ATP-LLaVA: Adaptive Token Pruning for Large Vision Language Models

Supplementary Material

1. Additional Implement Details

1.1. Experiment Details

We elaborate on the training details and hyperparameter design of ATP-LLaVA during the training process. Our method utlizes the LLaVA-1.5 [7] pre-trained projector and we perform fine-tuning based on this. Therefore, the reported hyperparameters are specific to the fine-tuning phase. We followed [7] to configure the majority of the model parameters. Tab. 1 provides a detailed list of hyperparameters involved in the model training phase. Under the aforementioned parameter settings, ATP-LLaVA is fine-tuned on 4 NVIDIA A100-SXM4-40GB GPUs.

Hyperparameter	Value
Overall batch size	128
Learning rate of LLM	2e-5
Learning rate of ATP module	1e-4
LR Scheduler	Cosine decay
DeepSpeed ZeRO Stage	ZeRO-3
Optimizer	AdamW
Warmup ratio	0.03
Epoch	1
Weight decay	0
Precision	bf16

Table 1. Hyperparameters of ATP-LLaVA.

1.2. Additional Ablation Details

In Sec. 4.3 and Tab. 3 of our main manuscript, we conducted experiments to explore efficient training strategies. In these experiments the language model is frozen and only the ATP modules are trainable. Besides, we remove the pruning positional encoding introduced in Sec 3.3.1 for the language model is frozen. It is worth noting that in these experiments, since we do not update the language model parameters during fine-tuning, we continue train ATP-LLaVA based on the fine-tuned LLaVA-1.5-7B model as the initialized model. Since LLaVA-1.5-7B already has vision understanding capability, the training process only needs to adjust the pruning strategy of the ATP module according to the loss after applying the pruning. We train the ATP module using the 665k visual instruction following data [7]. As the language model remains freezing, only the lightweight ATP module is updated during training, which greatly reduces the training resource overhead.

2. Additional Inference Efficiency Details

2.1. CUDA Time

We present a comparative efficiency analysis of ATP-LLaVA in Tab. 6 of our main manuscript. Here, we provide additional details on the CUDA time measurement during the inference phase. We primarily consider the following components that contribute to the reported CUDA time: image encoding time (if applicable) and transformers forward time. We exclude other computational times that are not dependent on the model itself and the caching strategy, such as model loading time, from the CUDA time measurement.

2.2. Storage Memory and FLOPs

The storage memory reported in Tab. 6 of our main manuscript corresponds to the full transformer activations on top of the vision tokens. This also refers to the size of the kv cache loaded during the inference process. Notably, the storage precision is bf16, which is consistent with the precision used in our training process.

2.3. Discussion about Flash Attention

The use of attention maps to compute importance scores for visual tokens has been demonstrated to be an effective method in both the ViT domain [4, 5] and the MLLM domain [2, 11], albeit with different implementation details. However, to reduce computational consumption, the introduction of the flash attention mechanism came at the cost of being unable to obtain the attention map of the current layer. To demonstrate the inference speedup achieved by our approach, we evaluated the CUDA time under the following strategies: (1) Not using flash attention at all. (2) Disabling flash attention at the pruning layer, while enabling it at other layers (as proposed by FastV). For intuitive comparison, we also report the CUDA time of the performance Upper Bound model (LLaVA-1.5-7B) when flash attention is applied uniformly.

As displayed in Tab. 3, we conduct an inference efficiency analysis on a single NVIDIA A100 using identical lengths of text prompts and single-image inputs. We found that the acceleration effect of flash attention is surprisingly insignificant in the inference phase. Specifically, the total CUDA time is only accelerated by 7.9% in the performance Upper Bound model LLaVA-1.5-7B, which is consistent with observations in the application of flash attention Huggingface Community and Qwen [1, 10]. Despite not employing flash attention, our approach still yields meaningful improvements in inference efficiency, incurring only a 1.9% performance loss while reducing inference CUDA

Method	Token	OCR & Chart				Knowledge		Spatial	Avg. Rate
		DocVQA	**OCRBench**	**ChartQA**	**TextVQA**	**AI2D**	**MMMU-val**	**RealWorldQA**	
LLaVA-1.5-7B [7]	576	21.8	297	17.8	58.2	55.5	37.0	54.8	100%
LLaVA-1.5-7B + **ATP**	144	21.4	286	17.5	57.3	54.9	36.6	53.1	97.9%

Table 2. Results on additional challenging benchmarks. Both the ATP module and LM are trainable. The percentage represents the average compression retention rates to Upper Bound model.

