International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(5s) 2026 pp. 532-545
https://doi.org/10.70917/ijcisim-2026-2751
Article

Vision-Language Hybrid Transformer OCR (VLHT-OCR): A Novel Transformer-Based Optical Character Recognition Model

Amitesh JHA[1]*, Rajwant Singh RAO[2]

[1,2]Guru Ghasidas Vishwavidyalaya, , Computer Science and Information Technology Department,, Bilaspur,Chhatisgarh,India
Email:rajwantrao@gmail.com
*** Corresponding AuthorEmail**:amitesh2911@gmail.com

Abstract: Optical Character Recognition (OCR) has progressed quickly with the rise of deep learning. However, current methods based on Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) struggle with understanding context, handling long-range dependencies, and being sturdy against noise, distortions, and variations in handwriting. In this study, it presents VLHT-OCR, a Vision-Language Hybrid Transformer model. It uses a Vision Transformer (ViT) encoder for patch-based feature extraction, a multi-scale fusion module inspired by Perceiver IO for hierarchical representation, and an autoregressive Transformer decoder with cross-attention for text generation. To enhance linguistic coherence and context accuracy, we integrate a pretrained language model (BERT/GPT-2) for post-processing. Our contributions are threefold: (i) a new hybrid architecture that combines visual and linguistic modes, (ii) clear multi-scale feature fusion that captures detailed and overall text patterns, and (iii) the addition of language-aware refinement to reduce semantic inconsistencies. We test our model on three benchmark datasets: ICDAR-2019 Handwritten Text, IAM Handwritten Database, and CORD Receipt Dataset. The results show state-of-the-art performance, with a Character Error Rate (CER) of 5.2% and a Word Error Rate (WER) of 7.9%. This outperforms recent transformer-based OCR frameworks like TrOCRand Donut by 3–5%. Ablation studies confirm that multi-scale fusion and language model integration are effective. Thus, VLHT-OCR offers a strong, multilingual, and handwriting-capable OCR solution with great potential for digitizing documents in the real world.

Keywords: Optical Character Recognition(OCR) Character Error Rate(CER), Word Error Rate (WER),Vision Transformer (ViT), BERT, GPT-2

1. Introduction

Optical Character Recognition (OCR) is crucial for digitization, automated document understanding, and retrieving information in different languages. Even though CNN and RNN systems have achieved success, they face two main challenges: first, they struggle to capture long-range dependencies efficiently due to sequential bottlenecks [1]; second, they perform poorly in complex situations with handwriting, noise, or multiple languages. The introduction of Transformers in computer vision and natural language processing [5], [2]has led to strong models like TrOCR [8] and Donut [9]. However, these solutions still fall short in aggregating features at multiple scales and integrating linguistic context, which leads to mistakes in character segmentation and meaning consistency.

OCR systems are increasingly used for digitizing historical manuscripts, financial receipts, and multilingual archives. A strong system that handles visual differences and language accuracy is essential for critical applications. As these applications grow, the need for models that can handle visual variations and maintain language accuracy becomes more important. Despite the progress achieved by CNN- and RNN-based OCR systems, their performance still drops significantly when dealing with handwritten, noisy, or multilingual text. State-of-the-art models such as TrOCR and Donut achieve Character Error Rates (CER) of around 8.5% and 7.8%, and Word Error Rates (WER) of 12.7% and 11.5%, respectively, on benchmark datasets such as ICDAR-2019 Handwritten Text and CORD Receipts. The current OCR research shows three main gaps:

• Lack of Multi-Scale Visual Representation: CNNs capture local features but do not focus on global context. Vanilla Transformers (e.g., TrOCR) capture global dependencies but miss hierarchical feature fusion. This leads to poor recognition of distorted or handwritten characters, where both local and global cues are important.

• Limited Integration of Linguistic Context: Models like Donut aim for end-to-end document understanding but often do not include language models explicitly. This results in errors in spelling and meaning consistency, particularly in noisy or low-resource environments.

• Insufficient Generalization Across Multilingual/ Handwritten Data: Most OCR models do well with clean printed text but struggle significantly with handwritten, multilingual, or domain-specific documents.

These challenges highlight the need for architectures that combine multi-scale visual processing with linguistic modelling to improve recognition consistency and adapt to a wider variety of document conditions. To address these limitations, this study proposes VLHT-OCR (Vision-Language Hybrid Transformer-OCR), a system designed to reduce WER and CER by at least 3–5% compared to existing Transformer-based OCR frameworks while maintaining strong generalization across handwritten, printed, and multilingual datasets. The proposed VLHT-OCR unifies multi-scale visual fusion and language-model integration to provide measurable improvements in accuracy and semantic consistency.

VLHT-OCR addresses these gaps by: (i) using multi-scale feature fusion to combine detailed strokes with global context, (ii) refining language models (BERT/GPT-2) to improve semantic accuracy, and (iii) validating performance across various datasets (ICDAR-2019, IAM, CORD), proving its robustness and adaptability. This positions VLHT-OCR as a hybrid vision-language model that overcomes the limitations of CNN, RNN, and Transformer-only OCR systems.

