
UAET: A Unified Adaptive-Expert Transformer for Efficient and Accurate Intelligent Document Processing

Debadatta S. Roy¹, Sreeza Chowdhury², and Bijoyini Bagchi³

*Computer Science Engineering, Institute of Engineering and Management,
debadatta.sinharoy2023@edu.in, chowdhurysreeza@gmail.com, Bijoyini.Bagchi@iem.edu.in*

Abstract.

Intelligent Document Processing(IDP) must make sure the results are correct but not too costly to use. Big vision-language models can think well. Cost a lot when used in real time. Small models have trouble with layouts. UAET tries to solve these problems by keeping the parts when combining different types of data. It uses a shared Transformer that works well together and sends types of information to different experts. The design makes sure that the overall thinking is still good but sends text parts to a Phi-style expert and visual parts to a strong Qwen-style expert. On one NVIDIA T4 GPU using 4-bit quantization, UAET's routing and training behave as intended, but its F1 scores currently fall well short of established baselines, reaching only 8.06% F1 on SROIE and 6.26% F1 on FUNSD; a 100% ANLS on InfographicVQA is also reported but is most likely due to benchmark contamination rather than genuine performance.

Keywords. Intelligent Document Processing (IDP), Mixture-of-Experts (MoE), document understanding, adaptive routing, Transformer architecture.

1. INTRODUCTION

Intelligent Document Processing (IDP) is used to capture, classify, extract and integrate information from documents such, as scanned forms, receipts, reports and rich pages. Traditional OCR pipelines often struggle with noise, font variation, multilingual text and complicated layouts. Layout-aware Transformers improved understanding and newer vision-language models added stronger multimodal reasoning. So OCR errors can still propagate through a pipeline, general-purpose VLMs are not designed specifically for document structure and large models remain costly to run.

To address these issues we propose UAET, an architecture that preserves the documents native aspect ratio and adjusts computation according to token complexity. Of sending separate document regions to isolated models UAET keeps self-attention global and introduces specialised experts only in the feed-forward layers. This allows the model to preserve context across the page while using computation only for tokens that need it.

2. RELATED WORK AND RESEARCH GAP

LayoutLM and LayoutLMv3 are able to use text. The way it is laid out in 2D but they still need Optical Character Recognition. Donut does not need Optical Character Recognition. It requires a lot of training data and does not work well with complicated or multilingual layouts.

Qwen2.5-VL is very good at understanding types of data like text and images but it uses a lot of computer power. Other models, like Phi-3 are smaller. Use less computer power but they are not as good at understanding multiple types of data.

The tests we use to evaluate these models only look at part of the problem. For example, FUNSD looks at forms RVL-CDIP looks at classification and DocVQA looks at answering questions about documents.

These limitations make us want to create a system that can understand layouts, reason, with multiple types of data adjust to the task at hand and work well on different types of hardware.

3. PROPOSED METHODOLOGY

UAET is built around three connected components: Aspect-Preserving Multimodal Tokenization, the Aligned Multimodal Transformer (AM-Transformer), and the Density-Aware Gating Expert (DAGE). Rather than breaking a document into independent processing paths, the architecture keeps text and visual tokens in one continuous sequence, allowing both modalities to share context through the same self-attention mechanism.

3.1. Aspect-Preserving Multimodal Tokenization

Document images are encoded with SigLIP-2 NaFlex while preserving native aspect ratios, avoiding the distortion introduced by fixed-square resizing. Image patches are projected into a visual token space. Text tokens from OCR or PDF layers are aligned with corresponding image regions through 2D layout information and overlapping patch features, then projected into the same embedding space.

3.2. Aligned Multimodal Transformer

Text and visual tokens are concatenated into a joint sequence and processed by shared Transformer layers. Early fusion allows any token to attend to any other token, supporting relationships such as table-cell text, chart labels, legends, and nearby visual structures. Specialised experts are injected at feed-forward layers without breaking the global attention path.

