.

SQA A dataset that features 76000 human-spoken questions and corresponding textual answers.

VQA^{V2} Extension of original Visual Question Answering (VQA) dataset that includes open-ended questions about images paired with ground-truth answers.

VQA^{Test} Dataset that consists of images and questions designed to evaluate visual question answering systems.

MMMU A dataset designed to evaluate multimodal models on college-level tasks across six disciplines, including Art, Business, Science, Health, Humanities, and Technology. Features 11500 questions that require both visual and textual understanding, aiming to assess expert-level reasoning capabilities in AI models.

SEED A collection of EEG data used for emotion recognition, provided by the BCMI laboratory.

MMVet A dataset designed for evaluating large multimodal models, focusing on the integration of various capabilities like recognition, language generation, and spatial awareness.

LLaVA-Bench A benchmarking suite designed to evaluate the performance of large multimodal models (LMMs) on various tasks, including visual instruction following and reasoning. It provides datasets and metrics to assess how well these models understand and generate responses based on visual and textual inputs.

3.2 Experimental Setup

We will use the CLIP vision encoder in combination with the LLaVa 1.5 LLM. For each experiement, we will use all applicable benchmarks listed above as evaluation metrics.

3.3 Planned Experiments

Baseline Comparison Test whether our Text-Specific Visual Token Selection and Compression framework outperforms the text-agnostic baseline given by VisionZip.

Dynamic K Comparison We will compare the performance difference when selecting a fixed number of top tokens and dynamically selecting K top tokens. We will also perform an efficiency analysis regarding the dominant visual token reduction and performance drop.

Hierarchical Contextual Token Merging For non-dominant tokens, we will try to replace naive averaging with hierarchical clustering (e.g., K-means, spectral methods). At each stage, clusters are merged iteratively with attention-weighted averaging, ensuring important context is preserved before reduction. We will evaluate whether hierarchical or one-shot merging performs better.

Multi-dimensional Dominant Token Scoring Instead of relying solely on attention, we will experiment with alternative scoring functions and use a hybrid token importance score. For example:

$$s_d = \alpha_1 * s_{attn} + \alpha_2 * H_{entropy} + \alpha_3 * I_{mutual} \quad (2)$$

where s_{attn} is the mean attention score, $H_{entropy}$ measures feature variability, and I_{mutual} represents mutual information with other tokens. The scoring reflects not only the salience of each token but also its contextual contribution to the overall semantics of the image.

We will also extract token features from multiple transformer layers (e.g., L-2, L-4, L-8, L-12) and fuse them via learned gating weights.

We will perform ablation tests on all used multi-dimensional scoring components.

4 Division of Responsibilities

We plan to reimplement the frameworks from the VisionZip and SparseVLM papers. We then plan to implement our proposed methodolgy and compare the performance to the VisionZip and SparseVLM baselines using the listed benchmarks. We will continue by running the experiments listed above, updating our implementation to include the best-performing design choices. As we progress, we will be thinking of further experiments and testing further hypotheses. By the deadline, we strive to have a framework that has comperable or better efficiency as compared with VisionZip while maintaining text-awareness.

4.1 Alexander Geppert & Shivam Mittal:

Implement vanilla VisionZip and SparseVLM & evaluate (with all datasets). Implement proposed methodology & evaluate (with all datasets). Implement Hierarchical Contextual Token Merging experiement and evaluate (with all datasets).

4.2 Qinxinghao Chen & Najib Akram Maheen Aboobacker:

Implement vanilla VisionZip and SparseVLM & evaluate (with all datasets). Implement Multi-Dimensional Dominant Token Scoring experiment & evaluate (with all datasets).

References

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuang Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024. An Image is Worth 1/2 Tokens After Layer 2: Plug-and-PLay Acceleration for VLLM Inference.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. 2025. VisionZip: Longer is Better but Not Necessary in Vision Language Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 19792-19802*.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. 2025. SparseVLM: Visual Token Sparsification for Efficient Vision-Language Model Inference. *ICML 2025*.