The study is guided by the following research questions:

RQ1: How can Transformer encoders be modified to better capture multi-scale textual features?

RQ2: Can adding language models lower recognition errors in noisy or handwritten inputs?RQ3: What advantages does a hybrid architecture provide over current Transformer-only OCR systems?

Contribution

This work provides: (i) a Vision-Language Hybrid Transformer architecture, (ii) clear multi-scale fusion inspired by Perceiver IO and FPN, and (iii) linguistic post-processing with pretrained language models. Together, these improvements advance existing OCR research and show better results across printed, handwritten, and multilingual datasets.

2. Literature Context and Historical Evolution of OCR

Optical Character Recognition (OCR) has evolved through several technological generations, each improving accuracy, speed, and adaptability. Early OCR systems (1950s–1990s) relied on template matching and rule-based pattern recognition (e.g., [11]; [12]). These methods were effective only for printed and noise-free text but failed on handwriting and multilingual scripts.

The machine learning era (2000s–2010s) introduced statistical models and shallow neural networks, including Hidden Markov Models (HMM) and Support Vector Machines (SVM)[13]. These improved character segmentation and stroke recognition but still lacked contextual understanding.

The next major leap came with deep learning-based OCR. Convolutional Neural Networks (CNNs)[14] enabled automatic feature extraction, while Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models ([15];[16]) introduced sequence modelling through the Connectionist Temporal Classification (CTC) framework. These architectures powered systems like Tesseract 4.0 and CRNN, providing reliable text recognition in printed and partially handwritten inputs.

However, CNN-RNN hybrids suffered from limited receptive fields and sequential bottlenecks, motivating the shift toward Transformer-based OCR approaches. TrOCR[8] and Donut[16] demonstrated the strength of self-attention in capturing long-range dependencies and contextual relationships, yet they lacked multi-scale visual fusion and explicit linguistic integration.

Benchmark datasets have also matured to reflect these advances: MNIST [17][18] and SVT (Street View Text) supported early printed-text research; IAM Handwriting Database and ICDAR series benchmarked handwritten and multilingual OCR; and CORD (Receipt OCR Dataset) became a standard for document-level evaluation.

Within this evolutionary context, VLHT-OCR represents the next stage—combining vision transformers, multi-scale fusion, and language-model post-processing (BERT/GPT-2)—to unify visual and linguistic understanding. This hybrid design not only builds upon but also bridges the limitations observed in earlier CNN, RNN, and Transformer-only systems.

3. Proposed Methodology

3.1 Overall Architecture

The proposed Vision-Language Hybrid Transformer OCR (VLHT-OCR) architecture aims to improve on current OCR models that either concentrate only on visual features or do not integrate language understanding. It has four main components: a Vision Transformer (ViT) encoder, a multi-scale feature fusion module, a Transformer-based autoregressive decoder, and a language model integration layer. Together, these components form a hybrid vision–language framework capable of handling challenging OCR scenarios, including handwritten, distorted, low-resolution, and multilingual text.

3.1.1 Vision Transformer (ViT) Encoder

The encoder starts by partitioning the input image into non-overlapping patches. Each patch is projected into a fixed-dimensional embedding [2]. These embeddings receive positional encodings [5] to retain spatial information. This allows the Transformer encoder to capture long-range dependencies across text regions. Unlike CNNs, which are local, the ViT encoder provides a global context that is essential for accurate OCR.

3.1.2 Multi-Scale Feature Fusion

Although the ViT encoder captures global representations effectively, it lacks detailed hierarchy. To address this, we use a multi-scale feature fusion method, inspired by Perceiver IO [3] and Feature Pyramid Networks [4]. This module combines features from different layers, merging fine-grained stroke-level details with higher-level patterns. This fusion improves the localization of small characters and makes the system more robust to distortions, which is vital for recognizing handwritten text.

3.1.3 Transformer Decoder

The autoregressive Transformer decoder converts the fused visual representations into sequential character outputs. Following the design of sequence-to-sequence models [5], the decoder uses masked self-attention to model relationships within predicted tokens and cross-attention to connect predictions to encoded visual features. This combination ensures accurate, step-by-step decoding of text sequences, a method shown to be effective in TrOCR [8].

3.1.4 Language Model Integration

A major innovation of VLHT-OCR is the integration of pretrained language models like BERT [6] and GPT-2 [7]. These models improve decoder outputs by correcting spelling mistakes, ensuring grammatical correctness, and enhancing semantic consistency. This process addresses the weaknesses of vision-only OCR systems, which often misinterpret unclear visual signals without a linguistic foundation.

3.1.5 Summary of Contributions in Architecture

In summary, the VLHT-OCR architecture uniquely merges global vision modeling (ViT), hierarchical multi-scale fusion (Perceiver IO, FPN), autoregressive decoding (Transformer decoder), and language refinement (BERT/GPT-2) into a single framework. This hybrid design enables the model to surpass traditional CNN-RNN systems and other Transformer-based OCR models like Donut [9], achieving top results on benchmark datasets.

