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 visual question answering (VQA). 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

The field of multimodal learning has experienced significant progress particularly in integrating CV and NLP to enhance cross-modal comprehension. Models such as CLIP [5] and SigLIP [6] have demonstrated impressive results in aligning visual and textual modalities enabling tasks like image-text matching and text-based image retrieval.

Despite these advances the challenges persist in extending such capabilities to structured, visually rich documents. Current approaches often rely heavily on OCR pipelines which can degrade performance when faced with noisy, handwritten or low-quality documents [7]. While models such as LayoutLM [8–10] and DocFormer [11] integrate text and layout information they struggle to generalize across diverse document types and layouts especially in multilingual and low-resource scenarios [12, 13].

These limitations highlight the need for scalable architectures that unify text, vision and layout understanding. Addressing this gap is crucial for enabling applications such as automated document classification, VQA and semantic search in real-world domains like legal, financial and healthcare documents.

Visual-to-text comprehension offers opportunities to create systems that can bridge the gap between textual and visual data enabling more natural and human-like understanding of information[8, 14].

The ability to integrate visual and textual information has already demonstrated its utility in various domains:

- Education and Social: Tools for generating accessible content such as image captioning for visually impaired users [15, 16].
- E-commerce and Marketing: Systems for automated cataloging, smart tagging and content-based search [17].
- Scientific and Legal Analysis: Solutions for parsing research papers, contracts and reports with complex layouts [18, 19].
- Healthcare: Automated processing of medical forms, reports and patient data for faster analysis and decision-making [11, 20].

As demonstrated by recent advancements in multimodal frameworks like LayoutLMv2 [9] and OCR-free approaches [7] the integration of layout-aware features and end-to-end architectures holds immense potential for document understanding tasks. These developments suggest a promising direction for leveraging cross-modal learning to tackle real-world challenges in structured document processing.

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 [21].

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 [22]. 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 [22]. 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 [23]. 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 significant milestone in aligning visual and textual representations. CLIP leverages a contrastive learning framework to pretrain image and text encoders enabling them to map visual and textual inputs into a shared embedding space. This design facilitates zero-shot learning for a variety of tasks including image classification, text-to-image retrieval and caption generation. Its ability to generalize across tasks without explicit task-specific training made it a foundational model for subsequent multimodal architectures.

Building upon this framework UNITER (Universal Image-Text Representation) [24] introduced multimodal transformers to combine visual and textual inputs through cross-attention mechanisms. UNITER succeeded in tasks requiring fine-grained alignment between vision and language.

Another prominent representative - the simple pairwise Sigmoid Loss for Language-Image Pre-training (SigLIP)[6] model, introduced the sigmoid loss that replace softmax-based contrastive loss. The sigmoid loss operates solely on image-text pairs and does not require a global view of the pairwise similarities for normalization. This allows further scaling up the batch size while also performing better at smaller batch size.

A more recent development, mPLUG (Multimodal Pre-trained Language Understanding and Generation) [12], extended vision-language learning by incorporating cross-modal skip connections to improve efficiency and representation quality. Its successor, mPLUG-DocOwl [25], further refined this approach for structured document understanding focusing on information extraction from visually rich layouts.

2.3 Document Understanding Models

Understanding structured documents requires capturing both visual and textual features along with their spatial relationships. Early approaches such as LayoutLM [8] introduced layout-aware transformers that incorporate spatial embeddings derived from OCR pipelines. This approach was extended by LayoutLMv2 [9] which combined visual and textual features using pretraining objectives tailored for multimodal alignment. These improvements enhanced performance in tasks like form parsing, invoice processing and table recognition.

Later LayoutLMv3 [10] further unified text and image pretraining using masking strategies and enabling stronger cross-modal reasoning. These advances demonstrated the importance of modeling layout information alongside text and visual features.

An alternative approach was proposed by DocFormer [11] which adopted a fully transformer-based design to jointly process textual, visual and spatial features. Unlike earlier models DocFormer avoided dependence on OCR pipelines providing end-to-end learning capabilities that improved robustness to noisy or handwritten data.

DiT (Document Image Transformer) [20] focused on self-supervised pretraining tailored for document analysis. Its design made it particularly effective for handling diverse document formats ranging from scanned PDFs to structured templates.

