Multimodal Visual Language Comprehension

Andriy Sukh[1] and Artur Kiulian[2]

[1] Ukrainian Catholic University, Lviv, Ukraine
sukh.pn@ucu.edu.ua
[2] San Francisco, California, United States
https://www.ukrainenow.org
https://www.openbabylon.com
akiulian@gmail.com

Abstract. Visual language comprehension is a pivotal field at the intersection of Computer Vision (CV) and Natural Language Processing (NLP) representing a specialized subset of Visual Document Understanding (VDU) that focuses on interpreting visual textual information. This field addresses tasks requiring a joint understanding of visual and textual modalities and has gained significant interest due to the limitations of Optical Character Recognition (OCR) particularly in handling noisy, handwritten, low-resolution or multilingual documents.
Recent advancements in multimodal learning have introduced OCR-free models such as PaliGemma [1, 2], Pix2Struct [3], Qwen2-VL [4] which leverage pre-trained visual-text encoders and language decoders to interpret document layouts and extract meaningful textual information directly.
In this study we conduct a systematic evaluation of OCR-free Vision-Language Models (VLMs) for Ukrainian document understanding. We benchmark the selected models across multiple document AI tasks including OCR-free text recognition, layout analysis, table extraction and document classification. Our approach involves generating semi-synthetic datasets of printed and handwritten Ukrainian text and instruction-tuning these models to assess their performance.
We anticipate that the results of this study will provide insights into the strengths and limitations of OCR-free document AI models contributing to their adaptation for low-resource languages and document processing.

Keywords: Visual Document Understanding · Vision-Language Model · OCR-Free Document AI

1 Introduction and Motivation

Multimodal learning has made significant progress in aligning visual and textual representations as demonstrated by models such as CLIP [5] and SigLIP [6]. While effective on natural images, these models fall short when applied to visually complex, text-dense documents that require precise layout and structural understanding.

Document understanding systems traditionally rely on OCR pipelines to extract text prior to downstream processing. However, OCR performance degrades under noisy, handwritten, low-resolution or multilingual conditions introducing critical errors that propagate through the pipeline [7]. Layout-aware models such as LayoutLM [8, 9], DocFormer [10] and UniDoc [11] attempt to address these issues by incorporating positional embeddings and visual features. Nonetheless, they remain dependent on OCR and show limited robustness in low-resource languages [12, 13].

This study is motivated by the emerging potential of OCR-free VLMs that directly generate structured text from document images. These models promise end-to-end reasoning across visual, textual and layout modalities offering robustness in challenging conditions and adaptability across languages. We focus on evaluating such models in Ukrainian, a low-resource language, to assess their viability in real-world document understanding tasks spanning OCR, layout analysis, table extraction and classification without reliance on OCR preprocessing.

2 Related Work

2.1 Method for Literature Search

To ensure that the literature review and analysis are grounded, unbiased and comprehensive we adopted a systematic approach for collecting relevant and representative research papers. The methodology combined established techniques in citation analysis, probabilistic topic modeling and snowball sampling to iteratively refine the selection of sources.

We began by collecting an initial seed of literature samples focusing on the work of academic and industry leaders. The most cited and prominent examples were selected ensuring that foundational and cutting-edge advancements in the field were covered. This approach aligns with prior studies emphasizing the importance of starting with highly influential seed papers in order to build robust citation networks [14].

The initial seed set was then fed into the Controlled Snowball Sampling tool which utilizes Probabilistic Topic Modeling (PTM) to moderate the snowball sampling process [15]. PTM facilitated the identification of thematically relevant papers by modeling topic distributions over short-text abstracts effectively refining the citation network. This iterative sampling approach allowed the collection to focus on high-relevance papers while filtering out unrelated or marginally related works.

The incorporation of PTM ensured that variations in the initial seed set did not significantly affect the identification of influential papers thereby improving the robustness and reliability of the collection process [15]. Citation analysis further enriched the dataset by identifying key relationships within the network enabling the discovery of seminal works across subfields.