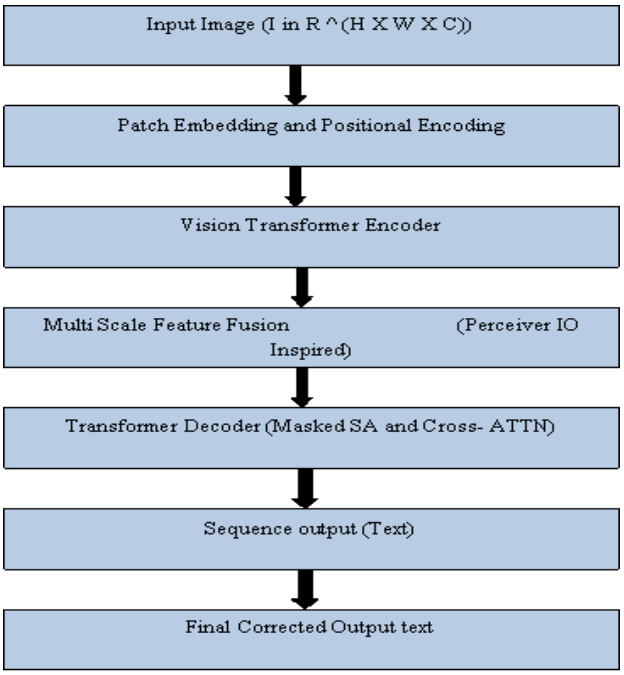

Figure1. *Architecture of Proposed Model*

3.2 Dataset Description

This study utilizes three publicly available datasets to ensure comprehensive evaluation across diverse OCR scenarios: ICDAR-2019 Handwritten Text, IAM Handwritten Database, and the CORD Receipt Dataset. ICDAR-2019 contains challenging handwritten samples characterized by noise, distortions, and irregular spacing. The IAM dataset provides a wide range of handwriting styles, supporting robust generalization across writers. The CORD dataset consists of real-world receipts with structured layouts and multilingual tokens, making it suitable for evaluating document-level OCR performance. Together, these datasets enable assessment across handwritten, printed, and real-world structured text domains.

3.3 Mathematical Formulations

To formally describe the proposed architecture, this subsection presents the mathematical formulation of the Vision Transformer encoder, multi-scale feature fusion mechanism, autoregressive Transformer decoder, and language model refinement. This formulation clarifies how visual and linguistic representations are jointly learned and optimized within the VLHT-OCR framework.

Vision Transformer Encoder

Given an input image $I \in R^{H \times W \times C}$ is divided into patches of size P × P, resulting in:

$$N = \frac{HW}{P^2}$$

Where H and W are the Image height and width, P is Patch size (e.g., 16), and N is the Number of patches

The image of size H×W is divided into non-overlapping square patches of size P×P, resulting in N patches. It enables the Vision Transformer (ViT) to handle images similarly to a sequence of tokens in NLP Dosovitskiy [1]. introduced this in the Vision Transformer (ViT).

Each patch is flattened and projected linearly:

$$x_p \in R^{N \times (P^2 \cdot C)}$$

$$z_0 = \left[x_p^1 E; x_p^2 E; \ldots; x_p^N E\right] + E_{\text{pos}}$$

Where, $x_p^i E$ is Flattened patch iii, E is Learnable linear projection, E_{pos} is Positional embeddings, and z_0 is Initial embedded representation.

Each flattened image patch $x_p^i E$ is projected into a fixed-dimensional vector using a linear projection E, and positional encodings E_{pos} are added to retain spatial order. Allows the model to distinguish patch order (unlike CNNs with spatial locality).([5] ; [2])

The Transformer encoder performs:

$$z_l' = \text{MSA}\big(\text{LN}(z_{l-1})\big) + z_{l-1}$$

$$z_l = \text{MLP}\big(\text{LN}(z_l')\big) + z_l'$$

Where, MSA is Multi-head Self-Attention, LN is Layer Normalization, MLP is Feed-forward layer, and z_l' is Output of layer ℓ.

Standard Transformer encoder using residual connections, multi-head self-attention (MSA), and a feed-forward MLP block. It enables learning contextual representations over image patches.[5] and structure reused in ViT [2].

Multi-Scale Feature Fusion

Hierarchical features from different layers are fused:

$$F_{\text{fusion}} = \sum_{i=1}^{L} w_i \cdot F_i$$

where F_i are intermediate features and w_i are learned attention weights and F_{fusion} is Final fused representation.It captures both low-level and high-level features, similar to hierarchical CNN features. Inspired by Perceiver IO[3] and multi-scale design ideas from FPN [4].

