

TopV: Compatible Token Pruning with Inference Time Optimization for Fast and Low-Memory Multimodal Vision Language Model

Supplementary Material

A. Memory Analysis

Fig. 6- 11 illustrate the GPU memory usage records of InternVL2-2B on the AI2D and OCRBench datasets, obtained using different methods. It is worth noting that conducting GPU Profiler analysis on all the tasks is challenging due to the substantial time and computational resources. We only performed GPU Profiler analysis after GPU warmup on a subset of the dataset. During the inference phase, we monitored the peak memory usage of GPU at the end of generating each token, with the peak memory usage for each dataset summarized in Tab. 1 and 2. Specifically, Fig. 6 and Fig. 7 show the results for TopV on the AI2D and OCRBench datasets, respectively; Fig. 8 and Fig. 9 present the results for Baseline on AI2D and OCRBench; and Fig. 10 and Fig. 11 display the results for the FastV on AI2D and OCRBench. Notably, TopV and FastV achieve a 48% and 47% reduction in vision token FLOPs.

In practical usage, the InternVL2-2B model typically consumes approximately 4.5GB of GPU memory, which is static and remains unaffected by token pruning techniques. This static memory allocation does not change with different pruning methods. However, pruning methods have distinct effects on dynamic memory usage. For instance, as illustrated in Fig. 6, 8, and 10, when performing inference on the same AI2D dataset, the dynamic memory usage for TopV, Baseline, and FastV is denoted as 0.98GB, 1.8GB, 2.07GB, respectively. Compared to Baseline and FastV, the TopV achieves dynamic memory savings of 45.6% and 52.7%. Similarly, as shown in Fig. 7, 9, and 11, when inferring the same OCRBench dataset, the dynamic memory usage for TopV, Baseline, and FastV is 0.71GB, 1.25GB, 1.26GB, respectively. In this case, the TopV method results in dynamic memory savings of 43.2% and 43.7% relative to Baseline and FastV, respectively.

B. Discussion

Orthogonal to token pruning, memory efficiency, model compression [12, 24, 44–46, 50–52, 54] and KV cache compression [31, 62] are another promising techniques to enhance inference speed and reduce memory usage. An important question for future research is how to effectively balance token pruning and model compression/KV cache compression, potentially leading to extremely efficient multimodal language models.

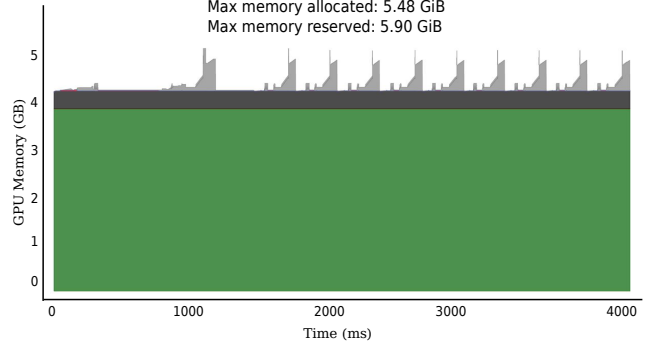


Figure 6. GPU Memory Usage of TopV on AI2D task for InternVL2-2B.

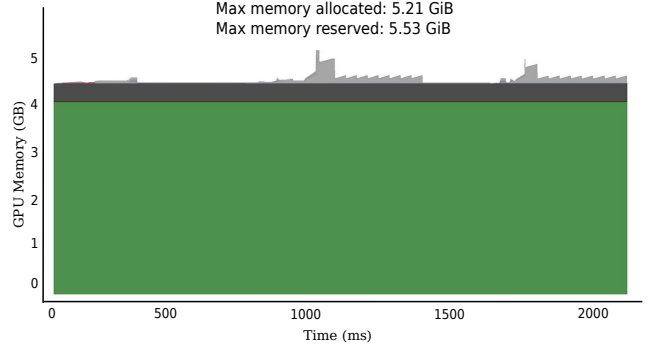


Figure 7. GPU Memory Usage of TopV on OCRBench task for InternVL2-2B.

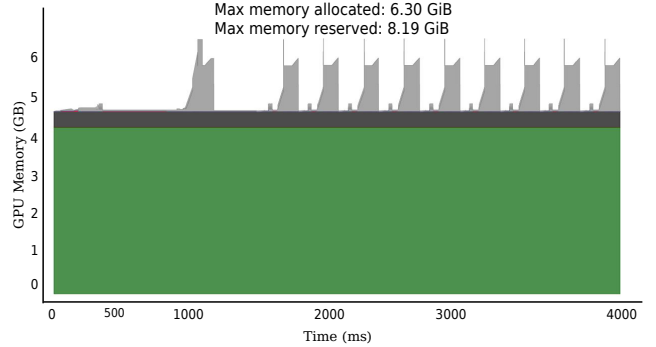


Figure 8. GPU Memory Usage of Baseline on AI2D task for InternVL2-2B.