To assess representativeness we employed terminology saturation analysis [16]. Automated term extraction techniques were used to evaluate the conceptual coverage within the selected literature. Terminology saturation the point at

which additional papers no longer introduce new concepts served as an objective indicator that the collection had achieved sufficient breadth and depth to capture the State-of-the-Art.

The iterative nature of this approach enabled the formation of a literature set that is both relevant containing papers directly influencing or informing the planned research and representative covering all significant developments within the chosen domain.

2.2 Vision-Language Alignment Models

The introduction of CLIP [5] marked a pivotal advancement in aligning visual and textual modalities through contrastive learning. By jointly training image and text encoders to map inputs into a shared embedding space, CLIP enables zero-shot generalization across tasks such as image classification, retrieval, and captioning without requiring task-specific fine-tuning.

Contrastive learning has become a cornerstone of multimodal representation learning. Methods like SimCLR [17] and SimCSE [18] demonstrated that semantic similarity can be effectively learned through augmentation-based objectives in both vision and language domains. Building on this, UNITER [19] introduced multimodal transformers with cross-attention, enabling fine-grained alignment between image regions and text.

SigLIP [6] advanced this paradigm by replacing the softmax contrastive loss with a sigmoid-based pairwise loss, which simplifies optimization and improves scalability by eliminating the need for global similarity normalization.

Recent extensions such as mPLUG [12] and mPLUG-DocOwl [20] introduced architectural innovations like cross-modal skip connections, enhancing representation quality for structured document understanding and information extraction from visually rich layouts.

2.3 Document Understanding Models

Effective document understanding requires modeling both textual content and visual layout. LayoutLM [8] pioneered this direction by introducing spatial embeddings derived from OCR outputs. LayoutLMv2 [9] extended this by integrating visual features and task-specific pretraining objectives improving performance on structured tasks such as form parsing and table extraction. LayoutLMv3 [21] further unified image and text pretraining using masked language and image modeling, enhancing cross-modal reasoning.

DocFormer [10] proposed a fully transformer-based architecture that jointly encodes text, vision, and layout features without relying on OCR improving robustness to noisy or handwritten inputs. DiT [22] adopted self-supervised pretraining on document images achieving strong performance across diverse formats. Most notably, the OCR-Free Document Understanding Transformer [7] eliminated OCR entirely directly processing raw pixels as an advantage in low-resource or degraded settings.

The DUE benchmark [23] provides a standardized framework for evaluating document understanding systems across tasks such as classification, text extraction, and layout analysis.

2.4 Vision-Language Pretraining Models

Multimodal pretraining has emerged as a foundation for general-purpose vision-language models. PaliGemma [1, 2] exemplifies this trend, offering strong generalization across tasks including VDU and VQA. Pix2Struct [3] focused on layout comprehension, pretraining on structured web and document screenshots for table and form extraction.

To address multilingual document scenarios, SynthDoc [24] introduced synthetic bilingual data to improve model robustness across underrepresented languages and layouts. More recently, Qwen2-VL [4] introduced Naive Dynamic Resolution to efficiently process variable image sizes along with Multimodal Rotary Position Embeddings (M-RoPE) for improved alignment of visual and textual modalities. These architectural innovations enhance model adaptability across complex document layouts and multilingual content.

3 Problem Setting and Approach to Solution

3.1 Problem Statement

Contemporary document understanding systems remain heavily dependent on OCR pipelines. One notable example is the *LIBRARIA*[3] project which, in collaboration with libraries, historical archives and scientific institutions in Ukraine and abroad, aims to digitize and provide online access to an extensive collection of Ukrainian historical newspapers. However, classical OCR methods exhibit significant limitations in terms of robustness and accuracy. In particular, OCR performance tends to degrade in the presence of noise, handwritten content, low-resolution scans or multilingual text-factors that are common in real-world documents [7].

To assess the extent of these limitations we conducted an evaluation of the open-source OCR engine *Tesseract*[4], developed and maintained by Google. Our evaluation was performed on a 10% subset of our dataset which contains documents in the Ukrainian language.

Tesseract was benchmarked under two conditions: (1) OCR applied to localized text regions (OCR Text) and (2) OCR applied to the entire document image (OCR Doc). We measured performance using two standard metrics: *Character Error Rate (CER)* and *Word Error Rate (WER)*.