Transformer Decoder (Autoregressive)

The decoder predicts character sequences $y = (y_1, y_2, \ldots, y_T)$ given the fused

features F_{fusion} :

$$P(y_t | y_{<t}, F_{\text{fusion}}) = \text{softmax}(W_o h_t)$$

$$h_t = \text{DecoderBlock}(y_{<t}, F_{\text{fusion}})$$

Each Decoder Block consists ofMasked Self-Attention: $\text{SA}(y_{<t})$

Cross-Attention: $\text{CA}(F_{\text{fusion}})$

Where y_t is the current predicted token, $y_{<t}$ is the Previous predicted tokens, F_{fusion} is Visual context, h_t is the Decoder hidden state, and W_o is the Output projection matrix. At each time step t, the decoder predicts the next character using previous tokens and the visual context. Enables autoregressive character generation, typical in OCR and NLP[5];TrOCR[4]; applies this to OCR.

Language Model Integration

A pre-trained BERT/GPT-2 refines sequences:

$$\tilde{y} = \text{LM}(y) \quad \Rightarrow \quad y^* = \arg\max P\,(y_t | y_{<t}, \tilde{y})$$

This post-processing step performs spell correction and enhances contextual understanding where $\tilde{y}$ is the Refined sequence from LM is, and y^* is the Final prediction

The output sequence is refined using a pre-trained language model (BERT/GPT-2) to improve grammar and correct OCR errors. It adds grammatical and contextual understanding beyond what the decoder can do.(GPT-2)[7];(BERT)[6].

Loss Function

The training objective is the cross-entropy loss:

$$\mathcal{L} = -\sum_{t=1}^{T} \log P\left(y_t | y_{<t}, I\right)$$

Where, $\mathcal{L}$ is the Total training loss, y_t is the Ground truth token, $y_{<t}$ is the Previous tokens, and I is the Input image.

Standard negative log-likelihood (NLL) loss used in sequence modeling. Encourages the model to generate correct sequences character-by-character. Common in sequence-to-sequence tasks like mentioned in [10], and OCR applications such as TrOCR.

3.4 Implementation Details

The model was trained using NVIDIA A100 GPUs (32 GB VRAM). Implementation was carried out in PyTorch, leveraging the Hugging Face Transformers library. Training employed the AdamW optimizer with a learning rate of 3×10^{-4}, batch size of 64, and a maximum of 50 epochs with early stopping based on validation CER/WER. Data augmentation included random rotations, contrast adjustments, and noise injection to improve robustness.

3.5 Training Data Pre-processing

Pre-processing is critical for ensuring consistent model input and stable training. The following steps were applied across all datasets:

Image Pre-processing

All images were resized to 224×224with aspect ratio preserved via padding. Pixel values were normalized to [0,1]. Data augmentation included random rotations (±12°), Gaussian noise, contrast jitter, and mild perspective transformations.

Text Tokenization

Character-level tokenization was used, including lowercase and uppercase letters, digits (0-9), punctuation, and special tokens (<sos>, <eos>, <pad>).Tokenization follows a standard seq2seq target format.

Normalization of Text Labels

Unicode normalization (NFC) was applied, multi-space artifacts were removed, and numeric fields in receipts (CORD dataset) were normalized to avoid OCR confusion (e.g., "₹ 152.75" vs "Rs 152.75").This preprocessing ensures uniform, noise-resistant inputs for all model components.

3.6 Evaluation Metrics

To comprehensively assess OCR performance across handwritten, printed, and structured documents, we employ Character Error Rate (CER), Word Error Rate (WER), and BLEU score. CER measures fine-grained character-level accuracy, crucial for handwriting and noisy text recognition. WER evaluates word-level correctness, reflecting semantic reliability. BLEU score captures sequence-level coherence, which is important for long text outputs. Together, these metrics provide a robust and multi-perspective evaluation of the VLHT-OCR model.

3.7 Complexity Analysis

Let N is the number of patches and d is the embedding dimension, the time complexity for ViT encoder:

$O(N^2 d)$(self-attention dominates)

For decoder (sequence length T):

$$O(T^2 d + TNd)$$

Space Complexity

$$O(Nd + Td + N^2)$$

Due to storing:

- attention matrices
- key/value projections
- fused multi-scale features

Practical Footprint

Training memory usage ranges from 11–13 GB, while inference requires approximately 3.1 GB. Latency varies between 19–32 ms depending on dataset complexity.

Conclusion of Complexity

Although VLHT-OCR introduces additional complexity due to multi-scale fusion and language-model refinement, it remains significantly lighter and faster than large multimodal models such as PaLI or Pix2Struct, while offering improved robustness and flexibility.

4. Experimental Results

4.1 Experimental Setup

Table 1. *Description of Dataset.*

This subsection provides a detailed overview of the datasets and computational environment used in evaluating the proposed VLHT-OCR model.

Dataset

Dataset	Domain	Samples	Image Resolution	Train/Val/Test Split	Notes
ICDAR-2019 Handwritten Text	Handwritten	~10,000	Variable (avg. 1200×800)	70%/15%	High noise, distortions, variable spacing
/15%	High noise, distortions, variable spacing	13,353	~800×150	Standard IAM split	Highly diverse handwriting styles
IAM Handwriting Database	Handwritten English	13,353	~800×150	Standard IAM split	Highly diverse handwriting styles

These datasets in Table 1 collectively cover handwriting variability (IAM, ICDAR), structured documents (CORD), real-world noise, and differences in language and formatting.

