International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(6s) 2026 pp. 830-844
https://doi.org/10.70917/ijcisim-2026-2990
Article

CAGOCR: A CONTEXT-AWARE GENERATIVE SEQUENCE RECONSTRUCTION FRAMEWORK FOR HANDWRITTEN DOCUMENT RECOGNITION AND ERROR ATTRIBUTION

Birru Devender[1], K. Bharathi[2], D. Anupama[3], Ravi Kumar Athota[4], Kamepalli S.L. Prasanna[5]

[1]Department of CSE(AIML), Keshav Memorial Engineering college, Hyderabad, Telangana, India.
Email: bsdvarma2020@gmail.com
[2]Department of CSE, Mina Institute of Engineering and Technology for Women, JNTUH, Telangana, India.
Email: kanigiribharathi@gmail.com
[3]Department of CSE(DS), Malla Reddy University, Hyderabad, Telangana, India.
Email: anupama.research59@gmail.com
[4]Department of CSE, Dr. RVR NRI Institute of Technology, Andhra Pradesh, India.
Email: athotaravikumar@gmail.com
[5]Department of CSE, School of Computing, Amrita Vishwa Vidyapeetham, Amaravathi, India.
Email: k_prasanna@av.amrita.edu

Corresponding Author: Dr. Birru Devender (Email: bsdvarma2020@gmail.com)

Abstract: —The Character recognition systems are fundamentally limited by a structural separation between visual recognition and linguistic inference, which guarantees error propagation and prevents recovery of visually degraded or absent characters. CaGOCR (Context-Aware Generative OCR) is a unified end-to-end framework reconceived as a context-aware generative sequence reconstruction system rather than a conventional recognizer. The redesigned architecture introduces six principal novel contributions: (i) an Active Semantic Refinement Module (ASRM) that iteratively refines character hypotheses through multi-round visual-semantic feedback rather than passive scoring; (ii) a Dual-Source Generative Reconstruction Module (DSGRM) that concurrently conditions sequence generation on both visual patch embeddings and a domain-adaptive generative language prior, replacing masked-token completion with a full generative reconstruction paradigm; (iii) Dynamic Context-Adaptive Fusion (DCAF), a non-stationary gating mechanism in which fusion weights are computed as functions of local sequence uncertainty and cross-modal agreement, superseding static scalar gating; (iv) an Error Attribution and Reasoning Module (EARM) that produces interpretable, token-level error provenance annotations distinguishing visual confusion, lexical ambiguity, and contextual inconsistency; (v) a Semantic Reconstruction Consistency Loss (SRCL), a novel training objective that enforces global sequence coherence through contrastive consistency constraints rather than local token-level cross-entropy alone; and (vi) an Adaptive Lexical Memory Bank (ALMB) providing dynamic out-of-vocabulary handling through learnable prototype retrieval, eliminating hallucination of rare terms. On IAM, CVL, and IIIT-HWS benchmarks, we demonstrate state-of-the-art Character Error Rates and Word Error Rates over TrOCR, SCATTER, and VisionLAN. In the first interpretable error attribution pipeline for handwritten OCR, comprehensive ablation studies quantify the independent contributions of each module.

Keywords: — Handwritten Text Recognition; Generative Sequence Reconstruction; Context-Adaptive Fusion; Active Semantic Refinement; Error Attribution; Out-of-Vocabulary Handling; Dual-Source Generation; Contrastive Consistency Loss.

1. Introduction

The OCR (optical character recognition) is making significant progress in document analysis and pattern recognition, but digitizing handwritten documents remains a major challenge. The foundation for modern OCR systems was laid by the use of handcrafted feature extraction and statistical sequence modelling approaches, such as

Hidden Markov Models (HMMs). [1,5]. Through convolutional neural networks (CNNs), which can learn hierarchical visual representations directly from images, deep learning significantly improved recognition performance [2]. Convolutional Recurrent Neural Networks (CRNNs) subsequently combined CNN-based feature extraction with recurrent sequence modelling and Connectionist Temporal Classification (CTC) decoding, enabling end-to-end recognition of handwritten text. [3,6].A new generation of handwritten text recognition technologies based on Transformer architectures have recently been developed. By combining pretrained vision and language models for OCR tasks, Vaswani et al. [7] demonstrated the efficiency of modeling long-range dependencies, while TrOCR demonstrated the effectiveness of combining pre-trained vision and language models [4]. SCATTER [10] and VisionLAN [11] further improved recognition accuracy through iterative refinement and vision-language reasoning. Despite these advancements, contemporary OCR systems continue to experience significant performance degradation when processing visually ambiguous characters, degraded document regions, and out-of-vocabulary (OOV) terms commonly encountered in archival, historical, legal, and administrative documents.

