a feature of the forgetting mechanism. When compressing tokens by nearly 20×, we find that precision can still approach 60%. These results indicate that optical contexts compression is a very promising and worthwhile research direction, and this approach does not bring any overhead because it can leverage VLM infrastructure, as multimodal systems inherently require an additional vision encoder.

Table 4 | Edit distances for different categories of documents in OmniDocBench. The results show that some types of documents can achieve good performance with just 64 or 100 vision tokens, while others require Gundam mode.

Mode \ Type	Book	Slides	Financial Report	Textbook	Exam Paper	Magazine	Academic Papers	Notes	Newspaper	Overall
Tiny	0.147	0.116	0.207	0.173	0.294	0.201	0.395	0.297	0.94	0.32
Small	0.085	0.111	0.079	0.147	0.171	0.107	0.131	0.187	0.744	0.205
Base	0.037	0.08	0.027	0.1	0.13	0.073	0.052	0.176	0.645	0.156
Large	0.038	0.108	0.022	0.084	0.109	0.06	0.053	0.155	0.353	0.117
Gundam	0.035	0.085	0.289	0.095	0.094	0.059	0.039	0.153	0.122	0.083
Guandam-M	0.052	0.09	0.034	0.091	0.079	0.079	0.048	0.1	0.099	0.077

4.2. OCR Practical Performance

DeepSeek-OCR is not only an experimental model; it has strong practical capabilities and can construct data for LLM/VLM pretraining. To quantify OCR performance, we test DeepSeek-OCR on OmniDocBench [27], with results shown in Table 3. Requiring only 100 vision tokens (640×640 resolution), DeepSeek-OCR surpasses GOT-OCR2.0 [38] which uses 256 tokens; with 400 tokens (285 valid tokens, 1280×1280 resolution), it achieves on-par performance with state-of-the-arts on this benchmark. Using fewer than 800 tokens (Gundam mode), DeepSeek-OCR outperforms MinerU2.0 [34] which needs nearly 7,000 vision tokens. These results demonstrate that our DeepSeek-OCR model is powerful in practical applications, and because the higher tokens compression, it enjoys a higher research ceiling.

As shown in Table 4, some categories of documents require very few tokens to achieve satisfactory performance, such as slides which only need 64 vision tokens. For book and report documents, DeepSeek-OCR can achieve good performance with only 100 vision tokens. Combined with the analysis from Section 4.1, this may be because most text tokens in these document categories are within 1,000, meaning the vision-token compression ratio does not exceed 10×. For newspapers, Gundam or even Gundam-master mode is required to achieve acceptable edit distances, because the text tokens in newspapers are 4-5,000, far exceeding the 10× compression of other modes. These experimental results further demonstrate the boundaries of contexts optical compression, which may provide effective references for researches on the vision token optimization in VLMs and context compression, forgetting mechanisms in LLMs.

4.3. Qualitative Study

4.3.1. Deep parsing

DeepSeek-OCR possesses both layout and OCR 2.0 capabilities, enabling it to further parse images within documents through secondary model calls, a feature we refer to as "deep parsing". As shown in Figures 7,8,9,10, our model can perform deep parsing on charts, geometry, chemical formulas, and even natural images, requiring only a unified prompt.