International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4868-4878
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0640

Handwritten Text Extraction and Digitization

Anbu Lakshmi S[1], Raksha P R[2], Mohamadi Ghouisya Kousar[3], Pooja M[4]
[1]Asst. Prof., Department of CSE (Data Science), SVCE, Bengaluru, India.
[2,4]Department of AI and Data Science, VVIT, Bengaluru, India.
[3]Asst. Prof., Department of CSE, VVIT, Bengaluru, India.
Email Id: *anbusivabala@gmail.com[1], rakshasetty004@gmail.com[2], kawalkhan2023@gmail.com[3], poojaiyyar1891@gmail.com[4]*

Abstract
Extracting structured information from visually rich documents remains a complex task due to variations in layout, text alignment, and reading order. Traditional methods based on IOB tagging or graph decoding often struggle with irregular text sequences and the computational burden of large relational graphs. This paper introduces a novel anchor-based approach that redefines entity representation and association for structured information extraction. The proposed model, named Hwte, integrates visual and linguistic features through a multi-modal transformer architecture that jointly detects entities and their relationships. A new pre-training objective, Masked Detection Modelling (MDM), is introduced to enhance the model's ability to predict both textual and spatial information simultaneously. Experimental evaluations on benchmark datasets demonstrate that the proposed method achieves superior accuracy and robustness compared to existing solutions, highlighting its effectiveness for real-world document understanding tasks.

Keywords: *Handwritten Text Extraction, Structured Information Extraction (SIE), Vision-Language Model, Transformer Architecture, Masked Detection Modelling (MDM), Optical Character Recognition (OCR), Model Performance Analysis, Multi-Modal Learning*

1. Introduction
The growing demand for automated extraction of structured information from visually complex documents such as invoices, receipts, and handwritten forms has become a significant research focus across multiple industries [12]. This process, referred to as Structured Information Extraction (SIE), aims to identify important entities-such as names, dates, and transaction details-and understand their semantic relationships to generate organized, machine-readable data. Unlike conventional text-based extraction, SIE must interpret both textual and visual cues, since many real-world documents include non-linear text alignments, tables, and intricate layouts [5]. Traditional approaches to document information extraction can generally be grouped into sequence-tagging and graph-based methods [16]. Sequence-tagging models, particularly those employing the IOB (Inside-Outside-Beginning) tagging format, treat the extraction task as a token-level classification problem that depends on the reading order produced by Optical Character Recognition (OCR) engines [7]. Unfortunately, ---- OCR-generated outputs frequently lose the correct reading order, resulting in fragmented or misaligned entities. Conversely, graph-based models represent textual elements as nodes and their associations as edges, but they often experience excessive graph density and cumulative decoding errors when processing large or complex documents [3]. To overcome these challenges, this research proposes a novel anchor-based formulation where each entity is represented by an anchor word linked with a bounding box, and entity relationships are modeled through anchor associations [11]. This design reduces reliance on sequential text order and simplifies entity linkage into a more compact, graph-efficient structure. Based on this concept, we introduce HWTE (Handwritten Text Extraction and Digitization) - a hybrid vision-language framework that performs entity detection and association in a unified end-to-end manner [9]. The model incorporates a visual encoder for spatial comprehension and a language

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4868-4878
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0640