Method	Avg. Token	Accuracy	CUDA Time (ms) ↓	Δ
U.B. w/o flash attn	576	100%	432.7	-
U.B. w/ flash attn	576	100%	398.6	7.9%
ATP w/o flash attn	144	**98.1%**	266.4	38.4%
ATP w/ flash attn	144	**98.1%**	237.7	45.1%

Table 3. Efficiency analysis of the flash attention mechanism on CUDA times. Δ denotes the reduction ratio. U.B. means Upper Bound model LLaVA-1.5 [7].

Token	Avg. Rate	Langauge Model	Training Wall-time
88	92.4%	*Frozen*	4,376 s
88	94.6%	*Trainable*	11,403 s
144	95.0%	*Frozen*	5,172 s
144	98.1%	*Trainable*	13,089 s

Table 4. Training wall-time of ATP-LLaVA using different training strategies.

ATP-LLaVA Win % over Upper Bound Model				
H1	H2	H3	Human (Average)	GPT-4V
51.0	46.5	51.5	49.7	48.5

Table 5. Human evaluation results of ATP-LLaVA.

N_{target}	λ_{ATP}	λ_{target}	Token	**Average Rate**
130	1	1	132	96.7%
130	0.1	0.2	144	98.1%
130	0.05	0.2	152	98.5%
130	0.001	0.001	435	98.7%

Table 6. Additional results of ATP-LLaVA using different hyper-parameter settings.

time by 38.4%. Furthermore, applying flash attention to the unpruned layers (layers without ATP module) can further accelerate inference while maintaining a constant performance loss. These experiments demonstrate that the inability to apply flash attention to the pruning layers (layers with ATP module) is orthogonal to the contribution of ATP-LLaVA to the field of LVLM inference acceleration.

3. Additional Training Details

We report the training time in Tab. 4, with all experiments conducted using four A100-40GB GPUs. The parameter and FLOPs overhead introduced by ATP module is negligible compared to the original LLM's parameter and the efficiency gains from token pruning.

4. Additional Results

4.1. Visualized Results of Output

Fig. 1 illustrates the visualized examples of ATP-LLaVA for vision compression. We compare ATP-LLaVA with the Upper Bound model (LLaVA-1.5-7B) (denoted as **U.B.** throughout the figure). Correct responses are highlighted in green, while incorrect responses are marked in red. We further show some of the failure cases. We observe that for certain fine-grained tasks, such as OCR and positional relation recognition, ATP-LLaVA can still provide accurate answers even after pruning a significant number (75%) of visual tokens. For image detail description, we find that using 144 tokens yields more fine-grained and accurate answers compared to pruning to 88 tokens, with performance comparable to the Upper Bound model.

4.2. Human Evaluation Results

Tab. 5 below presents the human evaluation results on 200 random samples. Three annotators (H1-H3) compare the outputs of ATP-LLaVA (144 tokens) and the Upper Bound model, selecting the better output or declaring a tie. Besides, GPT-4V is also used to compare two outputs as a supplementary validation.

4.3. Additional Benchmarks

In addition to the general benchmarks reported in our paper, Tab. 2 presents results on additional challenging benchmarks, including OCR & chart, knowledge, and vision spatial perspective tasks. Due to the lack of high-resolution training data in LLaVA-1.5, the upper-bound model performance is suboptimal.

4.4. Discussion about Hyper-parameters

We involve four hyper-parameters: λ_{sample}, T, λ_{ATP}, and λ_{target}. λ_{sample} maps spatial pruning scores to a range be-

