3.5.1. Training DeepEncoder

Following Vary [36], we utilize a compact language model [15] and use the next token prediction framework to train DeepEncoder. In this stage, we use all OCR 1.0 and 2.0 data aforementioned, as well as 100M general data sampled from the LAION [31] dataset. All data is trained for 2 epochs with a batch size of 1280, using the AdamW [23] optimizer with cosine annealing scheduler [22] and a learning rate of 5e-5. The training sequence length is 4096.

3.5.2. Training DeepSeek-OCR

After DeepEncoder is ready, we use data mentioned in Section 3.4 to train the DeepSeek-OCR. with the entire training process conducted on the HAI-LLM [14] platform. The entire model uses pipeline parallelism (PP) and is divided into 4 parts, with DeepEncoder taking two parts and the decoder taking two parts. For DeepEncoder, we treat SAM and the compressor as the vision tokenizer, place them in PP0 and freeze their parameters, while treating the CLIP part as input embedding layer and place it in PP1 with unfrozen weights for training. For the language model part, since DeepSeek3B-MoE has 12 layers, we place 6 layers each on PP2 and PP3. We use 20 nodes (each with 8 A100-40G GPUs) for training, with a data parallelism (DP) of 40 and a global batch size of 640. We use the AdamW optimizer with a step-based scheduler and an initial learning rate of 3e-5. For text-only data, the training speed is 90B tokens/day, while for multimodal data, the training speed is 70B tokens/day.

Table 2 | We test DeepSeek-OCR's vision-text compression ratio using all English documents with 600-1300 tokens from the Fox [21] benchmarks. Text tokens represent the number of tokens after tokenizing the ground truth text using DeepSeek-OCR's tokenizer. Vision Tokens=64 or 100 respectively represent the number of vision tokens output by DeepEncoder after resizing input images to 512×512 and 640×640.

Text Tokens	Vision Tokens =64		Vision Tokens=100		
	Precision	Compression	Precision	Compression	Pages
600-700	96.5%	10.5×	98.5%	6.7×	7
700-800	93.8%	11.8×	97.3%	7.5×	28
800-900	83.8%	13.2×	96.8%	8.5×	28
900-1000	85.9%	15.1×	96.8%	9.7×	14
1000-1100	79.3%	16.5×	91.5%	10.6×	11
1100-1200	76.4%	17.7×	89.8%	11.3×	8
1200-1300	59.1%	19.7×	87.1%	12.6×	4

4. Evaluation

4.1. Vision-text Compression Study

We select Fox [21] benchmarks to verify DeepSeek-OCR's compression-decompression capability for text-rich documents, in order to preliminarily explore the feasibility and boundaries of contexts optical compression. We use the English document portion of Fox, tokenize the ground truth text with DeepSeek-OCR's tokenizer (vocabulary size of approximately 129k), and select documents with 600-1300 tokens for testing, which happens to be 100 pages. Since the number of text tokens is not large, we only need to test performance in Tiny and Small modes, where Tiny mode corresponds to 64 tokens and Small mode corresponds to 100 tokens. We use the prompt