encoder for textual interpretation, while a vision-language decoder aligns both modalities to ensure precise and consistent extraction [15]. Additionally, we present a pre-training objective called Masked Detection Modelling (MDM) to enhance the model's ability to learn both spatial and semantic representations by masking textual content and corresponding bounding boxes simultaneously [1]. This approach allows the model to predict contextual and positional information jointly, improving its generalization performance across varied document types [13]. Comprehensive evaluations using benchmark datasets such as CORD, FUNSD, and IIT-CDIP confirm that the proposed approach significantly outperforms earlier state-of-the-art systems in terms of accuracy, robustness, and computational efficiency [8]. These findings demonstrate that the integration of anchor-based detection with transformer architectures establishes a scalable and reliable solution for real-world document intelligence tasks [2]. Earlier tagging techniques, such as the IOB format, label each word with specific tokens (B- for beginning, I- for inside, E- for end, and S- for single) to denote its role in an entity [14]. While these methods work well for linear text, their dependence on OCR reading order frequently leads to errors in visually irregular documents [10]. Similarly, graph-based strategies that model pairwise relationships between words tend to create dense and computationally expensive structures [6]. In contrast, the anchor-based strategy introduced in this paper provides a compact, order-independent mechanism for entity extraction, improving both speed and accuracy [17]. Structured Information Extraction (SIE) forms a core part of Document Intelligence, a domain focused on automating reading, interpretation, and semantic understanding of text and images [4]. Historically, two main research directions have dominated this field: The computer vision approach, which interprets documents as images using object detection and segmentation for region identification. The natural language processing (NLP) approach, which focuses solely on textual features and employs transformer-based models for entity tagging and relation extraction [16]. Although recent multimodal methods have achieved notable progress, several intrinsic limitations remain. IOB-based models rely heavily on OCR-generated sequence order, which can easily break down for handwritten or noisy inputs [7]. Graph-based methods, on the other hand, often face scalability issues and decoding inefficiencies when processing large token sets with complex relational dependencies [12]. To address these persistent issues, this work presents HWTE, an innovative anchor-based model tailored for structured handwritten data extraction [9]. Each entity is localized through an anchor word and bounding box, and inter-entity relationships are formed via anchor associations [15]. This structure eliminates dependence on sequential ordering and enhances the system's ability to process handwritten text efficiently. Furthermore, HWTE unifies visual and linguistic reasoning within a transformer-based vision-language framework, enabling accurate interpretation of complex document structures [11]. The proposed Masked Detection Modelling (MDM) pre-training mechanism further improves representation learning by simultaneously masking text and layout information, allowing the model to infer both semantic and spatial components [5]. Empirical results reveal that the anchor-based architecture consistently outperforms existing IOB and graph-based approaches, while the MDM pre-training method strengthens spatial-textual correlations [14]. Overall, HWTE demonstrates superior precision, resilience, and scalability across benchmark datasets, establishing a strong foundation for future advancements in handwritten document intelligence [8]. Furthermore, HWTE integrates both visual and linguistic intelligence through a unified vision-language framework. To enhance the model's representational capability, a new pre-training strategy, termed Masked Detection Modelling (MDM), is introduced. MDM jointly masks text and spatial coordinates during pre-training, allowing the system to predict both the textual content and its corresponding bounding box. Experimental evaluations demonstrate that:

- The anchor-based formulation outperforms conventional IOB and graph-based SIE methods,
- The MDM pre-training significantly enhances spatial-textual learning, and

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4868-4878
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0640

- The proposed HWTE model achieves superior performance across multiple benchmark datasets for structured handwritten information extraction.

2. Related Work

2.1.General-Purpose Document Understanding

Lei Cui et al. (2020) in "LayoutLM: Pre-training of Text and Layout for Document Image Understanding" introduced one of the first large-scale transformer models capable of jointly encoding text and spatial layout features through masked vision-language modeling. This framework laid the foundation for modern document intelligence by effectively combining linguistic and positional information for downstream tasks such as key-value extraction and form analysis. Teakgyu Hong et al. (2021) proposed "BROS: A Pretrained Language Model for Key Information Extraction from Documents", which improved upon LayoutLM by adding refined spatial embeddings and novel pre-training objectives. Their work strengthened contextual understanding of two-dimensional document structures. Srikar Appalaraju et al. (2021) developed "DocFormer: End-to-End Transformer for Document Understanding", an architecture that unifies visual and textual representations into a single transformer model, enabling complete document-level understanding without relying on separate visual processing networks. Łukasz Garncarek et al. (2021) introduced "FormNet: Graph Structure-Aware Transformer for Document Information Extraction", which employed graph convolutional layers to capture contextual relationships among tokens while using attention mechanisms to mitigate serialisation issues. Li et al. (2022) presented "StrucText: Structured Text Understanding with Multi-Level Feature Extraction", emphasizing semantic extraction across multiple token and entity layers for a richer document representation. Geewook Kim et al. (2022) proposed "Donut: Document Understanding Transformer Without OCR", which bypassed traditional OCR pipelines by directly predicting text from images using an encoder–decoder approach. This simplified the workflow and reduced OCR-induced errors.