And recently OCR-Free Document Understanding Transformer [7] eliminated reliance on OCR systems entirely by directly processing raw pixel inputs. This approach proved particularly effective in scenarios involving noisy, low-resolution or multilingual documents where OCR pipelines typically fail.

2.4 Contrastive Learning and Fine-tuning

Contrastive learning has played a critical role in multimodal systems as it enables representation learning without requiring labeled data. SimCSE (Simple Contrastive Sentence Embeddings) [26] demonstrated that contrastive loss could effectively capture semantic similarity between sentences by leveraging dropout as minimal augmentation during training. The same as SimCLR (Simple Framework for Contrastive Learning of Visual Representations) [27] aligned visual representation embeddings by maximizing agreement between augmented views of the same image.

Both approaches highlight the power of contrastive objectives in learning meaningful representations across modalities and lay the foundation for bridging visual and textual embeddings in multimodal architectures. These methods have influenced subsequent frameworks that integrate contrastive techniques to improve cross-modal alignment and scalability.

Expanding contrastive learning approach MoE-LoRA (Mixture of Experts with Low-Rank Adaptation) [28] integrated contrastive learning with parameter-efficient fine-tuning. It demonstrated scalability and adaptability especially in resource-constrained environments.

Vision-specific contrastive learning frameworks, such as BEiT (BERT Pre-training of Image Transformers) [29] and XCiT (Cross-Covariance Image Transformers) [30], adapted transformers for visual inputs. These models pre-trained visual encoders using masked image modeling, providing rich representations for multimodal architectures.

Self-supervised learning approaches like SimCLR [27] and MoCo [31] established foundational techniques for visual representation learning. Their influence extended to multimodal designs by demonstrating the value of contrastive loss in aligning visual and textual features.

2.5 Multimodal Pre-training Frameworks

Multimodal pretraining frameworks focus on transfer learning and generalization across diverse tasks. PaliGemma [1, 2] exemplified this approach by leveraging large-scale pretraining to create versatile models for visual-language tasks. Its adaptability made it suitable for both structured document processing and VQA.

Similar Pix2Struct [3] focused on layout understanding and pretraining on structured screenshots and tables. Its ability to extract hierarchical information demonstrated its effectiveness for tasks that involve forms and data grids.

SynthDoc [32] addressed the need for multilingual document processing by synthesizing bilingual datasets. This framework extended document understanding capabilities to languages and layouts that are often underrepresented in current datasets.

Recently Qwen2-VL [4] model series redefined the traditional fixed-resolution approach in visual processing by implementing the Naive Dynamic Resolution mechanism and enabling the model to dynamically process images of varying resolutions into different numbers of visual tokens. This method allows for more efficient and accurate visual representations closely aligning with human perceptual processes. Qwen2-VL integrates Multimodal Rotary Position Embedding (M-RoPE)[4] facilitating effective fusion of positional information across text, images and videos.

2.6 VQA Models

Tasks like VQA require models to interpret both textual and visual inputs to provide meaningful answers. DocVQA [18] served as an early benchmark for evaluating VQA systems on document images and exposed the challenges posed by complex layouts and non-standard text placements.

VisualMRC [19] extended traditional machine reading comprehension (MRC) approaches to visually-rich documents. It highlighted the importance of jointly reasoning over textual content, visual structure and layout information to accurately interpret and answer questions based on complex document formats.

Meanwhile, TextCaps [15] focused on generating captions enriched with textual information extracted directly from images. This approach highlighted the importance of integrating OCR outputs into downstream tasks, particularly for accessibility applications.

2.7 Datasets

The development of multimodal models has been supported by datasets specifically designed to evaluate vision-language tasks, document understanding and VQA.

DocVQA [18] is one of the foundational datasets for document-based VQA. It consists of scanned documents where models are required to answer questions by reasoning over textual content and layout information. DocVQA emphasizes challenges related to structured text and complex layouts and makes it a benchmark for evaluating document understanding systems.

VisualMRC [19] extends machine reading comprehension (MRC) to visually rich documents. It focuses on tasks requiring models to jointly interpret textual content, layout and visual elements by testing the ability of models to process structured data effectively.