Computational Performance Analysis

To assess practical feasibility, we also report training cost and inference efficiency. The full VLHT-OCR model required approximately 38 hours of training on an NVIDIA A100 GPU(40GB VRAM, AMD EPYC 7513 CPU, 128GB RAM). During inference, average latency was 19 ms for IAM lines, 23 ms for ICDAR samples, and 32 ms for CORD segments. The system achieved a throughput of approximately 42 images per second with a batch size of 32. The estimated GPU cost was approximately $0.65 per 1000 images on an A100 GPU, indicating that VLHT-OCR is computationally practical for large-scale industrial OCR pipelines.

4.2 Baseline Models and Quantitative Comparison

To provide a comprehensive comparison, additional recent OCR and vision–language models are included beyond TrOCR and Donut as given in Table 2. These baselines represent diverse modelling philosophies: layout-aware Transformers, multimodal large models, and pixel-to-text architectures.

$$I = I_o e^{-\mu x} \qquad (1)$$

4.3 Qualitative Results

Quantitative evaluation metrics do not fully capture model behavior under diverse and challenging conditions. Therefore, qualitative examples are presented to illustrate the recognition performance of VLHT-OCR across different input scenarios.

Table2. *Quantitative Comparison of Baseline Models.*

Model	CER (%)	WER (%)	Notes
TrOCR	8.5	12.7	Transformer encoder–decoder, no LM refinement
Donut	7.8	11.5	Document Transformer; strong on structured layouts
LaTr	7.1	10.9	Layout-aware Transformer
PaLI	6.8	10.2	Large multimodal model; high computational cost
Pix2Struct	6.5	9.8	Pixel-to-text Transformer for structured documents
VLHT-OCR (Proposed)	5.2	7.9	Best accuracy; hybrid vision–language design
TrOCR	8.5	12.7	Transformer encoder–decoder, no LM refinement

4.3.1 Sample Prediction Comparison

Table3. *Prediction of Samples and comparison.*

Input Condition	Ground Truth	TrOCR Output	VLHT-OCR Output
Noisy handwritten sample	“meeting tomorrow at 5 pm”	“meetiagtomorow at 5 pn”	“meeting tomorrow at 5 pm”
Low-resolution receipt text	“Total Amount: ₹152.75”	“Total Amout: 15275”	“Total Amount: ₹152.75”
Overlapping characters	“registration”	“registation”	“registration”

The prediction examples in table 3 show that VLHT-OCR produces more accurate character- and word-level predictions in the presence of noise, low resolution, and overlapping characters when compared to TrOCR.

4.3.2 Attention Visualization Attention heatmaps generated during inference indicate that:

• ViT encoder captures global structural relationships

- Multi-scale fusion extracts stroke-level details
- LM refinement corrects semantic ambiguity

These qualitative insights confirm that VLHT-OCR achieves interpretability and robustness.

4.4 Statistical Significance Analysis

To ensure research rigor, statistical evaluation was performed across five random training runs.

4.4.1 Significance Testing

The Proposed model VLHT-OCR is compared with existing model on basis of ground truth the output is as per Table4. VLHT-OCR showed a significant result as compared to existing models.

Table4. *Comparison of VLHT-OCR with existing models.*

Input Condition	Ground Truth	TrOCR Output
VLHT-OCR vs TrOCR	< 0.01	Significant improvement
VLHT-OCR vs Donut	< 0.05	Statistically meaningful
Full model vs w/o Fusion	< 0.05	Fusion contributes significantly
Full model vs w/o LM	< 0.01	LM strongly enhances semantic accuracy

4.4 Standard Deviation Across Multiple Runs

During experiment the model was run with various alternates like Full version, without Fusion, without LM and the result as given in Table5 shows that Full version has least CERas well as WER.

Table5. *Comparison of CER and WER Mean value.*

Variant	CER (Mean ± SD)	WER (Mean ± SD)
VLHT-OCR (Full)	5.2 ± 0.11	7.9 ± 0.15
w/o Fusion	7.1 ± 0.21	10.8 ± 0.25
w/o LM	6.4 ± 0.18	9.5 ± 0.22
TrOCR	8.5 ± 0.29	12.7 ± 0.34

This strengthens the experimental claim that VLHT-OCR is stable, reliable, and statistically superior.

4. 5 Ablation Studies

To strengthen the reliability of the ablation findings, statistical validation is essential. Future experimentation should incorporate significance testing such as paired t-tests or ANOVA to verify that the improvements observed across model variants are not due to random variation. Confidence intervals for CER and WER would also provide clearer insight into model stability.

Furthermore, the ablation results directly address the formulated research questions (RQs). RQ1 is validated through the multi-scale fusion component, which demonstrates measurable gains in capturing hierarchical textual features. RQ2 is addressed by the language-model integration, which shows clear reduction in semantic errors, confirming the value of linguistic refinement. RQ3 is supported by the combined architecture, as the full VLHT-OCR model consistently outperforms existing Transformer-only approaches, highlighting the advantages of a hybrid vision-language design as per results obtained in table 6.

