ZipVL: Efficient Large Vision-Language Models with Dynamic Token Sparsification and KV Cache Compression

Yefei He[1,2] * **Feng Chen**[3] **Jing Liu**[4] **Wenqi Shao**[2] **Hong Zhou**[1]† **Kaipeng Zhang**[2]†

Bohan Zhuang[1,4]

[1]Zhejiang University, China
[2]Shanghai AI Laboratory, China
[3]The University of Adelaide, Australia
[4]ZIP Lab, Monash University, Australia

Abstract

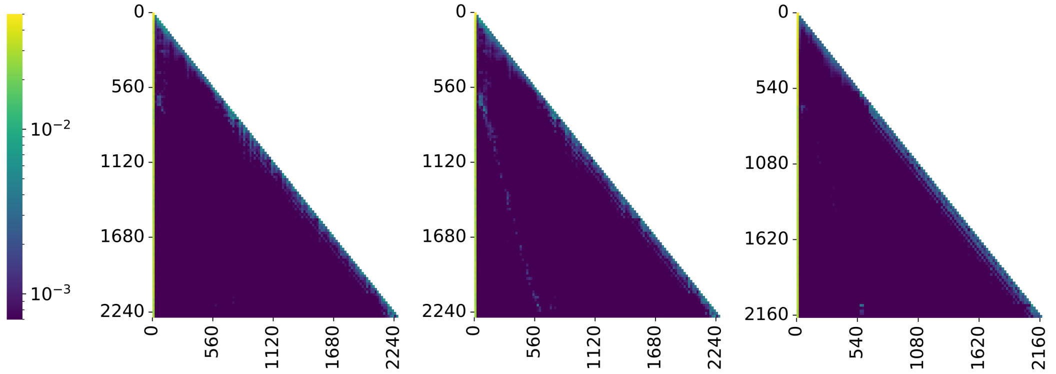
The efficiency of large vision-language models (LVLMs) is constrained by the computational bottleneck of the attention mechanism during the prefill phase and the memory bottleneck of fetching the key-value (KV) cache in the decoding phase, particularly in scenarios involving high-resolution images or videos. Visual content often exhibits substantial redundancy, resulting in highly sparse attention maps within LVLMs. This sparsity can be leveraged to accelerate attention computation or compress the KV cache through various approaches. However, most studies focus on addressing only one of these bottlenecks and do not adequately support dynamic adjustment of sparsity concerning distinct layers or tasks. In this paper, we present ZipVL, an efficient inference framework designed for LVLMs that resolves both computation and memory bottlenecks through a dynamic ratio allocation strategy of important tokens. This ratio is adaptively determined based on the layer-specific distribution of attention scores, rather than fixed hyper-parameters, thereby improving efficiency for less complex tasks while maintaining high performance for more challenging ones. Then we select important tokens based on their normalized attention scores and perform attention mechanism solely on those important tokens to accelerate the prefill phase. To mitigate the memory bottleneck in the decoding phase, we employ mixed-precision quantization to the KV cache, where high-bit quantization is used for caches of important tokens, while low-bit quantization is applied to those of less importance. Our experiments demonstrate that ZipVL can accelerate the prefill phase by 2.6× and reduce GPU memory usage by 50.0%, with a minimal accuracy reduction of only 0.2% on Video-MME benchmark over LongVA-7B model, effectively enhancing the generation efficiency of LVLMs.

1 Introduction

With the recent advancement of large language models (LLMs) (Achiam et al., 2023; Team et al., 2023; Vavekanand & Sam, 2024), many studies have extended their capabilities to comprehend and generate visual content. These models, commonly known as large vision-language models (LVLMs), have demonstrated remarkable performance in tasks such as image captioning and visual question answering (Ge et al., 2024b; Liu et al., 2024b; Team, 2024; Ge et al., 2024c; Lin et al., 2023). Typically, to remain compatible with the next-token-prediction generation scheme of LLMs, images or videos are encoded into visual tokens through a pre-trained visual encoder, and concatenated with text tokens for input into the model. For instance, LLaVA (Liu et al., 2024b) employs a pre-trained CLIP-ViT-L-336px model (Radford et al., 2021), which encodes an image of size 336×336 pixels

[*] Work done during an internship at Shanghai AI Laboratory.

[†] Corresponding authors. Email: zhouhong_zju@zju.edu.cn, kp_zhang@foxmail.com



(a) Layer 0, ChartQA (b) Layer 10, ChartQA (c) Layer 10, VQAv2

Figure 1: The attention maps exhibit distinct sparse patterns across different layers (subfigures (a) and (b)) and vary significantly between tasks (subfigures (b) and (c)). Data was collected from the LLaVA-Next-7B model using input samples from the VQAv2 and ChartQA datasets.

to 576 visual tokens. However, for high-resolution images or videos, the visual encoder generates excessive sequences of visual tokens, significantly limiting the generative efficiency of LVLMs. Specifically, the prefill phase suffers from the quadratic complexity of the attention mechanism, resulting in **computational bottleneck** and prolonged time-to-first-token (TTFT). In the decoding phase, each new token interacts with all preceding tokens, requiring to fetch the full key-value (KV) cache from memory. This process slows down decoding due to **memory bottleneck**. Improving generative efficiency in both phases is essential for the practical deployment of LVLMs.

To address computational complexity in the prefill phase, sparse attention (Pagliardini et al., 2023; Jiang et al., 2024; Zhu et al., 2024) has emerged as an effective strategy, particularly suitable for LVLMs where visual information exhibits considerable redundancy, leading to highly sparse attention maps (Wan et al., 2024; Chen et al., 2024). This sparsity can be implemented at various levels of granularity. Some studies pre-define several sparse patterns and assign them to the attention mask during inference (Jiang et al., 2024; Zhu et al., 2024). However, these predefined patterns are not compatible with efficient attention implementations such as FlashAttention (Dao et al., 2022) and require custom GPU kernels for each pattern. Alternatively, other approaches adopt token-level sparsity by identifying and discarding less important tokens (Chen et al., 2024; Arif et al., 2024), allowing seamless integration with off-the-shelf efficient attention implementations. However, the optimal retention ratio of important tokens may vary across different layers or tasks due to distinct attention patterns, as illustrated in Figure 1. These methods rely on a fixed token retention ratio and do not dynamically adjust based on task difficulty, leading to suboptimal performance on complex tasks.

To alleviate memory bottleneck, various efforts have been made to reduce KV cache size, including token dropping (Wan et al., 2024), token merging (Yang et al., 2024a), and quantization (Hooper et al., 2024; He et al., 2024b). However, these methods often rely on fixed compression ratios that are uniformly applied across all layers, failing to account for the distinct characteristics of attention maps in different layers. Moreover, despite the necessity of identifying important tokens for both sparse attention and KV cache compression, a unified inference optimization framework has yet to be developed.

In this paper, we present ZipVL, an efficient inference framework tailored for LVLMs that jointly optimizes the prefill and decoding phases with a unified ratio of important tokens, as shown in Figure 2. To start with, we introduce a layer-wise adaptive ratio assignment scheme for important tokens. This ratio is adaptively determined based on the distribution of attention scores in each layer, rather than relying on predefined hyper-parameters (Chen et al., 2024; Arif et al., 2024; He et al., 2024b; Zhang et al., 2023). This adaptive approach allows the ratio to be adjusted according to task complexity, enhancing efficiency for simpler tasks while preserving performance for more complex ones. After determining the ratio, we then select important tokens with the highest normalized attention scores,