TextCaps [15] addresses image captioning with text reading comprehension. It highlights the integration of OCR results into generated captions and enables models to describe images containing textual elements. This dataset is particularly useful for accessibility applications and content summarization.

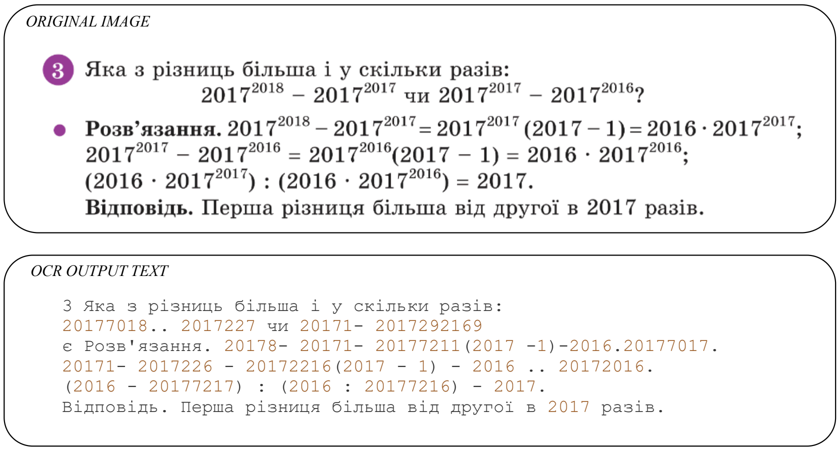
SynthDoc [32] introduces synthetic bilingual documents for visual document understanding. It is designed to address multilingual and multimodal scenarios by providing synthetic annotations that test model generalization across different languages and layouts.

The DUE (Document Understanding Evaluation) benchmark [33] evaluates end-to-end document understanding models and includes tasks such as text extraction, classification, and layout analysis. It offers a standardized framework for benchmarking multimodal systems.

3 Problem Setting and Approach to Solution

While significant progress has been made with the first genaration of VLM models such as LayoutLM [8–10] and DocFormer [11] the most current systems remain heavily reliant on OCR pipelines. This dependency introduces critical weaknesses when dealing with noisy, handwritten, low-resolution or multilingual documents as OCR quality often degrades under such conditions [7].

OCR may fail not only when there are quality issues but also when the text contains special syntax, mathematical expressions or structured elements as demonstrated in Figure 1. These failures occur because traditional OCR systems are primarily designed for plain text extraction and often struggle with complex layouts, superscripts, subscripts, equations or specialized notation.



ORIGINAL IMAGE

3 Яка з різниць більша і у скільки разів:

$$2017^{2018} - 2017^{2017} \text{ чи } 2017^{2017} - 2017^{2016}?$$

● **Розв'язання.** $2017^{2018} - 2017^{2017} = 2017^{2017}(2017 - 1) = 2016 \cdot 2017^{2017}$;
$2017^{2017} - 2017^{2016} = 2017^{2016}(2017 - 1) = 2016 \cdot 2017^{2016}$;
$(2016 \cdot 2017^{2017}) : (2016 \cdot 2017^{2016}) = 2017$.
Відповідь. Перша різниця більша від другої в 2017 разів.

OCR OUTPUT TEXT

```
3 Яка з різниць більша і у скільки разів:
20177018.. 2017227 чи 20171- 2017292169
є Розв'язання. 20178- 20171- 20177211(2017 -1)-2016.20177017.
20171- 2017226 - 20172216(2017 - 1) - 2016 .. 20172016.
(2016 - 20177217) : (2016 : 20177216) - 2017.
Відповідь. Перша різниця більша від другої в 2017 разів.
```

Fig. 1. OCR Error Output

Moreover, conventional OCR lacks post-processing capabilities for higher-level text understanding tasks such as recognition and retrieval, summarization and captioning of textual information. OCR systems typically output raw text without semantic interpretation, making it difficult to directly use the extracted content for downstream applications.