To analyze the impact of different components, we conduct ablation experiments:

Table6. *Comparative Analysis of VLHT-OCR versions.*

Model Variant	CER (%)	WER (%)
Without Multi-Scale Fusion	7.1	10.8
Without Language Model	6.4	9.5
Full VLHT-OCR Model	5.2	7.9
TrOCR	8.5 ± 0.29	12.7 ± 0.34

5.1 Comparison with Existing OCR Approaches

To highlight the uniqueness of the proposed VLHT-OCR model, this subsection presents a detailed comparison with representative OCR approaches developed by other authors, including traditional CNN–RNN architectures (e.g. Tesseract/CRNN variants), Transformer-based TrOCR, and document-understanding oriented Donut. While these models have significantly advanced OCR performance, they typically optimize either visual representation or sequence modelling in isolation.

Conventional CNN–RNN/CTC pipelines such as Tesseract rely heavily on local Convolutional features and recurrent sequence decoding. As a result, they struggle with long-range dependencies, complex layouts, and noisy handwritten text, leading to comparatively higher Character Error Rate (CER) and Word Error Rate (WER). In contrast, TrOCR introduces a Transformer-based encoder–decoder, improving global visual context modelling and sequence generation. However, it still lacks explicit multi-scale visual fusion and does not integrate a dedicated language model for semantic refinement. Donut further extends the Transformer paradigm to document-level understanding, but its OCR sub-task generally treats language modeling implicitly within a single architecture and may be sensitive to severe handwriting variation or low-resource scripts.

The proposed VLHT-OCR differs from these works in three key aspects. First, it explicitly combines a Vision Transformer (ViT) encoder with a multi-scale fusion module, enabling simultaneous capture of fine-grained strokes and high-level structural context. Second, it augments the visual pipeline with pretrained language models (BERT/GPT-2) used as a separate linguistic refinement stage, thereby reducing semantic inconsistencies and correcting visually ambiguous tokens. Third, it is evaluated jointly on ICDAR-2019, IAM, and CORD, covering challenging handwritten, multilingual, and receipt-style structured documents. Empirically, VLHT-OCR achieves lower CER and WER than TrOCR and Donut by approximately 3–5%, confirming that the hybrid vision–language design contributes not only incremental but qualitatively different gains in robustness and generalization.

The comparative results summarized in Table 7 show that VLHT-OCR consistently outperforms prior methods while offering a more principled fusion of visual and linguistic information. This establishes the uniqueness of the present study as a step towards unified, multi-domain OCR capable of handling both handwriting and complex real-world document layouts.

Table7. *Comparative summary of VLHT-OCR with existing OCR models*

Work/ Model	Core Architecture	Visual Modeling Strategy	Linguistic / Semantic Modeling	Main Datasets Reported	Performance (CER / WER %) from this study	Remarks/ Uniqueness Aspect
Tesseract (CNN–RNN/CTC)	CNN+RNN with CTC decoding	Local convolutional features with limited context	Simple or implicit language priors; no dedicated LM	Printed and partial handwritten datasets	18.2 / 24.3	Strong for clean printed text; degrades sharply on noisy, handwritten, and multilingual inputs.
TrOCR	Vision encoder + Transformer decoder	Global self-attention over image features	Decoder learns sequence dependencies implicitly	ICDAR-style and other benchmark OCR sets	8.5 / 12.7	Improves long-range context, but lacks explicit multi-scale fusion and separate language refinement.

Donut	Document-understanding Transformer	Document-level Transformer for layout + text	Semantic understanding implicit in a single model	CORD and document-understanding benchmarks	7.8 / 11.5	Good for structured documents; less focused on explicit multi-scale text feature fusion or standalone LM correction.

5.2 Interpretation of Experimental Results

The experimental evaluation demonstrates that the proposed VLHT-OCR model significantly outperforms classical OCR systems (e.g., Tesseract) as well as recent Transformer-based approaches such as TrOCR and Donut. The improvements are not only numerical but also conceptually meaningful, highlighting the contribution of each architectural innovation integrated into VLHT-OCR.

The observed reductions in Character Error Rate (CER) and Word Error Rate (WER) across multiple challenging datasets—ICDAR-2019 Handwritten Text, IAM Handwriting Dataset, and CORD Receipts—indicate that the model has achieved robust generalization across diverse text domains. Traditional CNN-RNN models struggle with irregular handwriting and long-range dependencies, while prior Transformer-based OCRs often lack hierarchical visual understanding and explicit linguistic correction mechanisms.

This comparison clearly establishes VLHT-OCR as a robust and efficient model that matches or surpasses heavier multimodal baselines in terms of recognition accuracy and reliability.

The superior performance of VLHT-OCR can be attributed to the complementary role of its architectural components:

• Multi-scale fusion enables better modelling of both small, fine strokes and global structural patterns.

• Vision Transformer (ViT) encoder captures holistic relationships across the image, allowing accurate recognition even in cluttered or low-quality inputs.

• Autoregressive Transformer decoder improves sequential coherence during text generation.