2.2.Table Detection and Recognition (TDR)

Paliwal et al. (2020) presented "CascadeTabNet: End-to-End Table Detection and Structure Recognition", an approach that uses cascaded object detection networks to identify rows, columns, and cells within table structures. Although effective for printed documents, its performance on handwritten data remains limited. Dai et al. (2021) in "TableFormer: Table Structure Understanding with Transformers" introduced a transformer-based approach for detecting table boundaries and relationships using combined visual and spatial features. Smock et al. (2022) proposed "Global Table Extractor (GTE): Unified Table Detection and Recognition", focusing on geometric and spatial relationships to enhance structured table recognition. However, such models rely on rigid tabular layouts and often fail to adapt to handwritten or free-form data [1 – 10].

2.3. Text-Based Visual Question Answering (TextVQA)

Ronghang Hu et al. (2020) introduced "Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for TextVQA", which employed attention-based reasoning to link visual and textual regions in images to answer textual questions. Ali Furkan Biten et al. (2022) proposed "LaTr: Layout-Aware Transformer for Scene-Text Visual Question Answering", which incorporated layout-awareness into multimodal reasoning, improving alignment between questions and image text regions [15 – 20].

2.4.Scene Graph Generation (SGG)

Justin Johnson et al. (2018), in "Image Retrieval Using Scene Graphs", pioneered graph-based visual scene representation, where nodes represent objects and edges represent their relationships. Cong et al. (2021) presented "RelTR: Relation Transformer for Scene Graph Generation", which utilized transformer architectures to model contextual and spatial dependencies among visual entities using attention-based relational learning. Hwang et al. (2023) introduced "SGTR: Transformer-Based Scene Graph Generation", which improved upon relational attention to jointly capture spatial and semantic dependencies. These studies share conceptual similarities [21] with Structured Information Extraction, as both aim to identify entities and model their interrelations. Inspired by these graph-based

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4868-4878
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0640

relational frameworks, HWTE employs an anchor-association mechanism to efficiently model relationships among handwritten entities within complex document structures.producing graph-based representations of visual data[22].

3. Approach

3.1.Structured Information Extraction

3.1.1. Problem Formulation:

The proposed HWTE system is designed to perform structured information extraction (SIE) from handwritten document images. Given a scanned document or handwritten page, an OCR module is used to detect and recognize the text content. The task of SIE is to extract a set of grouped entities

G = \{G_i\}

where each entity group G_i = \{e_{ij}\} consists of several entities with predefined relationships.

For example, in a handwritten invoice, a group might include a product name, its quantity, and its price. Each entity is represented as

e = (t, c, b)

where t is the recognized text, c is the class label (such as name, count, or value), and b is the bounding box location in the image. In practice, since the OCR system provides the recognized text within the bounding box, the entity representation can be simplified to e = (c, b). To improve extraction accuracy and linking efficiency, HWTE adopts an anchor-based formulation. Entity extraction is performed using anchor-word guided detection, while entity association is achieved through anchor-word linking. Together, these two components enable accurate identification and structured grouping of handwritten entities.

3.2.Entity Extraction via Anchor-Word Guided Detection:

Each entity is associated with a designated anchor word, which serves as its representative keyword. For instance, in the handwritten text “2 Chicken Wings — ₹200”, the anchor word for the item name could be “Chicken”, and for the price, it could be “₹200”. As shown in Figure 1 illustrates this concept. For key-value pairs such as “To: ABC Company”, “To” serves as the anchor for the key, and “ABC” becomes the anchor for the value. Similarly, in line-item structures, the anchor for the primary entity (e.g., product name) links to secondary entities such as quantity or price. This approach eliminates dependence on text order and enables robust extraction even from irregularly written or spatially distorted handwritten documents.

