DOCPRUNE: Efficient Document Question Answering via Background, Question, and Comprehension-aware Token Pruning

Joonmyung Choi[1] Sanghyeok Lee[2] Jongha Kim[1] Sehyung Kim[1] Dohwan Ko[1]
Jihyung Kil[3] Hyunwoo J. Kim[2]
[1]Korea University [2]KAIST [3]Adobe Research
{pizard, jonghakim, skim129, ikodoh}@korea.ac.kr
jkil@adobe.com {cat0626, hyunwoojkim}@kaist.ac.kr

1. Experimental settings

In this section, we delineate implementation details for applying DOCPRUNE to M3DocRAG [3] and outline the hyperparameter choices for both our method and the baseline token pruning approaches.

Implementation details. All results reported in the main paper are obtained by evaluating the model in a training-free manner, without any additional training or fine-tuning. When using Qwen2-VL for question answering in M3DocRAG [3], a spatial token merger with a 2×2 grid is inserted after the vision encoder. Note that applying BTP or QTP with arbitrary token pruning in the encoder stage would break this grid structure and cause the merger to behave incorrectly. To preserve compatibility, DOCPRUNE applies BTP and QTP at the encoder input by grouping spatial tokens into 2×2 blocks and pruning at the block level, so that the merger consistently receives a valid token layout. We also adapt DOCPRUNE to remain compatible with FlashAttention [4]. While FlashAttention reduces memory and computational overhead by avoiding storage of the full attention matrix in HBM, this makes token-wise attention scores unavailable to our CTP module. To address this, we simply recompute attention only for the last token using a reduced query only at the selected layer, providing the attention scores for token pruning at negligible additional cost.

Sensitivity analysis. In Tab. A, we present a sensitivity analysis of the hyperparameters. Notably, DOCPRUNE consistently surpasses all other pruning methods in all metrics, regardless of hyperparameter settings, demonstrating its robustness and superiority. In detail, adjusting τ_{bg} and τ_{qst} allows flexible control over the trade-off between throughput and QA accuracy (Tab. A-(a) and (b)). Moreover, varying τ_{comp} and τ_{att} results in negligible performance fluctuation (Tab. A-(c) and (d)), demonstrating the robustness of the model.

Hyperparameters for DOCPRUNE. For hyperparameter choice, since the number of pages affects the distribution of visual features and attention, we use separate hyperparameter sets for 1, 2, and 4 page inputs. Specifically, we perform a grid search over τ_{bg} and τ_q with a step size of 0.1, and over τ_{info} with a step size of 5. For τ_{att}, we adopt finer step sizes of [0.1, 0.05, 0.025] as pages increase, as attention scores become more dispersed when the number of visual tokens increases. The final hyperparameters are summarized in Tab. B.

Hyperparameters for previous pruning methods. For a fair comparison with previous works, we tune FastV [2], DivPrune [1], and VTW [5] using the search ranges specified in their original configurations. For clarity, we here use the same notation used in previous works, where K denotes the drop layer and R denotes the pruning ratio. FastV is originally evaluated with $l \in \{0, 2, 3, 5\}$ and $r \in \{0.5, 0.75, 0.9\}$ in their paper, and $l = 2$, $r = 0.5$ con-

Value	Throughput		Overall		Value	Throughput		Overall	
	ENC	DEC	EM	F1		ENC	DEC	EM	F1
(a) Background threshold τ_{bg}					**(b) Relevance threshold τ_{qst}**				
1.0	4.9	5.5	28.1	32.1	0.1	4.5	5.0	27.9	31.9
0.9	5.3	5.8	27.9	32.0	0.2	4.8	5.5	27.7	31.9
0.8	6.1	6.4	27.3	31.3	0.3	5.3	5.8	27.9	32.0
0.7	7.9	7.5	26.9	30.9	0.4	5.8	6.3	27.4	31.2
(c) Comprehension threshold τ_{comp}					**(d) Attention threshold τ_{att}**				
60	5.0	5.9	27.8	32.0	0.1	5.2	5.2	27.7	31.8
65	5.3	5.8	27.9	32.0	0.3	5.3	5.7	27.7	31.8
70	5.3	5.7	27.9	32.0	0.5	5.3	5.8	27.9	32.0
75	5.3	5.7	27.9	31.9	0.7	5.3	5.9	27.8	32.0

Table A. **Sensitivity analysis on M3DocVQA.** The highlighted row indicates the default settings for DOCPRUNE.