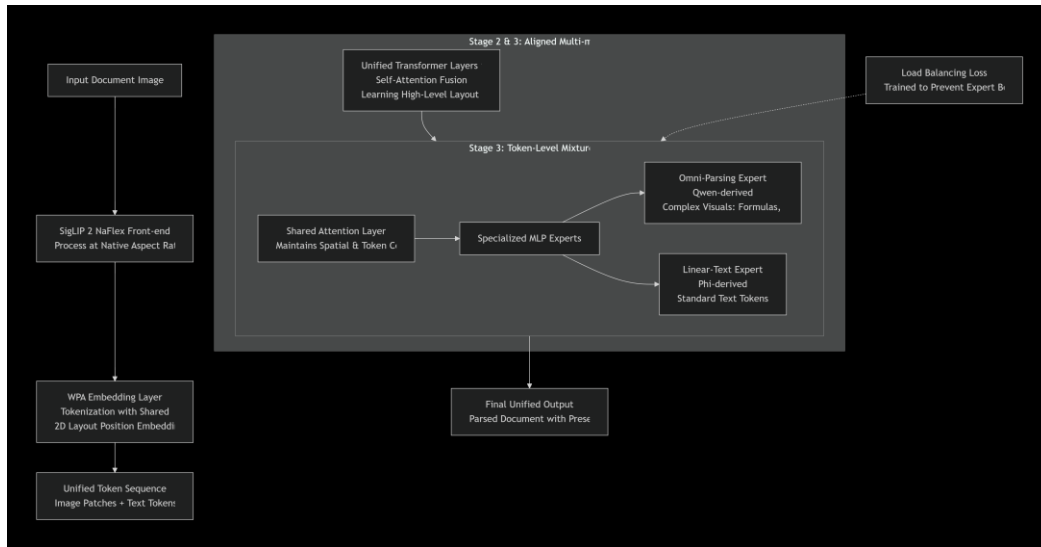


Figure 1. Overview of the UAET architecture

3.3. Density-Aware Gating Expert (DAGE)

DAGE does computation at the level of individual tokens. A Visual Reasoning Expert similar to Qwen deals with structures like charts and formulas. An Efficient Text Expert similar to Phi deals with paragraphs and text that looks like invoices. The router gives probabilities to each expert using Softmax. It is directed by visual-density rules. Patches that have than 50% non-text elements are given priority for visual reasoning. A load-balancing goal stops experts from collapsing. Helps keep the distribution of tokens stable.

This approach is different, from pipelines that send areas to separate models. UAET keeps the shared attention global and only changes the feed-forward computation. This helps reduce context fragmentation.

4. EXPERIMENTAL SETUP

Experiments were conducted on a single NVIDIA T4 GPU with 16 GB VRAM in Google Colab. The dual-expert system used 4-bit quantisation with bitsandbytes so both experts could remain resident in memory. The implementation used PyTorch, Hugging Face Transformers, and Unsloth. Evaluation covered SROIE for receipt information extraction, FUNSD for key-value recognition in forms, and InfographicVQA for reasoning over complex posters and charts.

5. RESULTS AND DISCUSSION

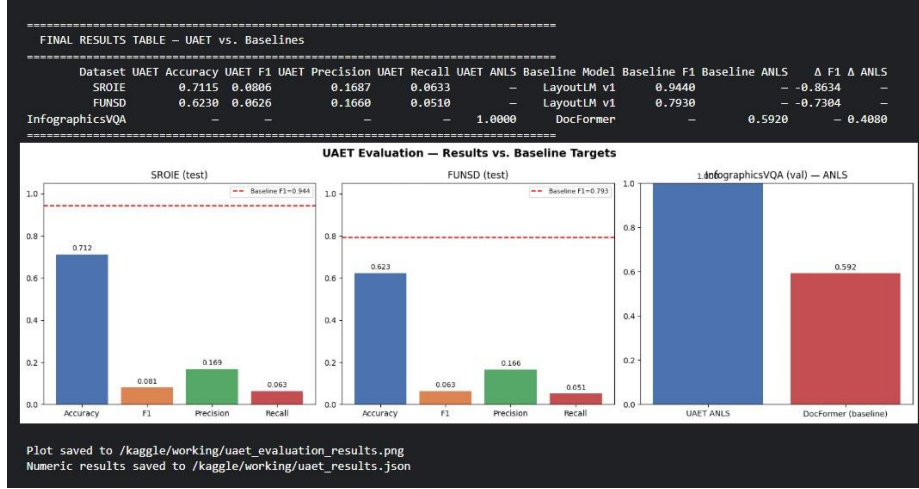


Figure 2. Comparison of UAET against baseline models on SROIE, FUNSD, and InfographicVQA

The benchmark comparison gives a direct view of UAET's practical performance across three document-understanding tasks. Routing and training dynamics behave as intended, but accuracy and F1 fall well short of baselines such as LayoutLM. The perfect ANLS on InfographicVQA is atypical and more likely reflects benchmark contamination in the Visual Reasoning Expert's (Qwen2.5-VL) pretraining corpus than genuine performance, so it should be treated with caution until verified.