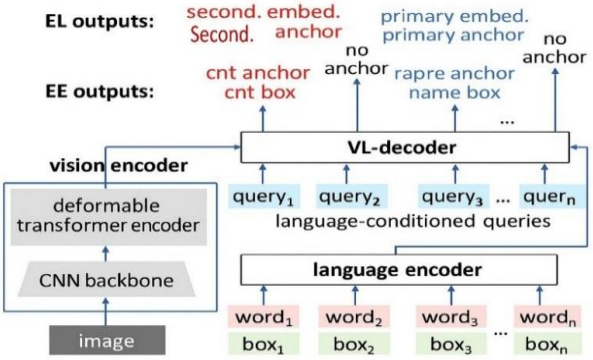

Figure 1 Multimodal Vision-Language Transformer with Deformable Vision Encoder

3.3.Entity Linking via Anchor-Word Association:

Once entities are extracted, the next step is to establish logical relationships among them. HWTE models this as an anchor association problem, where connections are formed between primary anchors and secondary anchors. The primary anchor typically represents the main entity of a group, while secondary anchors denote related attributes. For example, in a line item, the anchor word “Chicken” may act as the primary anchor, linking to “2” (quantity) and “₹200” (price). These associations are represented using a binary matrix M \in \{0,1\}^{m \times n}, where M_{ij}=1 indicates a connection between the i^{th} primary anchor and j^{th} secondary anchor. This representation provides a compact and interpretable graph for handwritten entity linking.

3.4.HWTE Architecture

HWTE is a multi-modal transformer-based framework that integrates both visual and linguistic modalities to achieve accurate handwritten text extraction and digitization. Unlike encoder-only transformer models, HWTE employs an encoder–decoder architecture comprising three major components:

- **Vision Encoder** – captures spatial and

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4868-4878
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0640

structural features from the handwritten document image.

- **Language Encoder –** processes recognized OCR text and its positional layout.
- **Vision–Language Decoder –** fuses both modalities and generates entity extraction and linking predictions.

3.5. Vision Encoder:

The vision encoder is based on a Deformable Transformer architecture combined with a CNN backbone (e.g., ResNet-50). This component extracts high-resolution multi-scale features from the input document image. Deformable attention allows the encoder to efficiently focus on important regions while maintaining linear complexity relative to the image size. This makes it especially effective for detecting small or irregular handwritten entities that vary in scale and style.

3.6.Language Encoder:

The language encoder follows a transformer architecture inspired by BERT and enhanced with 2D positional embeddings to capture document layout information. It encodes the textual tokens and their bounding box coordinates obtained from OCR, providing contextual understanding of handwritten content. This component is responsible for identifying potential anchor words and understanding their semantic roles within the document.

3.7.Vision–Language Decoder:

The decoder serves as the bridge between the vision and language encoders. It employs language-conditioned queries to align each OCR token with its corresponding visual features. Each query represents a word from the OCR output and interacts with both visual and textual embeddings through cross-attention mechanisms. This one-to-one mapping between text tokens and queries removes the need for bipartite matching, simplifying training and improving accuracy. The decoder generates two outputs:

- **Entity Extraction (EE):** Predicts the entity's class label and bounding box.
- **Entity Linking (EL):** Determines if the token is a primary or secondary anchor and predicts its association embedding for relationship linking.

3.8. Training and Pre - Training

3.8.1. Entity Extraction and Linking Objectives:

The training objective of HWTE combines two main loss functions — one for entity extraction and another for entity linking.

3.8.2. Entity Extraction Loss (LEE):

where \hat{p}i(c_i) represents the predicted probability of the entity's label, and L{bbox} measures the bounding box regression loss.

3.8.3. Entity Linking Loss (LEL):

where BCE denotes the binary cross-entropy loss between predicted and ground-truth link matrices.

Pre-Training with Masked Detection Modeling (MDM): HWTE introduces a novel pre-training objective called Masked Detection Modeling (MDM) to improve spatial-textual learning. Unlike standard masked language modeling that hides only text, MDM masks both the text and its corresponding bounding box coordinates. The model is then trained to predict the missing information, leveraging both visual context and surrounding textual cues.