Page	RET	QA				
	τ_{bg}	τ_{bg}	τ_e	τ_q	τ_{info}	τ_{att}
1	0.9	0.9	1	0.3	65	0.5
2	1.0	1.0	1	0.3	60	0.25
4	1.0	0.8	1	0.4	45	0.075

Table B. **Hyperparameters for M3DOCVQA.**

sistently perform best in our setting. In DivPrune, the original pruning ratio of $r = 0.902$ is heuristic and yields performance similar to random pruning in our document setting. Therefore, we additionally explore $r \in \{0.3, 0.5, 0.7, 0.9\}$, and report the result with $r = 0.5$, which achieves the best results. DivPrune does select a drop-layer l because it prunes tokens immediately before the decoder. For VTW, the results are reported with $l \in \{16, 20\}$ with an original pruning ratio of $r = 1.0$, and we find that $l = 20$ yields the best results across all page counts.

2. Qualitative analysis of removing irrelevant regions for QA

We qualitatively analyze how removing irrelevant regions affects QA predictions. Fig. A presents the examples illustrating how removing irrelevant regions improves question-answering performance. Given the original document in the left column, the middle and right columns show the attention map referenced by the last token, visualized with and without token pruning, respectively. The baseline often focuses on noisy or irrelevant areas of the document, leading to incorrect predictions. In contrast, our method suppresses unrelated regions and emphasizes areas aligned with the question, enabling the model to produce the correct answer.

3. Comparison of comprehension metrics

Fig. B analyzes the relationship between F1 scores and the comprehension criteria in Tab.4 of the main paper. For each metric, the left heatmap shows the average F1 within each value interval, while the right heatmap shows the number of samples in the corresponding interval. (a) **L2 Norm** shows a clear and stable relationship with accuracy. Samples with higher norm values consistently achieve higher F1 scores, and samples that reach such values at earlier layers also tend to perform better. This is consistent with our hypothesis that easier samples attain a confident representation more quickly, so high L2 norms emerging in shallow layers indicate cases that the model can answer correctly with fewer computation steps. Next, (b) **Feature** Δ shows a weaker relationship, with many intervals exhibiting high metric values but low accuracy. For example, in Layer 21, the 80-90 interval achieves an average F1 of 45.5, which is higher than the 90-100 and 100-110 intervals (43.1 and 42.6), despite having a lower metric value. The last metric, (c) **Entropy**, is the least informative, showing the weakest correlation with the F1 score. Although lower entropy values often correspond to higher F1, this relationship is inconsistent across intervals, and most samples cluster in the 9–10 range across layers, making entropy a poor discriminator of comprehension. Overall, the L2 norm provides the most reliable indicator of layer-wise comprehension.

4. Qualitative results of DocPrune

Fig. C presents additional qualitative results. The figure illustrates how the number and proportion of remaining tokens change as BTP, QTP, and CTP are sequentially applied. These examples show that DocPrune adaptively adjusts the pruning ratio at each stage for each document–question pair, removing substantial redundancy while preserving the semantic evidence necessary for accurate answers.