Our results (Table 1) indicate that while CER remain relatively low across most categories, excluding handwritten inputs, WER are significantly higher in particular for handwritten and full-document OCR. This disparity underscores

[3] https://libraria.ua/
[4] https://github.com/tesseract-ocr/tesseract

Table 1. Tesseract OCR Performance

Metric	OCR Text	OCR Doc
Mean CER	9.81%	30.82%
Mean CER (print)	3.68%	28.08%
Mean CER (hand)	20.41%	36.71%
Mean CER (scan)	4.30%	27.17%
Mean WER	58.07%	212.13%
Mean WER (print)	14.30%	192.24%
Mean WER (hand)	140.03%	265.89%
Mean WER (scan)	12.16%	174.26%

a key limitation of traditional OCR systems: their reliance on purely visual character recognition without contextual reasoning. Minor character-level errors can propagate and distort entire words severely impacting downstream tasks.

In addition to performance degradation on noisy or handwritten inputs OCR systems often fail when processing documents with non-standard syntax such as mathematical expressions, structured elements or complex layout features. Designed primarily for plain-text extraction conventional OCR struggles with superscripts, subscripts, equations and specialized notations further limiting its applicability in real-world document understanding scenarios.

3.2 VLM as OCR Tool

VLMs take an end-to-end approach learning to directly map images to structured textual representations without an intermediate text detection step. A typical VLM architecture (Figure 1) consists of visual encoder, cross-modal connector and auto-regressive Large Language Model (LLM) [12, 4].

Vision Encoder. This module extracts visual features from the input image and aligns them with textual representations. Models such as CLIP and SigLIP are commonly used for this purpose, as they are trained on large-scale vision-language datasets and can effectively capture both textual and visual context.

Cross-Modal Connector. The extracted visual features need to be integrated with language representations before being processed by an LLM. The cross-modal connector serves this purpose by aligning, fusing and transforming the visual embeddings into a format compatible with LLM [12, 20]. This can be achieved through linear projection layers, cross-attention mechanisms or multi-modal fusion transformers.

Auto-regressive LLM. Once the visual features are extracted, LLM (like GPT-style transformers) processes them to generate structured text output. This component ensures that the generated text preserves formatting, mathematical notation or other complex structures present in the original document [12, 20].

By leveraging this end-to-end approach, VLMs provide several advantages: (1) better handling of complex layouts (e.g., multi-column documents, tables, equations), (2) improved robustness to noisy or distorted text and (3) higher-

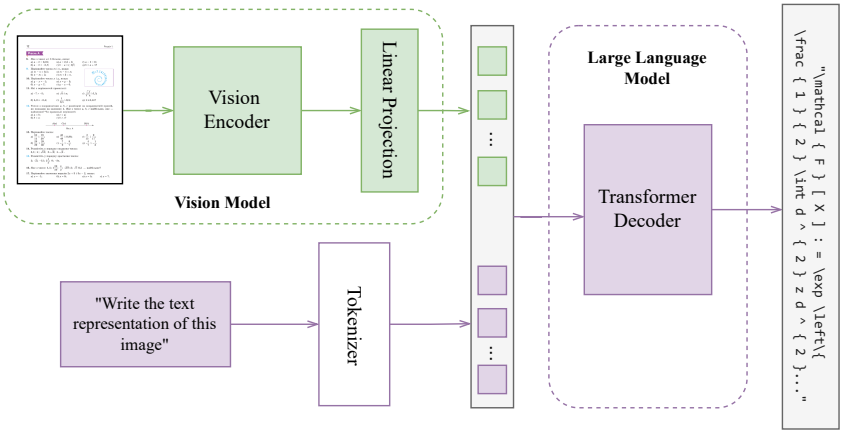

Fig. 1. VLM Architecture

level language understanding that allows for contextual extraction and summarization.

3.3 A Semi-Synthetic Dataset

We introduce a semi-synthetic dataset of approximately 5,000 visually structured documents, with a particular focus on the Ukrainian language. Each document is generated using a Retrieval-Augmented Generation (RAG) pipeline: textual content is retrieved from Ukrainian Wikipedia[5] by category, then assembled via a structured prompt using a large language model (LLM) into coherent document sections including paragraphs, tables and charts.