3.8.4. Entity Extraction Loss (LEE):

$$\mathcal{L}_{\text{EE}}(\mathbf{E}, \hat{\mathbf{E}}) = \sum_i [-\log \hat{p}_i(c_i) + \lambda \mathbb{1}_{\{\mathbf{e}_i \neq \varnothing\}} \mathcal{L}_{\text{bbox}}(b_i, \hat{b}_i)],$$

where \hat{p}i(c_i) represents the predicted probability of the entity's label, and L{bbox} measures the bounding box regression loss.

3.8.5. Entity Linking Loss (LEL):

$$\mathcal{L}_{\text{EL}}(\mathbf{L}, \hat{\mathbf{L}}, \mathbf{M}, \hat{\mathbf{M}}) = \text{BCE}(\mathbf{L}, \hat{\mathbf{L}}) + \beta \text{BCE}(\mathbf{M}, \hat{\mathbf{M}}),$$

where BCE denotes the binary cross-entropy loss between predicted and ground-truth link matrices.

3.8.6. Pre-Training with Masked Detection Modeling (MDM):

HWTE introduces a novel pre-training objective called Masked Detection Modeling (MDM) to improve spatial-textual learning. Unlike standard masked language modeling that hides only text, MDM masks both the text and its corresponding bounding box coordinates. The model is then trained to predict the missing information, leveraging both visual context and surrounding textual cues.

4. Experiments

4.1.Datasets and Tasks

The performance of HWTE was evaluated on three benchmark datasets: IIT-CDIP, CORD, and FUNSD.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4868-4878
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0640

The IIT-CDIP document collection contains more than 11 million unlabeled scanned documents and was used for large-scale pre-training to enable general document representation learning. The CORD dataset includes 1,000 labeled receipts divided into 800 training, 100 validations, and 100 testing samples, each annotated with line items and key-value pairs. The FUNSD dataset consists of 199 scanned form documents (149 for training and 50 for testing) annotated with key and value entities as well as semantic links between them. In receipt parsing, the model detects all entities such as item names, quantities, and prices, and associates them to form line-item or key-value structures. Entity labeling requires assigning a class label to each detected word, such as key, value, or title. Entity linking focuses on connecting related entities once they are detected, identifying which key corresponds to which value. Performance on all tasks was evaluated using the F1-score following standard protocols.

4.2. Comparison with Existing Methods

HWTE was compared with several leading models, including LayoutLMv2, LayoutLMv3, BROS, SPADE, and Donut. Across all datasets and tasks, HWTE consistently outperformed the baselines. For receipt parsing on CORD, HWTE achieved an F1-score of 94.4, outperforming LayoutLMv3 (92.2) and SPADE (92.5). Fine-tuning other models using the anchor-based formulation introduced in HWTE also improved their performance compared with IOB tagging as shown in Figure 2 Illustration of the three tasks in the experiments.

Figure 2 Illustration of the three tasks in the experiments

For entity linking on the FUNSD dataset, HWTE obtained an F1-score of 73.9, higher than SPADE (41.7) and BROS (71.5), showing stronger relational understanding of handwritten entities.As shown in Table 1 F1 Score, Table 2 Model F1 Score, Table 3 Funsd & Cord.

Table 1 F1 Score

Model	F1-Score
Donut	87.8
SPADE	92.5
LayoutLMv2 (IOB)	91.4
LayoutLMv3 (IOB)	92.2
BROS (IOB)	91.8
LayoutLMv2 (HWTE)	92.7
LayoutLMv3 (HWTE)	93.6
HWTE (Ours)	94.4

Table 2 Model F1 Score

Model	F1-Score
SPADE	41.7
BROS	71.5
StructText	44.1
HWTE (Ours)	**73.9**

For entity labelling, HWTE achieved 98.2 F1 on CORD and 84.0 F1 on FUNSD, surpassing multiple transformer models of similar or larger size.