Method (***Trainable LM***)	Token	**VQA^{v2}**	**GQA**	**SEED**	**SQA^{I}**	**POPE**	**MME**	**MMB**	**Avg. Rate**
VILA-7B [6]	576	79.9	62.3	61.1	68.2	85.5	1533.0	68.9	100%
VILA-7B + **ATP**	144	77.8	60.9	60.2	66.8	84.1	1492.9	70.2	98.5%
Method (***Frozen LM***)	Token	**DocVQA**	**OCRBench**	**ChartQA**	**InfoVQA**	**AI2D**	**MMMU-val**	**MMB-EN**	**Avg. Rate**
InternVL-2-8B [3]	1122*	91.6	794	83.3	74.8	83.8	51.2	81.7	100%
InternVL-2-8B + **ATP**	256	88.5	786	81.2	73.4	81.0	48.3	79.4	96.7%
Qwen2-VL-7B [10]	1076*	94.5	845	83.0	76.5	-	54.1	83.0	100%
Qwen2-VL-7B + **ATP**	256	91.2	809	79.5	73.7	-	51.4	81.2	96.2%

Table 7. Results of applying ATP module to advanced VLMs, including VILA [6], InternVL-2 [3], and Qwen2-VL [10]. The percentage represents the average compression retention rates to Upper Bound model. * means dynamic resolution.

Method	Token	**MMB**	**MME**	**POPE**	**SQA^{I}**	**VQA^{v2}**	**Avg. Rate**
LLaVA-1.5 [7]	576	64.3	1510.7	85.8	71.6	78.5	100%
LLaVA-PruMerge+ [9]	144	64.9	1462.4	84.0	68.3	76.8	97.8%
ATP-LLaVA	144	68.2	1522.9	86.3	71.2	79.0	**101.6%**

Table 8. Comparison with LLaVA-PruMerge+ [9] on several benchmarks under the same training setting with LLaVA-PruMerge+.

tween 0 and 1. T is set to a sufficiently large value 10^6 to ensure the sigmoid functions as a differentiable mask. These two parameters are fixed and remain unchanged during our design. λ_{ATP} and λ_{target} regulate the impact of computational budget constraints during training. Setting these values too high or low may result in the model being overly constrained or insufficiently influenced by the computational budget. Within a reasonable range, the model is **in**sensitive to these parameters and the training stage is stable. Tab. 6 presents their impact.

4.5. Advanced VLMs

Tab. 7 presents the results of applying the ATP module to advanced VLMs (e.g., VILA [6], InternVL-2 [3], and Qwen2-VL [10]). For VILA, we introduce the ATP module based on its pretrained weights from Stage 2 and proceed with Stage 3 (SFT), training both the ATP module and the LM. For InternVL-2 and Qwen2-VL, due to the lack of public data, we integrate the ATP module with finetuned models and train only the ATP module using the SFT data of LLaVA-1.6 [8], keeping the LM frozen. Our method demonstrates generalization ability across advanced methods and challenging benchmarks.

4.6. Comparison with LLaVA-PruMerge+

In Tab. 1 of our main paper, the improvement of our method over LLaVA-PruMerge+ [9] is not significant, primarily due to the differences in training strategies. LLaVA-1.5 comprises two stages: Stage 1 (pretraining, where the projector is trained for alignment) and Stage 2 (SFT, for fine-tuning the LLM). ATP-LLaVA, based on the weights from Stage 1, introduces the pruning module and completes Stage 2 training using SFT data. In contrast, LLaVA-PruMerge+ is based on LLaVA-1.5 (weights after Stage 2 SFT) and further trained for one epoch using SFT data. We opted not to adopt this strategy initially, as most VLMs are fine-tuned for only one epoch. The following table presents the results of training ATP-LLaVA using the LLaVA-PruMerge+ strategy. Under the same pruning ratios, ATP-LLaVA outperforms the LLaVA-PruMerge+.

5. Broader Impacts

Our research is related to Large Language Models and Large Vision Language Models. Although we use publicly available and compliant training data, it is essential to consider the potential negative societal implications. One concern is the possibility of generating unintended or malicious outputs that could be exploited to disseminate misinformation. To address these risks, we recommend that future research prioritize the development of mechanisms for detecting and mitigating misuse, as well as implementing controlled release strategies to ensure responsible model deployment.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 1

[2] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models, 2024. 1

[3] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 3

[4] Cao Jianjian, Ye Peng, Li Shengze, Yu Chong, Tang Yansong, Lu Jiwen, and Chen Tao. Madtp: Multimodal alignment-guided dynamic token pruning for accelerating vision-language transformer. *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 1