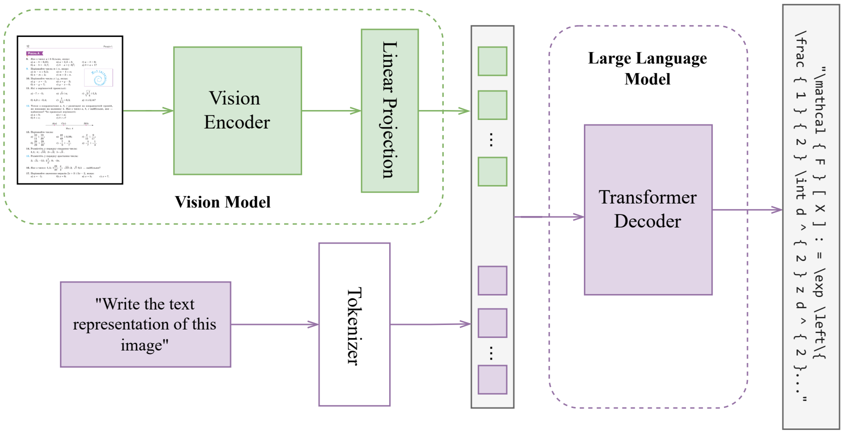
To overcome these limitations OCR-free VLMs have emerged as a more robust solution. Unlike traditional OCR, which explicitly detects and extracts text regions before processing, OCR-free VLMs take an end-to-end approach, learning to directly map images to structured textual representations without an intermediate text detection step.

A typical VLM architecture (Figure 2) 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, 25]. This can be achieved through linear projection layers, cross-attention mechanisms or multimodal fusion transformers.

Auto-regressive LLM. Once the visual features are extracted, an 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, 25].



```
"\mathcal { F } [ X ] : = \exp \left\{
\frac { 1 } { 2 } \int d ^ { 2 } z d ^ { 2 }..."
```

Fig. 2. VLM Architecture

By leveraging this end-to-end approach, OCR-free VLMs provide several advantages:

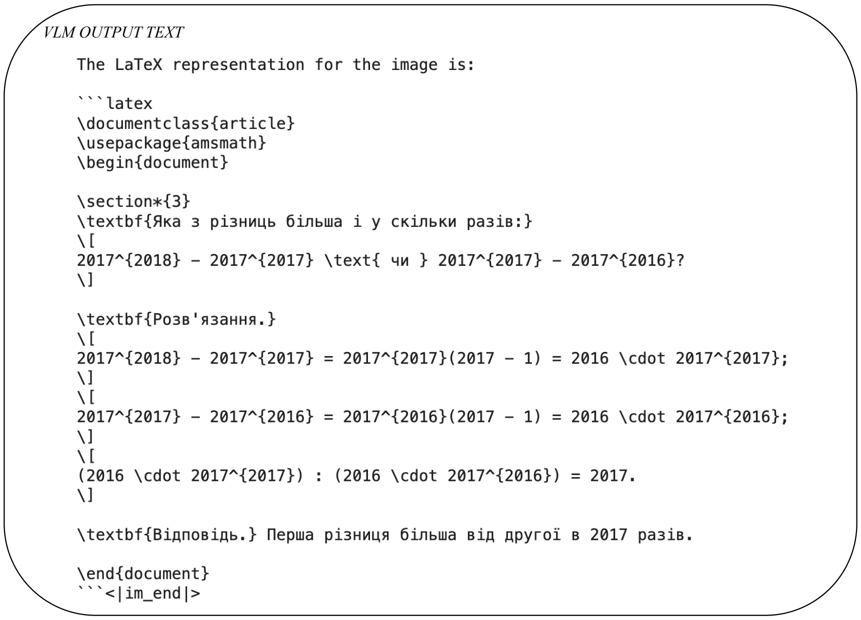
- **better handling of complex layouts** (e.g., multi-column documents, tables, equations).
- **improved robustness** to noisy or distorted text that traditional OCR might misinterpret.
- **higher-level language understanding** that allows for contextual extraction, summarization and interactive question-answering directly from images.

These advancements make OCR-free VLMs a more powerful alternative to traditional OCR, especially in domains where preserving structure, meaning and formatting is critical, like in academic/scientific works, legal documents and handwritten manuscripts, etc.

4 Early Results and Discussions

During the early stages of the research we conducted a review of VLMs to identify those best suited for multilingual text understanding across various downstream document AI tasks. To ensure a fair selection we performed a zero-shot evaluation testing of each model without fine-tuning.