To increase structural diversity we incorporate HTML tables sampled from the PubTables-1M dataset [25]. The generated content is embedded into randomly selected HTML templates and styled with varying fonts (including handwritten styles), colors and background textures. Documents are rendered as images to simulate real-world appearance.

To enhance visual realism we apply controlled degradation techniques including Gaussian blur, perspective warping and low-intensity noise. Additional distortions, such as sinusoidal warping and contrast/brightness adjustments, simulate scanning artifacts and layout imperfections.

Each document is paired with rich metadata including language, category, title, rendering style ("print", "hand" or "scan"), template markup and grounding information (element type, content, bounding boxes and reading order).

For evaluation the dataset is partitioned using an 80/20 split: 1,000 images are reserved for benchmarking and the remaining 4,000 are allocated for visual instruction tuning and pretraining on benchmark-style prompts.

[5] https://uk.wikipedia.org/w/api.php

3.4 Methodology and Evaluation

We evaluate Document AI models using a structured benchmarking pipeline comprising three stages: (1) prompting the model with task-specific queries, (2) parsing outputs into standardized formats and (3) computing evaluation metrics. To ensure prompt consistency and output validation we employ the `pydantic`[6] library to define structured data schemas. These schemas enforce type validation and facilitate reliable parsing for metric computation.

Each benchmark task, including OCR accuracy, document classification, table structure extraction and layout analysis, is supported by tailored prompts designed to elicit outputs in the appropriate structured format for downstream evaluation.

OCR Benchmark evaluates the model's ability to extract text from document images at both character and word levels.

Character Error Rate (CER) measures character-level transcription accuracy and is defined as:

$$CER = \frac{S + D + I}{N_{char}} \tag{1}$$

where S, D and I are the number of substitutions, deletions, insertions and N_{char} is the total number of characters in the ground truth.

Word Error Rate (WER) evaluates word-level accuracy using the same edit operations:

$$WER = \frac{S + D + I}{N_{word}} \tag{2}$$

where N_{word} is the total number of words in the ground truth. Both metrics rely on Levenshtein distance to quantify the deviation from the reference transcription.

Document Classification. This task involves assigning Ukrainian Wikipedia-sourced documents to predefined thematic categories. The VLM is prompted to classify each document based on its content.

Performance is measured using *accuracy*, defined as:

$$Accuracy = \frac{CorrectPredictions}{N} \tag{3}$$

where $CorrectPredictions$ is the number of documents correctly classified and N is the total number of evaluated samples.

Table Structure Extraction involves recovering the structural layout of tables in markup formats such as HTML or XML.

Evaluation is based on *Tree Edit Distance-based Similarity (TEDS)* [26], which measures structural accuracy as:

$$TEDS(T_a, T_b) = 1 - \frac{EditDist(T_a, T_b)}{\max(|T_a|, |T_b|)} \tag{4}$$

[6] https://docs.pydantic.dev/latest/

where $EditDist(T_a, T_b)$ is the tree edit distance between predicted and reference trees and $|T|$ denotes the number of nodes.

TEDS incorporates a normalized Levenshtein similarity to allow partial matching of cell content, enabling fine-grained evaluation of both structure and textual fidelity [27].

Document Layout Analysis focuses on extracting structured elements from documents in reading order (top-to-bottom, left-to-right) using VLMs. Evaluation compares the predicted sequence of elements to ground truth, considering both content and order.

We propose *Mean Ordered Sequence Similarity (MOSS)* to assess layout accuracy, extending traditional text similarity by incorporating sequence constraints:

$$MOSS(T_p, T_g) = \frac{1}{N} \sum_{i=1}^{N} S(T_p^i, T_g^i) \tag{5}$$

where T_p^i and T_g^i are the predicted and ground truth sequences for document i and $S(\cdot)$ is a similarity score. MOSS uses TF-IDF vectorization over character-level n-grams and cosine similarity to evaluate sequence alignment.