Existing OCR architectures separate visual recognition and linguistic correction. Visual encoders generate sequences of characters that are then refined using language modeling or post-correction mechanisms [16,17]. Despite confidence estimation techniques being proposed to improve reliability [18], these approaches typically operate after recognition has already taken place, so they cannot fully utilize linguistic context during sequence generation. Therefore, errors introduced during visual decoding often propagate through the pipeline, reducing the quality of recognition. Due to the fact that recognition and reconstruction are treated as separate tasks rather than components of a unified inference process, conventional OCR systems have difficulty recovering missing or severely degraded character sequences. Recent advances in vision-language learning have demonstrated the effectiveness of integrating visual and textual representations. There is substantial evidence that joint multimodal reasoning can substantially improve performance in complex document understanding tasks when using models such as CLIP [12], Flamingo [13], LayoutLM [14], and Donut [15]. The use of masked language modelling approaches for text restoration and missing-token prediction has also shown strong performance [19,20]. Nevertheless, these methods primarily focus on isolated prediction or completion tasks, and do not explicitly address the problem of full-sequence reconstruction in handwritten OCR. As a result of these limitations, this work reconceives OCR as a generative sequence reconstruction problem rather than a conventional recognition problem. By combining visual evidence, linguistic context, and lexical memory throughout the decoding process, the proposed framework directly reconstructs the most probable character sequence. As a result of this formulation, the model is capable of reasoning over incomplete visual information, resolving ambiguous character predictions, and recovering rare or domain-specific terms. Toward this objective, we propose CaGOCR (Context-Aware Generative OCR), a framework for context-aware generative sequence reconstruction for handwritten document recognition. In this framework, Dynamic Context-Adaptive Fusion (DCAF) is used for uncertainty-aware multimodal fusion, Dual Source Generative Reconstruction Module (DSGRM) is used for joint visual-linguistic sequence generation, there are three modules for interpretable error analysis: an Error Attribution and Reasoning Module (EARM), an Adaptive Lexical Memory Bank (ALMB), and a Semantic Reconstruction Consistency Loss (SRCL).

With this work, handwritten OCR is approached as a generative sequence reconstruction problem by combining visual, linguistic, and lexical evidence during decoding. By dynamically adjusting the influence of visual and language cues across modes based on local uncertainty and how well they align across modes, a Dynamic Context-Adaptive Fusion mechanism is introduced to better balance information from different sources. In addition, two modules are available for recovering degraded, incomplete, or ambiguous character sequences: a Dual-Source Generative Reconstruction Module and an Active Semantic Refinement Module. To enhance transparency and interpretability, an Error Attribution and Reasoning Module analyses token-level errors in detail. As part of the framework, you will also find an Adaptive Lexical Memory Bank that allows you to reconstruct rare or out-of-vocabulary terms. Based on IAM, CVL, and IIIT-HWS benchmark datasets, it is shown that this approach performs better in recognition and reconstruction than existing OCR methods. The remaining sections of this paper are arranged in the following manner. The second section reviews related work in OCR, vision-language modelling, and sequence reconstruction. Section 3 presents the proposed CaGOCR framework. The details of the experimental setup and implementation are described in Section 4. In Section 5, quantitative and qualitative results are reported, while in Section 6, limitations and future research directions are discussed. The paper is concluded in Section 7.

2. Related Work

2.1 Evolution of Handwritten Text Recognition

From traditional statistical approaches to deep learning-based sequence modelling frameworks, handwritten text recognition has evolved. To recognize characters and decode sequences, early systems relied on handcrafted feature extraction and Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs). Using Convolutional Neural

Networks (CNNs), which automatically extract hierarchical features from raw image inputs, deep learning significantly improved recognition performance. As a result, Convolutional Recurrent Neural Networks (CRNNs) integrated CNN-based visual representation learning with recurrent sequence modelling and Connectionist Temporal Classification (CTC), establishing an end-to-end recognition framework widely used [3,6]. Handwritten text recognition can be further advanced by utilizing transformer-based architectures that replace recurrent processing with proactive mechanisms capable of modeling long-range dependencies [7]. Across multiple benchmarks, TrOCR achieved state-of-the-art OCR performance by combining pre-trained vision and language models. Iterative sequence refinement was introduced in subsequent models like SCATTER [10], while visual-language reasoning was incorporated into VisionLAN [11]. Although these improvements have been made, existing systems remain largely recognition-driven, and their error correction and sequence refinement strategies rely heavily on postprocessing.