Our zero-shot prompting of the VLM models indicates that they can effectively handle OCR tasks particularly in cases where traditional OCR struggles. Figure 3 demonstrates the text output of the VLM on the same example where classical OCR produced incorrect results (Figure 1).



VLM OUTPUT TEXT

~~~
The LaTeX representation for the image is:

```latex
\documentclass{article}
\usepackage{amsmath}
\begin{document}

\section*{3}
\textbf{Яка з різниць більша і у скільки разів:}
\[
2017^{2018} - 2017^{2017} \text{ чи } 2017^{2017} - 2017^{2016}?
\]

\textbf{Розв'язання.}
\[
2017^{2018} - 2017^{2017} = 2017^{2017}(2017 - 1) = 2016 \cdot 2017^{2017};
\]
\[
2017^{2017} - 2017^{2016} = 2017^{2016}(2017 - 1) = 2016 \cdot 2017^{2016};
\]
\[
(2016 \cdot 2017^{2017}) : (2016 \cdot 2017^{2016}) = 2017.
\]

\textbf{Відповідь.} Перша різниця більша від другої в 2017 разів.

\end{document}
```<|im_end|>
~~~

Fig. 3. Qwen2 VL Text Output

From this evaluation we selected models: Qwen2 VL[3], Llama3.2 Vision[4], Phi-3.5 Vision[5] and Eagle2[6] that exhibited promising initial performance in processing multilingual text. While none were perfect in zero-shot mode, their pretrained capabilities provided a solid foundation for instruction-tuning and benchmarking on document AI tasks for Ukrainian language.

The selected models are detailed in Table 1 highlighting their architectural strengths and suitability for OCR-free document understanding in Ukrainian. Given the rapid advancements in AI, we reserve the right to update the list of selected VLMs if superior models emerge during the course of our research.

[3] https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct

[4] https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct

[5] https://huggingface.co/microsoft/Phi-3.5-vision-instruct

[6] https://huggingface.co/nvidia/Eagle2-9B

Table 1. VLMs for Multilingual Visual Text Comprehension

Model	**Description**
Qwen2 VL	A multimodal model excelling in document QA, OCR-free text recognition and multilingual understanding. Supports long video analysis and autonomous agent tasks [4].
Phi-3.5 Vision	A lightweight multimodal model designed for commercial and research applications. It is optimized for memory-constrained environments, low-latency scenarios and tasks requiring general image understanding, OCR, chart/table interpretation, multiple image comparison and video clip summarization.[7]
Llama3.2 Vision	Built on Llama 3.1, an autoregressive transformer-based model, Llama 3.2-Vision integrates a separately trained vision adapter with cross-attention layers to feed image encoder outputs into the LLM. Tuned versions use supervised fine-tuning (SFT) and reinforcement learning with human feedback RLHF for alignment with human preferences in helpfulness and safety.[8]
Eagle2	As part of the Eagle2 series, it focuses on data-centric post-training strategies, aiming to enhance reproducibility and innovation in open-source VLMs. The model supports 4K high-definition input, long-context video analysis, and grounding tasks. It has demonstrated strong performance across various benchmarks, including DocVQA, ChartQA, InfoVQA, TextVQA, and MathVista.[9]

5 Summary and Future Work

While our initial evaluations demonstrate the robustness of VLMs in OCR tasks it remains unclear which models prevail in specific document AI tasks. To address this we expand the scope of evaluation by defining a broader set of document understanding tasks and systematically benchmarking the selected models.

Our approach involves generating semi-synthetic datasets of Ukrainian text images covering both printed and handwritten text. This allows for a controlled and scalable evaluation framework. We will benchmark the selected models (Table 1) with a primary focus on prompt optimization and / or instruction tuning using Low-Rank Adaptation (LoRA) [34] for downstream document AI tasks as outlined in Table 2.

The goal of this study is to assess the performance of the image-text-to-text VLMs on Ukrainian text-based tasks, identify their strengths and limitations, and analyze technical aspects that impact their effectiveness. We anticipate that these findings will offer valuable insights into model adaptation strategies, potential enhancements for low-resource languages and guidelines for future OCR-free document processing pipelines.

