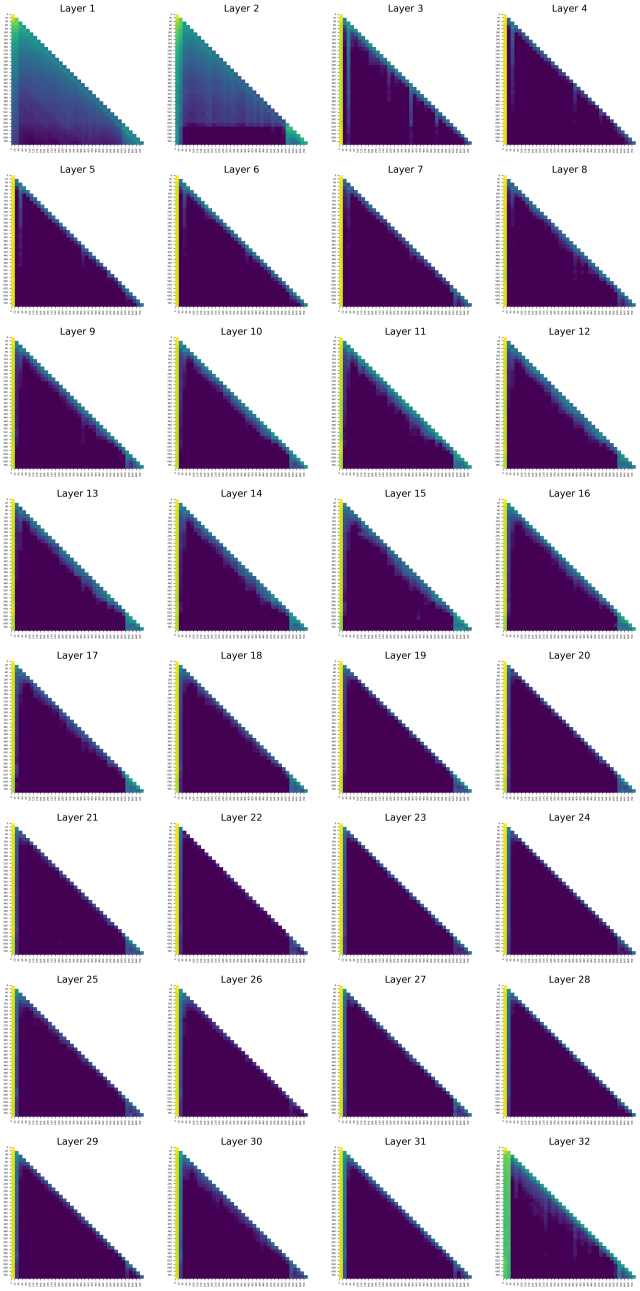

Fig. 7: Full Attention Maps of Each Layer of LLaVA.

A Evaluation Tasks Description

Image Captioning. Image captioning requires the model to generate a description for a given image. We choose Nocaps [1] and Flickr30k [33] as benchmarks and report CIDEr score [38] as metric. For image captioning tasks Nocaps and Flickr30k, we adopt prompt as "Describe the image in one sentence."

Visual Question Answering (VQA). VQA requires the model to generate an answer for a given image-question pair. We select the development set of A-OKVQA [35] and the test set of OCR-VQA [31] as the benchmark and the report the multiple choice (MC) score of AOKVQA and Rouge-L score of OCR-VQA. For AOKVQA, we adopt the the multiple choice version of evaluation and use prompt as: "Analyse the image and choose the best answer for the following question: {question} Options: {options}. Output the letter of the correct answer." For OCRVQA, we use the default question as prompt for each example as provided in the official dataset.

Multimodal Reasoning. Compared with VQA, multimodal reasoning requires more advanced perception, knowledge and reasoning skills of the model, which are more suitable benchmarks to evaluate the integrated abilities of LVLMs. We choose MMMU and PCA-Bench [6] as benchmarks. MMMU is a multimodal benchmark featuring multi-discipline tasks demanding college-level subject knowledge and reasoning skills. PCA-Bench is a complex embodied reasoning benchmark with error localization, which features three different domains including autonomous driving, robot and game. We report the multiple choice accuracy for the development set of MMMU and Perception, Cognition, Action, Genuine PCA scores for both the open and closed test set of PCA-Bench. We use the default prompts for each example as provided in the official dataset MMMU and PCA-Bench.

Video Question Answering. Similar to VQA for single image, Video Question Answering requires the model to generate answer given a video-question pair. Current LVLMs usually deal with video question answering tasks by sampling multiple frames as input, resulting in longer image token sequences. We choose TGIF-QA [12], MSVD-QA [44] and MSRVTT-QA [43] as benchmarks following the evaluation pipeline of Video-ChatGPT [30] and report the accuracy and chatgpt-score as metrics. We use the first 1K examples in each benchmark in our experiments due to the limited commercial API usage in evaluation. For all video QA tasks, we use the default question as the prompt as provided in Video-LLaVA, and use the same tool from Video-ChatGPT to conduct GPT evaluation.

Fine-grained Benchmarks For the evaluation of the influence of FastV on LVLM performance, we incorporate four distinct Fine-grained benchmarks: MME [10], Seed-Bench [16], SciQA-IMG [29], and MMVet [45]. MME offers a comprehensive evaluation of models' perception and cognition abilities across a diverse set of tasks, focusing on intuitive and quantifiable analysis without extensive prompt engineering. SEED-Bench, on the other hand, evaluates generative comprehension across multiple dimensions, ensuring question relevance and quality

through a mix of automated filtering and manual verification. While MME and SEED-Bench cover general abilities of LVLMs, SciQA-IMG and MMVet focus on the advanced aspects of multi-modal understanding. SciQA-IMG is a large-scale multimodal science question dataset annotated with detailed lectures and explanations. MMVet evaluates LVLMs on complex multimodal tasks, emphasizing multi-modal understanding and free-form answering capabilities, thus offering a comprehensive view of model performance.

B Computing Cost Estimation

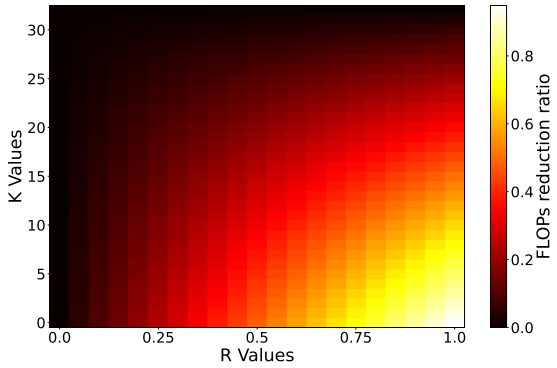

Fig. 8: The heat map of theoretical FLOPs reduction ratio. The color in the figure represents the reduction ratio in different K and R in FastV.

C Limitations

The FLOPs reduction ratio is based on the theoretical calculation considering the removal of image tokens, while actual inference cost can be influenced by a variety of factors such as inference framework optimization, specific CUDA kernels and hardwares. We aim at integrating FastV into mainstream LLM inference frameworks such as vLLM [15] for broader application in the future.