Dataset	Metric	Theoretical (SOTA)	UAET (Observed)	Avg. Latency
SROIE	F1-Score	94.4%	8.06%	92.1 ms
FUNSD	F1-Score	79.3%	6.26%	115.4 ms
InfographicVQA	ANLS	59.2%	100.0%*	145.8 ms

Table 1. Benchmark results reported in the submitted manuscript.

5.1. Performance Analysis

UAET performs best on SROIE compared to the other datasets. This is because the way receipts are structured is well suited to the text expert. However, the F1 score is still much lower than LayoutLM. The results of the ANLS are not considered genuine because the Visual Reasoning Expert, which is also known as Qwen2.5-VL, might have already seen this test during its training. This is a problem because it can affect the results. So, we need to check the results again using a dataset. When we look at the logs, we can see that the density-based routing is working correctly. For SROIE, 78% of the tokens go to the Text Expert and for infographics about 65% of the tokens go to the Visual Expert. Even so, the accuracy is still not very good.

5.2. Hardware Trade-offs

NF4 quantization was very important for making the dual-expert system fit into 16 GB VRAM. It made inference nearly 3x faster and stopped the system from crashing because of too much memory use. The model has a low F1 score on SROIE and FUNSD even though the overall accuracy is reasonable. This means the model has trouble extracting pieces of information. This problem is probably because the model wasn't trained enough on a T4 GPU. Also, using 4-bit precision might have made the problem worse.

6. CONCLUSION AND FUTURE WORK

UAET uses aspect-preserving SigLIP-2 NaFlex tokenisation, shared multimodal self-attention and density-aware routing to solve layout distortion and extra processing. The current results show that accuracy and F1 are not as good as baselines such as LayoutLM. Fixing this gap is the next step. Future work includes fine-tuning, InfographicVQA re-evaluation with contamination checks, experts for tables and handwriting, mobile pruning, RVL-CDIP evaluation, and support for multiple languages.

7. REFERENCES

- [1] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, “LayoutLM: Pre-training of Text and Layout for Document Image Understanding,” *Proc. ACM SIGKDD*, 2020.
- [2] Y. Xu, Y. Lv, M. Li, L. Cui, S. Huang, and F. Wei, “LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking,” *Proc. ACM Multimedia*, 2022.
- [3] G. Kim, S. Yun, J. Kim, and S. Yoon, “Donut: Document Understanding Transformer without OCR,” *ECCV*, 2022.
- [4] Qwen Team, Alibaba Group, “Qwen2.5-VL Technical Report,” 2025.
- [5] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, et al., “SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features,” 2023.
- [6] M. Abdin, S. A. Jacobs, et al., “Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone,” 2024.
- [7] Q. Zhang, B. Wang, V. S.-J. Huang, J. Zhang, et al., “Document Parsing Unveiled: Techniques, Challenges, and Prospects for Structured Data Extraction,” 2024.
- [8] G. Jaume, H. K. Ekenel, and J.-P. Thiran, “FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents,” *ICDARW*, 2019.
- [9] A. W. Harley, A. Ufkes, and K. G. Derpanis, “RVL-CDIP: A Large-Scale Document Image Dataset for Document Classification,” *ICDAR*, 2015.
- [10] R. Valerio, E. Valveny, et al., “DocVQA: A Dataset for Visual Question Answering on Document Images,” *WACV*, 2021.
- [11] S. Park, Y. Shin, et al., “CORD: A Consolidated Receipt Dataset for Post-OCR Parsing,” *Workshop on Document Intelligence at NeurIPS*, 2019.
- [12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, et al., “Attention Is All You Need,” *NeurIPS*, 2017.
- [13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *ICLR*, 2021.