Table 2. Document Understanding Tasks and Metrics for Image-Text-to-Text Models

Task	**Typical Metrics**
Optical Character Recognition (OCR)	Character Error Rate (CER), Word Error Rate (WER)
Key Information Extraction (KIE)	Accuracy, F1-score, Edit Distance
Document Visual Question Answering (DocVQA)	Average Normalized Levenshtein Similarity (ANLS), Exact Match (EM), ROUGE
Document Classification (via Prompting)	Zero-shot Accuracy, F1-score
Table Structure Extraction	Tree Edit Distance (TED), Cell-level F1-score
Layout-to-Text Conversion	BLEU, ROUGE, Structural Similarity Metrics

At the time of writing this position paper, we have completed the literature review, defined our research objectives and conducted the initial model selection for benchmarking on downstream tasks involving Ukrainian text images. Figure 4 below outlines the remaining steps required to complete the project.

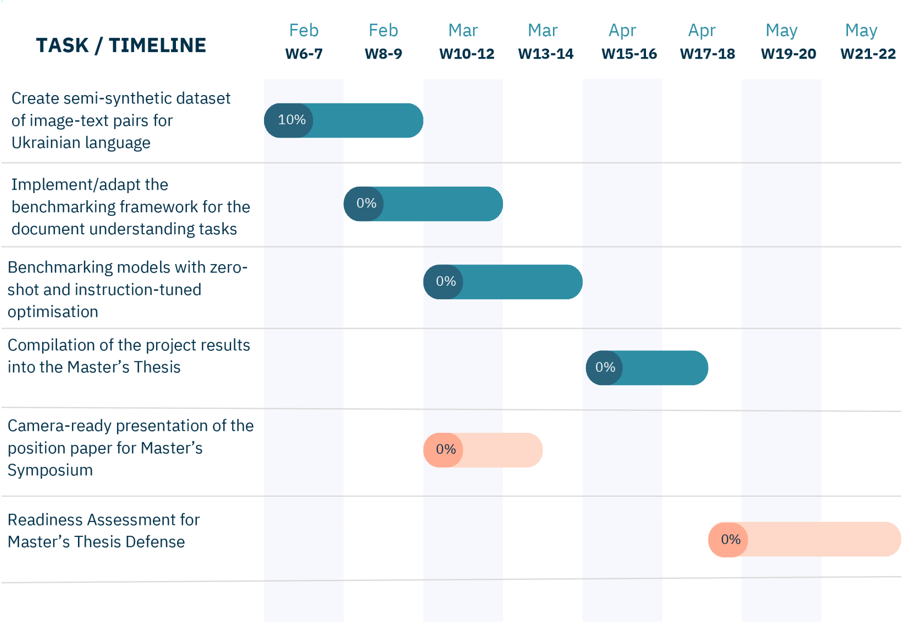


Fig. 4. Research Roadmap

In this paper we have presented a comprehensive research plan detailing the scope, objectives, methodology and timeline of our study. We have defined the research objectives, introduced the semi-synthetic dataset of Ukrainian text images and selected the VLMs for benchmarking. As well as, we have outlined the planned experiments, including zero-shot and instruction-tuned model evaluations and described the benchmarking framework for document understanding tasks. This structured approach ensures a systematic evaluation of model performance providing insights into their capabilities and limitations for OCR-free document processing in low-resource languages.

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. 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.
11. 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.
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. 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.
15. Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020.
16. Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.
17. Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Michał Turski, and Filip Graliński. Lambert: layout-aware language modeling for information extraction. In *International conference on document analysis and recognition*, pages 532–547. Springer, 2021.
18. Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.
19. Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13878–13888, 2021.
20. 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.
21. 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.
22. 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.
23. 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.

24. 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.
25. 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.
26. Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.
27. 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.
28. Tongxu Luo, Jiahe Lei, Fangyu Lei, Weihao Liu, Shizhu He, Jun Zhao, and Kang Liu. Moelora: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models. *arXiv preprint arXiv:2402.12851*, 2024.
29. Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
30. Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. *Advances in neural information processing systems*, 34:20014–20027, 2021.
31. Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.
32. 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.
33. 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.
34. Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.