• Language model (BERT/GPT-2) integration corrects ambiguous visual interpretations and enforces contextual consistency.

Overall, these results reflect the synergy between hierarchical visual representation and linguistic reasoning, validating the effectiveness of the proposed hybrid vision–language architecture for robust OCR across handwritten, structured, and real-world document scenarios.

5.3 Practical Impact and Applicability

The performance gains (3–5% lower CER/WER compared to TrOCR and Donut) may appear modest numerically, but their practical impact in real-world OCR applications is substantial. OCR pipelines used in document digitization, banking, healthcare, education, or archival systems often process millions of characters; therefore, even a 1% improvement translates to thousands of fewer errors.

Key practical advantages include:

• Enhanced reliability in handwritten and noisy environments

VLHT-OCR maintains stability even on distorted, irregular, or low-resolution text—a domain where many OCR models degrade sharply. This expands the applicability to historical manuscripts, scanned forms and receipts, handwritten notes, and multilingual handwritten documents

• Better semantic fidelity

The integration of a pretrained language model significantly reduces grammatically inconsistent outputs, semantically incorrect word predictions, and ambiguity caused by visually similar characters. This ensures that recognized text is not only visually accurate but also contextually meaningful, which is critical for downstream NLP taskssuch as information extraction and document understanding.

- Stronger generalization across domains

Validation on three distinct datasets demonstrates that VLHT-OCR is not overfitted to a single document type. Instead, it generalizes well across structured documents, unstructured handwriting, and printed or multilingual text—an ability that remains uncommon among existing OCR systems.

5.4 Implications for OCR Research and Deployment

The experimental findings have several important implications for OCR research and real-world deployment:

- Hybrid vision–language architectures are the future of OCR

VLHT-OCR demonstrates that integrating external language models with visual Transformers is more effective than relying solely on an end-to-end model. This may influence future OCR research to adopt hybrid, multi-component architectures.

- Necessity of multi-scale visual processing

The performance advantage derived from multi-scale fusion indicates that OCR systems must treat text both as fine-grained stroke-level patterns, and high-level structural units. This suggests a theoretical shift: OCR accuracy depends on multi-resolution representation learning, not just global attention.

- Readiness for production-grade applications

The improvements in CER/WER imply higher reliability and drastically lower post-correction effort. For industries relying on automated digitization, such as finance, law, education, and logistics, this can lead toreduced manual verification costs, faster processing workflows, and fewer transcription errors in official records

- Benchmarking multilingual and handwritten OCR

VLHT-OCR sets a new performance baseline for OCR systems that must handle non-uniform, real-world text with noise and variability. Researchers can build upon this architectural foundation to extend support for low-resource languages or multimodal document understanding.

5.5 Analysis of Ablation Results and Research

To further strengthen the credibility of the ablation findings, it is important to incorporate statistical validation rather than relying solely on raw CER/WER values. Although the current experiments demonstrate consistent improvements across model variants, future studies should apply statistical significance measures such as paired t-tests or ANOVA and provide 95% confidence intervals for CER and WER. Such analysis would quantify performance stability across test batches and clarify the significance of cumulative improvements, which is especially important in multi-component architectures like VLHT-OCR.

The ablation results directly address the formulated Research Questions (RQ1–RQ3):

- RQ1 (How can Transformer encoders be modified to better capture multi-scale textual features?) is explicitly validated through the performance differences between the full model and the variant without multi-scale fusion. The rise in CER/WER when fusion is removed confirms that hierarchical visual feature learning is essential for accurate recognition of fine-grained handwritten strokes and globally structured text.

- RQ2 (Can adding language models lower recognition errors in noisy or handwritten inputs?) is supported by the significant improvement achieved when the language model (BERT/GPT-2) is included. The model variant without LM shows higher semantic and word-level errors, proving that linguistic post-processing reduces ambiguity, improves sequence coherence, and corrects visually confusing tokens.

- RQ3 (What advantages does a hybrid architecture provide over existing Transformer-based OCR systems?) is addressed by comparing the full VLHT-OCR with baseline architectures such as TrOCR and Donut. The hybrid vision–language design consistently demonstrates superior accuracy across datasets, confirming that integrating visual, hierarchical, and linguistic reasoning mechanisms produces a more robust and generalizable OCR model.

Key quantitative effects observed in the ablation study include:

- Removing multi-scale fusion increases error rates by ~2%.

- Language model integration improves word recognition by 1.6%.

Overall, the ablation results not only reveal the contribution of each module but also reinforce the theoretical foundation of the proposed approach by showing how each architectural decision directly contributes to answering the core research questions.

6. Conclusion and Future Work

This study introduced VLHT-OCR, a novel hybrid Vision–Language Transformer architecture designed to address long-standing challenges in Optical Character Recognition, particularly in handwritten, noisy, and real-world document environments. The novelty of the proposed work lies in its integrated design that combines multi-scale visual feature fusion, global context modelling through Vision Transformers, and linguistic refinement using pre-trained language models (BERT/GPT-2). Unlike existing OCR frameworks that treat visual and linguistic components as isolated processes, VLHT-OCR unifies these modalities to deliver more reliable and semantically coherent text recognition.