[5] Sehoon Kim, Sheng Shen, David Thorsley, Amir Gholami, Woosuk Kwon, Joseph Hassoun, and Kurt Keutzer. Learned token pruning for transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 784–794, 2022. 1

[6] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2023. 3

[7] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 1, 2, 3

[8] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 3

[9] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models, 2024. 3

[10] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 3

[11] Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024. 1

User: What is the relation between the soccer player and the football player?
A. They are teammates
B. Cannot be determined
C. They are not interacting
D. They are opponents

U.B. (576 tokens): C. They are not interacting.
ATP-LLaVA (144 tokens): C. They are not interacting

User: How many people are present in the image?
A: 1
B: 2
C: 3
D: 4

U.B. (576 tokens): C: 3.
ATP-LLaVA (144 tokens): D: 4.

User: What are the people in the picture doing?

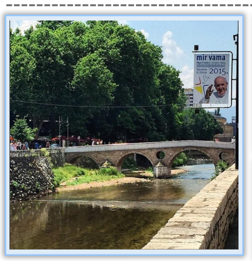
U.B. (576 tokens):
The people in the picture are playing soccer, kicking a soccer ball on a field.
ATP-LLaVA (144 tokens): The people in the picture are playing soccer on a field, with a group of men in yellow and blue uniforms competing against each other.

User: What is the name of the player in blue on the right?

U.B. (576 tokens): The name of the player in blue on the right is "Sila".
ATP-LLaVA (144 tokens): The name of the player in blue on the right is "Shea."

User: Do you know who drew this painting? What's so strange about this painting?

U.B. (576 tokens):
The painting is a creative and unusual artwork, featuring a dog as the main subject, dressed as the famous Leonardo da Vinci's Mona Lisa. The artist who drew this painting is likely a skilled and imaginative individual, as they have managed to capture the essence of the original painting while incorporating a dog as the central figure. The strange aspect of this painting is the combination of the recognizable Mona Lisa pose and the dog's presence, which creates a unique and unexpected visual experience for the viewer.
ATP-LLaVA (144 tokens):
The painting is a copy of Leonardo da Vinci's famous painting "Mona Lisa", which features a woman with a long, flowing dress. However, in this particular painting, the woman is wearing a dog's face, which is quite unusual and humorous. The dog's face is painted over the woman's face, creating a unique and amusing twist on the original artwork.



User: How does the water appear in the image?
A. Crystal clear
B. Muddy and murky
C. Blue and calm
D. Rough and wavy

U.B. (576 tokens):
B. Muddy and murky
ATP-LLaVA (144 tokens):
B. Muddy and murky

User: What are the numbers on the billboard on the right?

U.B. (576 tokens):
The numbers on the billboard on the right are 2015.

ATP-LLaVA (144 tokens):
The numbers on the billboard on the right are 2015.

User: Describe the scene in detail.

U.B. (576 tokens):
The image features a picturesque scene of a river flowing through a city, with a bridge crossing over it. The bridge is adorned with a large sign, possibly featuring Pope Francis, who is also present in the scene. The river is surrounded by trees, creating a serene atmosphere. There are several people scattered throughout the scene, enjoying the view or going about their daily activities. Some of them are located near the bridge, while others can be seen further away from the river. The combination of the river, bridge, trees, and people creates a lively and inviting urban environment.

ATP-LLaVA (88 tokens):
The image features a picturesque scene of a river flowing under a bridge, surrounded by trees and a lush green hillside. The bridge is located near a city, and the river is lined with trees, creating a serene and peaceful atmosphere. There are several people scattered throughout the scene, some standing near the river and others walking along the bridge. A few cars can be seen parked or driving near the river, and a bench is placed along the side of the river, providing a spot for people to sit and enjoy the view.

ATP-LLaVA (144 tokens):
The image features a picturesque scene of a river flowing under a bridge, surrounded by trees and a lush green hillside. The bridge is located near a city, and the river is lined with trees, creating a serene and peaceful atmosphere. There are several people scattered throughout the scene, some standing near the river and others walking along the bridge. Some are situated near the bridge, while others are scattered at a distance from the river. The river, bridge, trees, and people come together to form a vibrant and welcoming urban space.

Figure 1. Examples of instructions and model outputs demonstrating the vision compression capabilities of ATP-LLaVA. Green and red represent the correct and incorrect responses, respectively.