C. Evaluation Examples

Tab. 8, 9, and 10 provide additional results for the dialogue tasks. While maintaining performance, we reduced the vision token FLOPs by 51%, and 48% for the LLaVA-v1.5-7B and InternVL2-2B models, respectively. As shown in Tab. 8, 9, 10, which correspond to the OCR, Captioning,

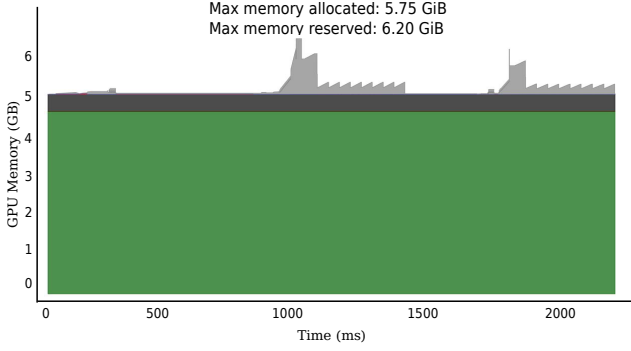


Figure 9. GPU Memory Usage of Baseline on OCRBench task for InternVL2-2B

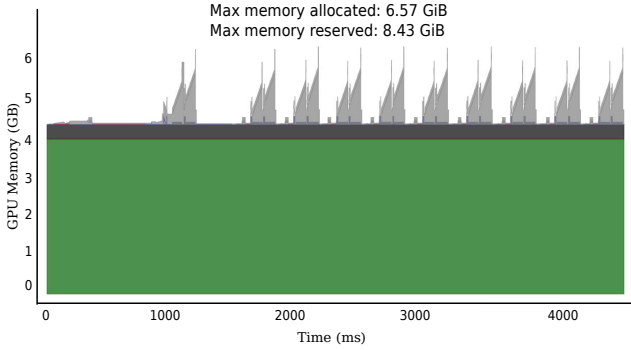


Figure 10. GPU Memory Usage of FastV on AI2D task for InternVL2-2B.

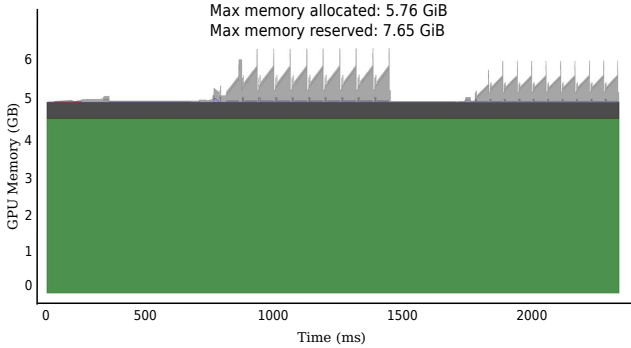


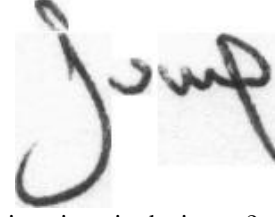
Figure 11. GPU Memory Usage of FastV on OCRBench task for InternVL2-2B.

and Writing tasks, it is clear that applying our TopV method yields results that are closer to the ground truth.

In the OCR task presented in Tab. 8, the model needs to recognize the word "Jump". The TopV method, applied to the LLaVA-v1.5-7B model, successfully identifies the word. while the Baseline fails to recognize it, and FastV only detects the letter "J". For the Writing task, as detailed in Tab. 10, the model is tasked with providing a comprehensive description of the image content. Both the TopV and Baseline generate similar descriptions, offering a more detailed description than FastV. In the Captioning task, TopV correctly identifies the rabbit, whereas both the Baseline

Table 8. Comparison of TopV, Baseline, and FastV in the OCR Task for LLaVA-v1.5-7B.

Visual input example, OCR Task:



User: what is written in the image?

TopV: The image features a handwritten signature, which appears to be a cursive "Jump" written in black ink.

Baseline: The image features a handwritten signature, which appears to be a cursive or script-style writing.

FastV: The image features a handwritten signature, which is a cursive letter "J" written in black ink.

and FastV methods erroneously classify it as a baby.

Table 9. Comparison of TopV, Baseline, and FastV in the Captioning Task for InternVL2-2B.

Visual input example, Captioning Task:



User: Provide a one-sentence caption for the provided image.

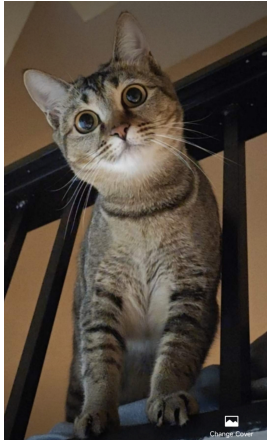
TopV: A man is sitting on a couch, reading a book, while a small rabbit lies on the floor, seemingly unaware of the man's presence.

Baseline: A man and a baby are sitting on a couch, with the man reading a book and the baby sleeping.

FastV: A man and a baby are sitting on a couch, with the man reading a book and the baby sleeping.

Table 10. Comparison of TopV, Baseline, and FastV in the Writing Task for LLaVA-v1.5-7B.

Visual input example, Writing Task:



User: Describe this picture in detail.

TopV: The image features a small, fluffy cat sitting on a metal railing, looking upwards with a curious expression. The cat appears to be looking at something above it, possibly a camera or another object of interest. The cat is positioned in the center of the scene, occupying a significant portion of the image. The railing it is sitting on is located near the top of the image, giving the cat a vantage point to observe its surroundings.

Baseline: The image features a small, fluffy cat sitting on a stair rail, looking upwards with a curious expression. The cat appears to be looking at something above it, possibly a camera or a person taking the photo. The cat is positioned in the center of the image, with its body facing the viewer. The stair rail is located in the middle of the scene, providing a clear view of the cat’s position.

FastV: The image features a cat sitting on a stair rail, looking upwards with a curious expression. The cat is positioned in the center of the scene, and its gaze is directed towards the top of the staircase. The stair rail is located near the top of the image, and the cat appears to be comfortably perched on it.