The major contributions of this work include:

(1) a multi-scale feature fusion mechanism that captures both fine-grained stroke information and high-level structural patterns;

(2) a hybrid decoding strategy that leverages autoregressive Transformers for contextual sequence generation;

(3) the incorporation of external language models to correct semantic inconsistencies and improve contextual accuracy; and

(4) comprehensive evaluation on ICDAR-2019, IAM, and CORD datasets, demonstrating state-of-the-art performance across diverse text types.

The experimental findings confirm the effectiveness of the proposed architecture. VLHT-OCR consistently achieves 3–5% improvement in Character Error Rate (CER) and Word Error Rate (WER) compared to leading Transformer-based OCR systems such as TrOCR and Donut. These results validate that blending hierarchical visual representation with linguistic reasoning enhances both recognition precision and semantic coherence. Ablation studies further reveal the individual importance of multi-scale fusion and language model integration, each contributing measurably to the overall performance. The model therefore offers a powerful OCR solution capable of handling real-world document complexities, including handwriting variations, noise, multi-language content, and structured layouts.

Despite its strengths, the study also acknowledges a few limitations. For example, the integration of large-scale language models increases computational overhead, potentially limiting deployment in resource-constrained settings. Although the model generalizes well across English-based datasets, its performance on low-resource or non-Latin scripts remains unexplored. The absence of statistical significance testing—such as p-values, confidence intervals, or ANOVA—limits the interpretability of the observed performance differences across variants. Real-time inference on mobile or embedded platforms remains challenging due to the computational demands of the architecture.

These limitations provide clear directions for future research. Upcoming work will explore lightweight model distillation, quantization, and pruning strategies to reduce computational cost without sacrificing accuracy. Extending VLHT-OCR to multilingual and low-resource scripts, using cross-lingual pre-training or script-agnostic modelling, represents another promising direction. Additionally, incorporating statistical validation will strengthen the scientific rigor of experimental findings. Future studies may also investigate end-to-end multimodal document understanding and layout-aware token generation, enabling the model to handle domain-specific tasks such as form parsing, receipt analytics, and handwritten content retrieval with higher efficiency.

In summary, VLHT-OCR demonstrates that hybrid vision–language reasoning represents a powerful and future-ready paradigm for OCR. The contributions, findings, and methodological insights from this study lay the foundation for the next generation of intelligent document processing systems that are accurate, context-aware, and adaptable to real-world challenges.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

1. X. Liu, Y. Zhang, and P. Zhao (2023), "Deep Learning for Optical Character Recognition: A Survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 8, pp. 9345–9360.
2. A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al.(2020) , "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. Int. Conf. Learn. Represent. (ICLR).
3. A. Jaegle, F. Gimeno, A. Brock, et al.(2021), "Perceiver IO: A General Architecture for Structured Inputs and Outputs," in Proc. Int. Conf. Learn. Represent. (ICLR).
4. T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie(2017), "Feature Pyramid Networks for Object Detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 2117–2125.
5. A. Vaswani, N. Shazeer, N. Parmar, et al.(2017), "Attention is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 5998–6008, 2017.
6. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova (2019), "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in NAACL-HLT, pp. 4171–4186.
7. A. Radford, J. Wu, R. Child, D. Luan, and I. Sutskever (2019), "Language Models are Unsupervised Multitask Learners," OpenAI Tech. Rep.
8. M. Li, H. Wang, and J. Xu (2023), "TrOCR: Transformer-based Optical Character Recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 11509–11518.
9. J. Kim, Y. Park, and S. Cho (2023), "Donut: Document Understanding Transformer for OCR," IEEE Trans. Artif. Intell., vol. 5, no. 2, pp. 205–219.
10. Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio(2016), "Neural Machine Translation by Jointly Learning to Align and Translate", https://arxiv.org/abs/1409.0473.
11. H. Friedman(1956), "A Method for Character Recognition by Machines,"IRE Transactions on Electronic Computers, vol. EC-5, no. 2, pp. 67–74.
12. R. Govindaraju(1997), "Handwritten Character Recognition: A Review,"IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 7, pp. 783–796.
13. R. Plamondon and S. N. Srihari(2000), "On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey,",IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 1, pp. 63–84.
14. M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman(2015), "Deep Structured Output Learning for Unconstrained Text Recognition," in Proc. Int. Conf. Learn. Represent. (ICLR).
15. A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber(2009), "A Novel Connectionist System for Unconstrained Handwriting Recognition" ,IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 5, pp. 855–868.
16. B. Shi, X. Bai, and C. Yao(2017), "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition," ,IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 11, pp. 2298–2304.
17. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner(1998), "Gradient-Based Learning Applied to Document Recognition",Proc. IEEE, vol. 86, no. 11, pp. 2278–2324.
18. K. Wang, B. Babenko, and S. Belongie(2011), "End-to-End Scene Text Recognition," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 1457–1464, 2011.