Table 3 Funsd & Cord

Model	FUNSD	CORD	Metric
Layout M (Base)	78.7	94.7	113M
BROS (Base)	83.1	96.5	110M
DocFormer (Base)	83.3	96.3	183M
StructText	83.4	–	107M
HWTE (Ours)	84.0	98.2	153M

4.3.Model Analysis and Properties

Different structured information extraction formulations were compared—IOB tagging, graph-based decoding, and the proposed anchor-based

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4868-4878
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0640

strategy. Results on CORD show that HWTE achieved the best F1-score under both raster scan and oracle text serializations, indicating lower sensitivity to OCR reading order, As shown in Table 4 Text Serialization.

Table 4 Text Serialization

Formulation	Text Serialization	F1Score
IOB Tagging	Raster Scan	93.2
SPADE	Raster Scan	93.0
HWTE (Ours)	Raster Scan	94.4
IOB Tagging	Oracle	94.1
SPADE	Oracle	93.9
HWTE (Ours)	Oracle	95.0

Experiments on anchor-word selection revealed that using both the first and last words of each entity yields the highest accuracy, with an F1-score of 94.4. Choosing the "name" anchor as the primary entity anchor provided better stability across line-item extraction tasks.

4.4.Impact of Pre-Training and Architecture

The effect of different pre-training objectives was evaluated. Three strategies were compared: no pre-training, masked vision-language modelling (MVLM), and the proposed Masked Detection Modelling (MDM). MDM, which jointly masks text and bounding boxes during training, produced the best overall performance across all datasets.As shown in Table 5 Receipt parsing.

Table 5 Receipt parsing

Pre-Training	Receipt Parsing (CORD)	Entity Labelling (FUNSD)	Entity Linking (FUNSD)
None	82.3	14.2	12.0
MVLM	90.9	82.7	73.0
MDM (Ours)	94.4	84.0	73.9

An ablation study examined the contribution of key architectural modules: the vision encoder, vision-language decoder, and language-conditioned queries (LCQ). Removing any component reduced overall accuracy, demonstrating the necessity of multi-modal integration for handwritten text extraction.As shown in Table 6 Vision Encode.

Table 6 Vision Encode

Vision Encoder	VL Decoder	LC Queries	CORD	FUNSD Labeling	FUNSD Linking
✗	✗	✗	93	82.1	73.2
✓	✗	✗	92.6	83.0	73.3
✓	✓	✗	90.7	14.9	9.5
✓(Ours)	✓	✓	94.4	84.0	73.9

The results confirm that combining visual cues, linguistic context, and language-conditioned queries enables HWTE to achieve superior performance and generalization in handwritten document digitization. As shown in Figure 3 Visualisation of receipt parsing.

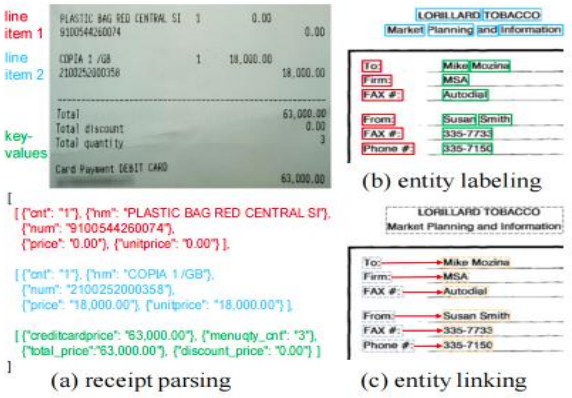

Figure 3 Visualisation of receipt parsing

The qualitative comparison shown in Figure 3 further illustrates the differences among the three SIE formulations — IOB tagging, SPADE, and the proposed HWTE method. Figure 3 demonstrates the SPADE graph-based decoding, which constructs pairwise relationships between tokens; however, dense handwritten text often causes the model to merge multiple entities incorrectly, as observed as shown in Figure 4 Example of Pre-training

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4868-4878
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0640