The metric is (1) sensitive to element order, (2) robust to minor OCR noise and (3) scalable for large document corpora.

4 Early Results and Discussion

4.1 Model Selection

While comprehensive quantitative results are still pending at the time of writing, we base our initial model selection on available performance indicators and relevant considerations. Our primary focus is on multilingual VLMs with support for the Ukrainian language particularly in the context of image text-to-text understanding tasks.

We distinguish between two main categories of such models. The first category comprises high-performance Tier 1 models accessible via provider APIs, such as Gemini 2.0 Flash[7] and Mistral OCR[8]. Both models have been benchmarked on Ukrainian-language documents and demonstrated outstanding OCR performance.[9]

Table 2. Tier 1 Model Performance on Ukrainian Language

Language	Gemini-2.0-Flash	Mistral OCR
uk	96.70%	99.29%

[7] https://deepmind.google/technologies/gemini/flash/
[8] https://docs.mistral.ai/capabilities/document/
[9] https://mistral.ai/news/mistral-ocr

The second category includes open-source VLMs available through platforms such as Hugging Face[10]. During the early stages of our study, we conducted a review of available models to identify those best suited for Ukrainian image text-to-text understanding across the selected benchmarking tasks.

Based on these observations we selected the following models for further evaluation: **Qwen2 VL**[11] and **Phi-3.5 Vision**[12]. Both models demonstrated strong initial performance in processing Ukrainian-language documents. Their pretrained capabilities offer a robust foundation for visual instruction tuning and benchmarking in the context of document understanding tasks.

Given the rapid pace of advancement in the field we anticipate the emergence of new more capable models. As such the list of selected VLMs remains subject to revision over the course of our study.

4.2 Research Hypotheses

This study investigates the capabilities of VLMs for document understanding with a focus on Ukrainian-language documents. We formulate the following hypotheses to guide our evaluation:

1. *Effectiveness of OCR-Free VLMs.* Modern OCR-free VLMs can outperform traditional OCR systems particularly on noisy, handwritten or low-quality documents where OCR typically fails due to lack of contextual reasoning or layout robustness.
2. *Competitiveness of Open-Source Models.* Open-source OCR-free VLMs (e.g., those available via Hugging Face) can achieve performance comparable to Tier 1 proprietary models (e.g., Gemini-2.0-Flash, Mistral OCR) across OCR, classification and layout tasks.
3. *Structural Understanding.* VLMs can accurately extract document structures, including reading order and table layouts as measured by sequence and tree-based metrics such as MOSS and TEDS.
4. *Classification Robustness.* VLMs can serve as effective classifiers for Ukrainian-language documents requiring little or no task-specific fine-tuning.
5. *Resilience to Image Degradation.* VLMs demonstrate greater robustness than traditional OCR under challenging image conditions such as low resolution, noise and visual distortion.

These hypotheses aim to assess whether VLMs can serve as viable OCR-free alternatives for document processing, particularly in low-resource and degraded settings.

[10] https://huggingface.co/
[11] https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct
[12] https://huggingface.co/microsoft/Phi-3.5-vision-instruct

5 Summary and Future Work

At the time of writing we have completed a comprehensive literature review, defined key research hypotheses and selected candidate models for evaluation on core document understanding tasks. These include OCR benchmarking, document classification, table structure extraction and layout analysis using Ukrainian-language document images. The remaining phases involve systematic benchmarking, comparative analysis and synthesis of results. An overview of the research timeline is shown in Figure 2.