5. Resource Availability

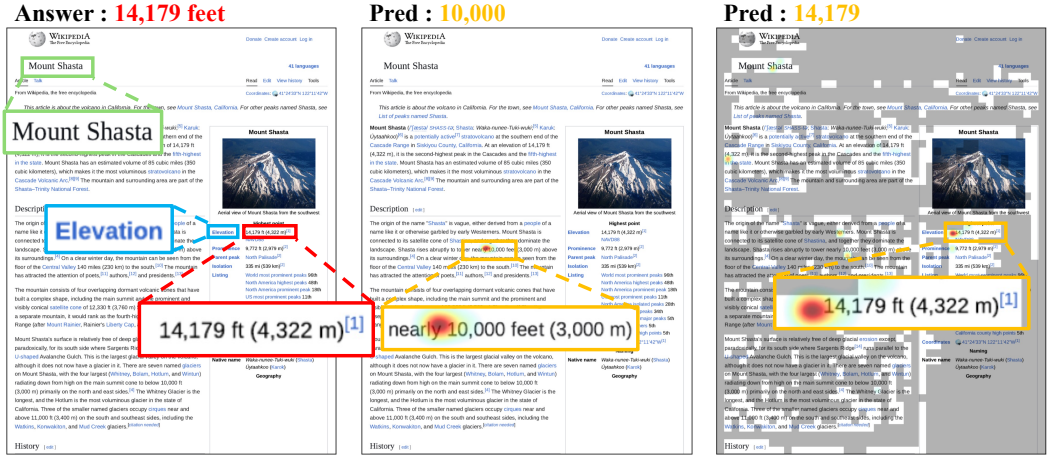
To support the public release and ensure reproducibility, we provide the official links to the models and datasets utilized in our experiments:

- **Qwen2-VL-7B**: https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct
- **Qwen2.5-VL-7B**: https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct
- **ColPali-v1**: https://huggingface.co/vidore/colpali-v1
- **VDocRetriever**: https://huggingface.co/NTT-hil-insight/VDocRetriever-Phi3-vision
- **M3DocVQA**: https://huggingface.co/datasets/m3docrag/m3docvqa
- **MMLongBench-Doc**: https://huggingface.co/datasets/m3docrag/mmlongbench-doc
- **OpenDocVQA**: https://huggingface.co/datasets/NTT-hil-insight/OpenDocVQA

: answer : pred : cue1 : cue2

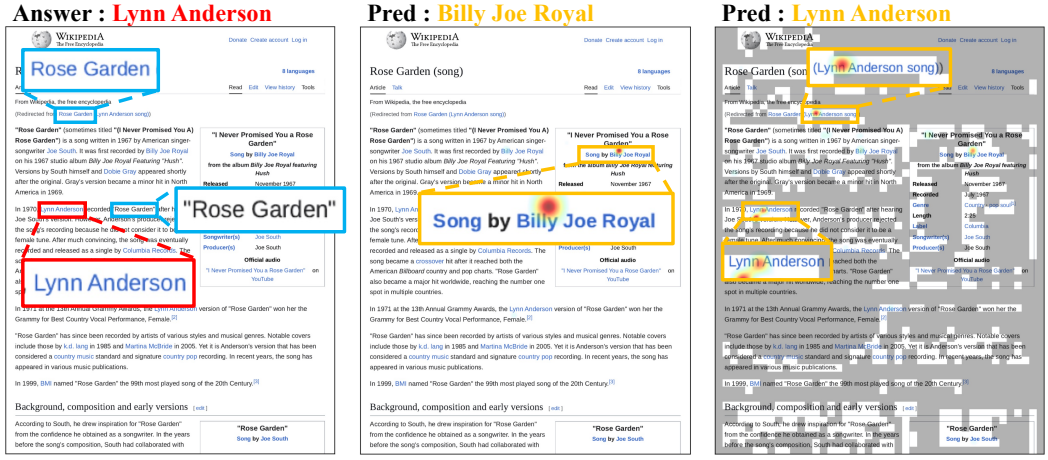
Q1. Mount shasta is a volcano in northern california that towers more than how many feet