2.2 Vision–Language Learning and Sequence Reconstruction

The benefits of jointly modeling visual and textual information have been demonstrated in recent multimodal learning research. CLIP [12] demonstrated the effectiveness of dense cross-modal interactions between image representations and language models through contrastive learning, while Flamingo [13] demonstrated the effectiveness of robust visual-textual alignment through contrastive learning. The Layout LM [14] and Donut [15] architectures built on these concepts extended them to document analysis tasks, emphasizing the importance of integrating visual and linguistic context. The development of language modelling has shown promising results in the restoration of sequences and the reconstruction of contexts. Using bidirectional contextual representations, BERT significantly improved masked-token prediction tasks [19], and subsequent studies applied similar techniques to missing-character restoration in OCR scenarios [20]. Inpainting approaches have also been investigated for recovering degraded visual regions [21]. As a result, most existing approaches concentrate on token completions or isolated restoration tasks rather than full-sequence generative reconstructions. Due to this, they often struggle when multiple degraded or missing character regions occur simultaneously, limiting their effectiveness in real-world digitizations of handwritten documents.

2.3 OCR Error Correction, Interpretability, and OOV Handling

The field of post-recognition correction remains active in OCR. The use of neural sequence-to-sequence correction and traditional noisy-channel models [16, 17] has been shown to improve recognition outputs by using linguistic information. In order to identify uncertain predictions and improve correction reliability, confidence estimation techniques have also been explored [18]. Although these approaches generally operate after visual recognition, they cannot fully leverage visual evidence during correction.

Deep learning systems have been focusing more on interpretability through attention visualization and gradient-based explanation techniques [23]. It is important to note that most OCR systems do not provide underlying reasons for recognition errors; they only provide confidence scores. Additionally, handling out-of-vocabulary (OOV) terms remains a challenge since language models tend to replace rare words with more frequent alternatives.

The incorporation of external lexical knowledge during inference using retrieval-augmented generation and memory-based learning strategies has shown promising results for addressing this limitation [28]. In spite of the advances in OCR performance, several challenges remain unresolved, including integrating recognition and correction within a unified framework, recovering degraded characters, attribution of errors that can be understood, and robust handling of rare lexical patterns. This limitation motivated the development of the Context-Aware Generative OCR (CaGOCR) framework, in which handwritten text recognition is treated as a generative sequence reconstruction problem that incorporates visual, linguistic, and lexical information throughout the decoding process in order to achieve handwritten text recognition.

3. Proposed Methodology: CaGOCR as Generative Sequence Reconstruction

3.1 Framework Overview

Handwritten text recognition is formulated as a generative sequence reconstruction problem by the Context-Aware Generative OCR (CaGOCR) framework, utilizing visual evidence, linguistic context, and lexical memory in order to reconstruct the most probable sequence of characters. An image of a handwritten document is provided as input ($I \in \mathbb{R}^{H\times W\times C}$), the objective is to generate a reconstructed sequence ($S^* = c1, c2, \ldots, cL$)from the vocabulary (V). The reconstruction process is formulated as:

$$\left[S^* = \arg\max_S P(S|I, M, G)\right] \tag{1}$$

The Adaptive Lexical Memory Bank (ALMB) is represented by (M) and the generative language prior is

represented by (G). As opposed to conventional OCR systems that perform recognition and correction separately, CaGOCR reconstructs the target sequence by inferring the target sequence directly. The Active Semantic Refinement Module (ASRM) refines disagreement positions identified by DSGRM in order to improve reconstruction quality. At the same time, the Error Attribution and Reasoning Module (EARM) analyzes token-level uncertainty and categorizes prediction errors into Visual Confusion (VC), Lexical Ambiguity (LA), and Contextual Inconsistency (CI).