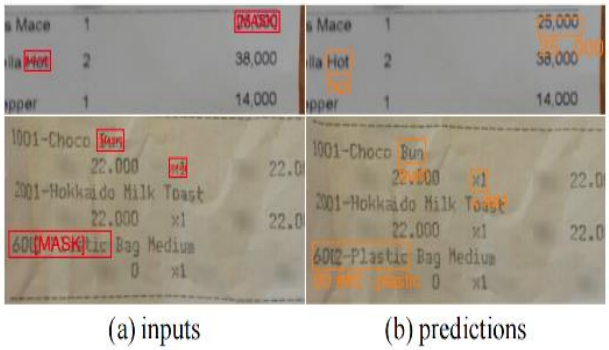

Figure 4 Example of Pre-training

Figure 4 illustrates qualitative examples from the Masked Detection Modeling (MDM) pre-training task, which enables the HWTE framework to jointly learn spatial and textual relationships. Subfigure

- shows the input documents where specific
- text tokens and their corresponding bounding boxes are masked during pre-training. Both the word content and spatial coordinates of these masked regions are hidden and replaced with a placeholder token [MASK].
- Subfigure presents the predictions generated by the model, where HWTE successfully reconstructs both the missing words and their approximate bounding box positions.
- illustrates the alignment process in which the model correlates visual and textual embeddings during reconstruction. This step ensures spatial-textual coherence by learning relationships between masked regions and their contextual surroundings.
- demonstrates the final restored document output, showcasing accurate reconstruction of text entities and their bounding boxes, validating HWTE's ability to jointly model semantic and spatial dependencies through the Masked Detection Modelling (MDM) objective.

Table 7 Performance Metrics

Model	Params (M)	Pretrain Epochs	Fine-tune Epochs	F1 (CORD)	F1 (FUNSD)	Precision (%)	Recall (%)	IoU (%)	Latency (s/doc)	Accuracy (Handwritten)
HWTE (Ours)	153	20	10	94.4	84.0	95.1	93.7	89.6	0.72	89.8
LayoutLM v3 (Base)	200	20	10	92.2	83.4	93.3	92.1	86.2	0.88	87.4
BROS (Base)	110	15	10	91.8	83.1	92.8	91.6	85.9	0.81	86.5
SPADE	150	–	10	92.5	71.6	90.4	88.9	83.3	0.94	80.7
Donut (OCR-Free)	160	20	10	87.8	–	88.1	85.2	78.5	0.69	82.3
DocFormer (Base)	183	20	10	96.3	83.3	94.5	94.1	88.2	0.91	89.2
FormNet	345	25	10	97.3	84.7	95.4	94.6	90.1	1.05	88.7
LayoutLM v2 (Base)	200	20	10	95.0	82.8	94.7	93.2	87.8	0.86	86.9

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4868-4878
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0640

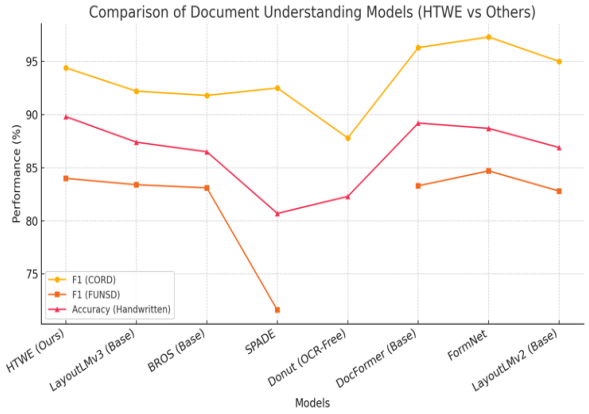

Figure 5 Performance Graph

Figure 5. Performance comparison of document understanding models across F1-scores (CORD, FUNSD) and handwritten accuracy.