Answer : 14,179 feet | **Pred : 10,000** | **Pred : 14,179**



Q2. Who sang i never promised you a rose garden

Answer : Lynn Anderson | **Pred : Billy Joe Royal** | **Pred : Lynn Anderson**



Q3. When is the new season of daredevil coming out

Answer : 2018 | **Pred : March 1, 2022.** | **Pred : 2018**

(a) Origin **(b) Baseline** **(c) Ours**

Figure A. **Qualitative results on irrelevant-region suppression.**

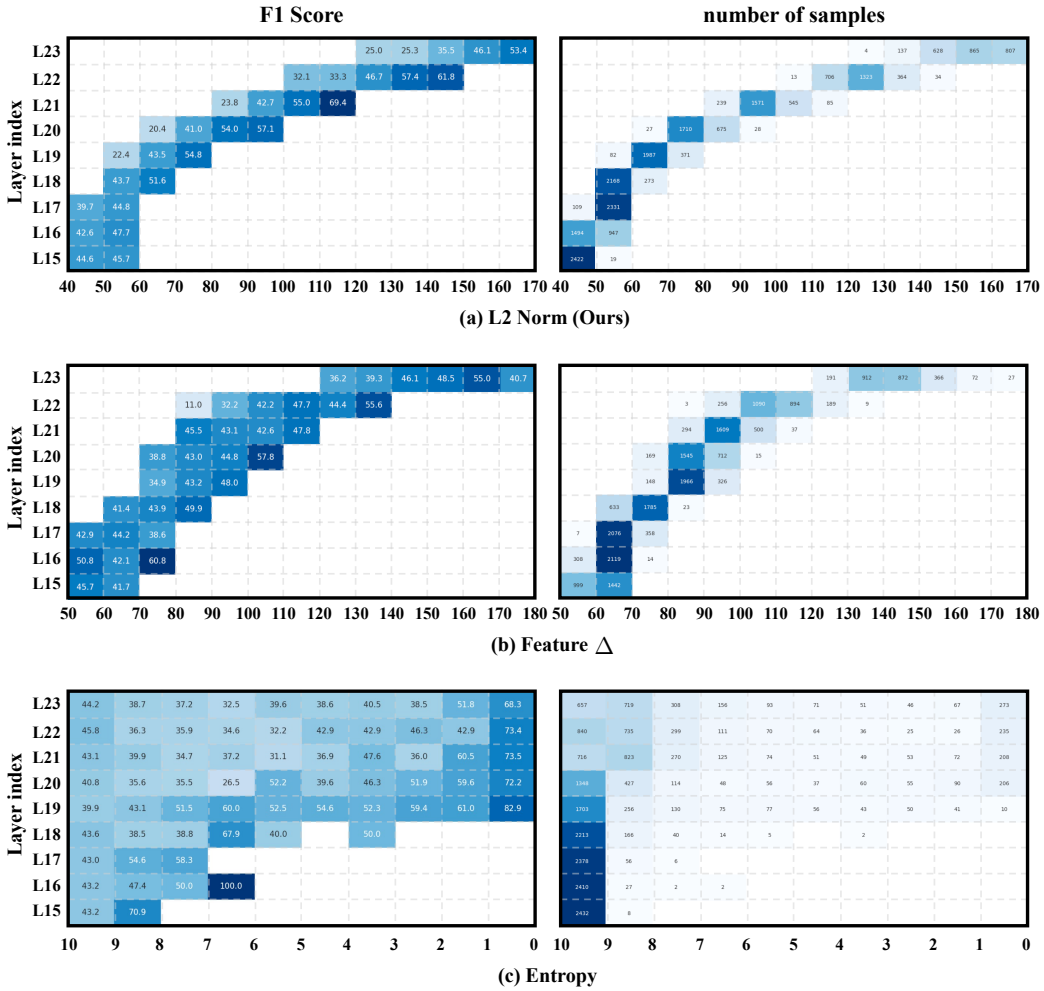
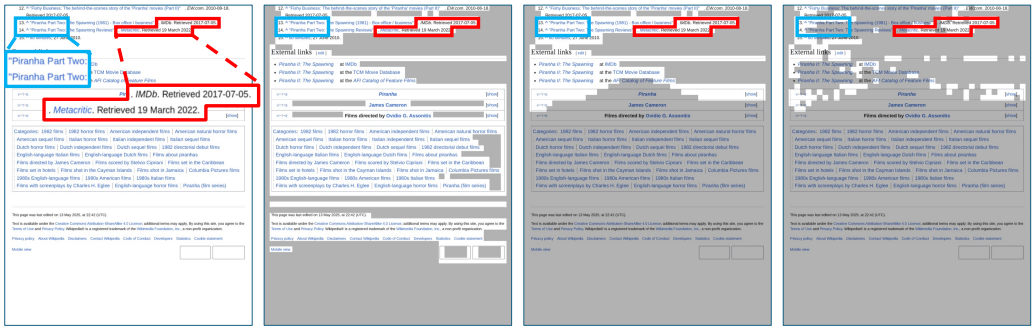