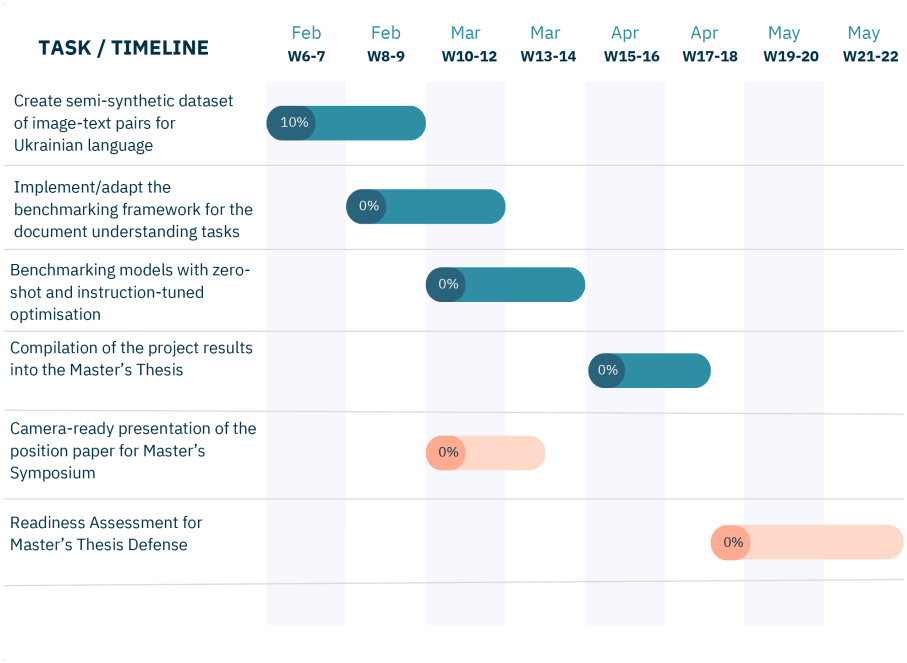

Fig. 2. Research Roadmap

This position paper presents the overall research plan detailing the objectives, dataset design, model selection and evaluation methodology. We introduced a semi-synthetic dataset tailored for Ukrainian, outlined the benchmarking tasks and described both zero-shot and instruction-tuned evaluation protocols. This structured framework enables a rigorous assessment of VLMs and their applicability to document understanding in low-resource language settings.

References

1. Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.
2. Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*, 2024.
3. Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023.
4. Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
5. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
6. Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.
7. Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022.
8. Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1192–1200, 2020.
9. Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. Layoutlmv2: Multimodal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*, 2020.
10. Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1003, 2021.
11. Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. Unidoc: Unified pretraining framework for document understanding. *Advances in Neural Information Processing Systems*, 34:39–50, 2021.
12. Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022.

13. A Conneau. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.
14. Hennadii Dobrovolskyi and Nataliya Keberle. Collecting the seminal scientific abstracts with topic modelling, snowball sampling and citation analysis. In *International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications*, 2018.
15. Hennadii Dobrovolskyi, Nataliya Keberle, and Olga Todoriko. Probabilistic topic modelling for controlled snowball sampling in citation network collection. In Przemysław Różewski and Christoph Lange, editors, *Knowledge Engineering and Semantic Web*, pages 85–100, Cham, 2017. Springer International Publishing.
16. Victoria Kosa, David Chaves-Fraga, Hennadii Dobrovolskyi, Egor Fedorenko, and Vadim Ermolayev. Optimizing automated term extraction for terminological saturation measurement. In *International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications*, 2019.
17. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
18. Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.
19. Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.
20. Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023.
21. Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091, 2022.
22. Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. Dit: Self-supervised pre-training for document image transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3530–3539, 2022.
23. Ł ukasz Borchmann, MichałPietruszka, Tomasz Stanislawek, Dawid Jurkiewicz, MichałTurski, Karolina Szyndler, and Filip Graliński. Due: End-to-end document understanding benchmark. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.
24. Chuanghao Ding, Xuejing Liu, Wei Tang, Juan Li, Xiaoliang Wang, Rui Zhao, Cam-Tu Nguyen, and Fei Tan. Synthdoc: Bilingual documents synthesis for visual document understanding. In *Proceedings of the 2nd Workshop on Large Generative Models Meet Multimodal Applications*, pages 16–25, 2024.
25. Brandon Smock, Rohith Pesala, and Robin Abraham. Pubtables-1m: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4634–4642, 2022.
26. Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation. In *European conference on computer vision*, pages 564–580. Springer, 2020.
27. Brandon Smock, Rohith Pesala, and Robin Abraham. Grits: Grid table similarity metric for table structure recognition. In *International Conference on Document Analysis and Recognition*, pages 535–549. Springer, 2023.