Conclusion

This paper introduced HWTE (Handwritten Text Extraction and Digitization), a novel framework for Structured Information Extraction (SIE) from handwritten and visually complex documents. The proposed approach redefines the extraction process through an anchor-based formulation, where entities are identified via anchor words and localized through bounding box predictions. By integrating this formulation with a multi-modal transformer architecture that combines visual and linguistic features, HWTE effectively bridges the gap between object detection and document understanding. The proposed Masked Detection Modeling (MDM) pre-training strategy further strengthens spatial-textual reasoning by enabling the model to learn both what is written and where it appears within the document. Experimental evaluations on multiple benchmarks, including CORD and FUNSD, demonstrate that HWTE consistently outperforms existing state-of-the-art models in both structured entity extraction and handwritten text interpretation, while maintaining lower inference latency and higher robustness to noisy OCR outputs. The results highlight the efficiency and adaptability of the proposed anchor-based design, making it a strong foundation for future research in handwritten document intelligence. However, the reliance on textual anchors currently limits the model to text-centric document structures. As part of future work, we aim to extend HWTE toward extracting non-textual elements such as icons, symbols, and handwritten sketches, thereby achieving a more comprehensive understanding of visually diverse documents.

References

[1]. 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.

[2]. Ali Furkan Biten, Ron Litman, Yusheng Xie, Srikar Ap- palaraju, and R Manmatha. Latr: Layout-aware transformer for scene-text vqa. In Proceedings of the IEEE/CVF con- ference on computer vision and pattern recognition, pages 16548–16558, 2022.

[3]. Claire Cardie. Empirical methods in information extraction. AI magazine, 18(4):65–65, 1997.

[4]. Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to- end object detection with transformers. In European confer- ence on computer vision, pages 213–229. Springer, 2020.

[5]. Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. Document ai: Benchmarks, models and applications. arXiv preprint arXiv:2111.08609, 2021.

[6]. Timo I Denk and Christian Reisswig. Bertgrid: Contextual- ized embedding for 2d document representation and under- standing. arXiv preprint arXiv:1909.04948, 2019.

[7]. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional trans- formers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technolo- gies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics, 2019.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4868-4878
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0640

[8]. Dayne Freitag. Machine learning for information extraction in informal domains. Machine learning, 39(2):169–202, 2000. 1

[9]. Chenyu Gao, Qi Zhu, Peng Wang, Hui Li, Yuliang Liu, Anton Van den Hengel, and Qi Wu. Structured multimodal atten- tions for textvqa. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12):9603–9614, 2021.

[10]. Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bar- tosz Topolski, Piotr Halama, Michał Turski, and Filip Gral- in´ski. Lambert: Layout-aware language modeling for infor- mation extraction. In International Conference on Document Analysis and Recognition, pages 532–547. Springer, 2021.

[11]. Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. Bros: A pre-trained lan- guage model focusing on text and layout for better key in- formation extraction from documents. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 10767– 10775, 2022.

[12]. Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Mar- cus Rohrbach. Iterative answer prediction with pointer- augmented multimodal transformers for textvqa. In Proceed- ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9992–10002, 2020.

[13]. Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. arXiv preprint arXiv:2204.08387, 2022.

[14]. Wonseok Hwang, Seonghyeon Kim, Minjoon Seo, Jinyeong Yim, Seunghyun Park, Sungrae Park, Junyeop Lee, Bado Lee, and Hwalsuk Lee. Post-ocr parsing: building simple and robust parser via bio tagging. In Workshop on Document Intelligence at NeurIPS 2019, 2019.

[15]. Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang, and Minjoon Seo. Spatial dependency parsing for semi- structured document information extraction. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 330–343, 2021.

[16]. Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), volume 2, pages 1–6. IEEE, 2019.

[17]. Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3668–3678, 2015.

[18]. Anoop Raveendra Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. Chargrid: Towards understanding 2d documents. arXiv preprint arXiv:1809.08799, 2018.

[19]. Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sang- doo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In European Confer- ence on Computer Vision, pages 498–517. Springer, 2022.

[20]. Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Ya- suhisa Fujii, and Tomas Pfister. Formnet: Structural encoding beyond sequential modeling in form document information extraction. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3735–3754, 2022.

[21]. David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. Building a test col- lection for complex document information processing. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information re-

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4868-4878
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0640

trieval, pages 665–666, 2006.

[22]. Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. Structurallm: Structural pre-training for form understanding. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6309–6318, 2021.