Figure B. **Performance and number of samples by layers and multiple criteria.**

: answer : cue1 : cue2

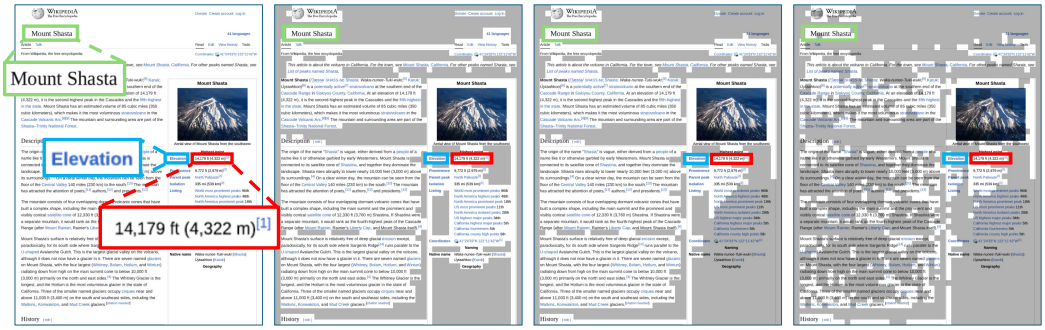
Q1. Is there going to be a piranha 2 A1. : Yes

2508 (100%) 1340 (53%) 406 (16%) 197 (8%)



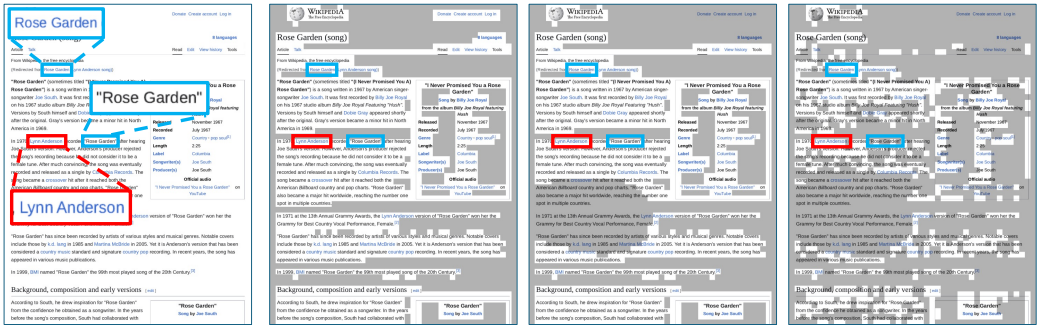
Q2. Mount Shasta, a northern California volcano, rises above how many feet? A2. : 14,179 feet

2508 (100%) 1633 (65%) 1457 (58%) 775 (31%)



Q3. Who sang i never promised you a rose garden A3. : Lynn Anderson

2508, 100% 1520 (61%) 1279 (51%) 759 (30%)



Q4. When is the new season of daredevil coming out A4. : 2018

2508 (100%) 1476 (59%) 1086 (43%) 549 (22%)

(a) Origin (b) BTP (c) BTP & QTP (d) BTP & QTP & CTP

Figure C. **Additional Qualitative results of DocPrune.**

References

[1] Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *CVPR*, 2025. 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. In *ECCV*, 2024. 1

[3] Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *arXiv*, 2024. 1

[4] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023. 1

[5] Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *AAAI*, 2025. 1