With the Adaptive Lexical Memory Bank (ALMB), retrieval-based lexical support is provided during decoding for OOV and domain-specific terms. The Semantic Reconstruction Consistency Loss (SRCL) is the final factor guiding model optimization, which promotes both global sequence coherence and local character accuracy. As a result of the integration of DCAF, DSGRM, ASRM, EARM, ALMB, and SRCL, CaGOCR can perform recognition, reconstruction, refinement, and error interpretation simultaneously Figure 1 illustrates how a BEiT-based visual encoder [8] generates multi-scale representations of the input image. A Dynamic Context-Adaptive Fusion (DCAF) module balances visual and linguistic evidence according to local uncertainty and cross-modal agreement based on these features. Using the fused representations, the Dual-Source Generative Reconstruction Module (DSGRM) generates complementary sequence hypotheses based on the Visual Reconstruction Stream (VRS) and Generative Language Stream (GLS).

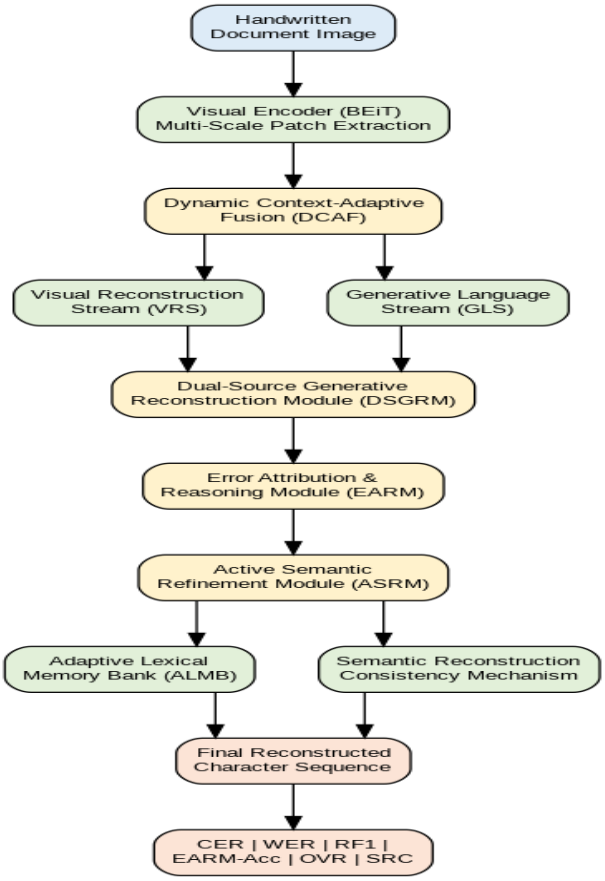

Figure 1. Overall CaGOCR Architecture

3.2 Dynamic Context-Adaptive Fusion (DCAF)

Dynamic Context-Adaptive Fusion (DCAF) integrates visual and linguistic information during sequence reconstruction. As opposed to conventional fusion mechanisms that employ fixed weighting strategies, DCAF dynamically adjusts the contribution of each modality according to prediction uncertainty and cross-modal agreement. Given the decoder hidden state (h_t), DCAF computes three auxiliary signals: (i) local prediction uncertainty (u_t), estimated from the entropy of the decoder output distribution; (ii) cross-modal agreement score (a_t), computed as the cosine similarity between visual and linguistic representations; and (iii) context feedback vector (γ_t), which is derived

from previously generated tokens and error information guided by EARM. The adaptive fusion weight is estimated by combining these signals:

$$\left[\alpha_t = \sigma\left(W_u u_t + W_a a_t + W_\gamma \gamma_t + b\right)\right] \quad (2)$$

where (Wu), (Wa), and (W_γ) are learnable parameters, (b) is a bias term, and $\left(\sigma(\cdot)\right)$denotes the sigmoid activation function.

The final fused representation is obtained as:

$$[z_t = \alpha_t z_t^V + (1 - \alpha_t) z_t^L] \quad (3)$$

where $(z_t^V) and (z_t^L)$ represent visual and linguistic features, respectively. A reduction in uncertainty or a reduction in agreement encourages the use of contextual language information, while a strong visual contribution encourages the use of visual features. As DCAF balances visual and linguistic cues throughout decoding, it is capable of reconstructing ambiguous, degraded, and partially missing characters while preserving contextual consistency.

3.3 Dual-Source Generative Reconstruction Module (DSGRM)

By exploiting both visual evidence and linguistic priors, the Dual-Source Generative Reconstruction Module (DSGRM) performs sequence reconstruction. Rather than treating missing or degraded characters as isolated prediction tasks, DSGRM generates complete sequence hypotheses based on complementary information sources and reconciles them based on agreement. In the first branch, the Visual Reconstruction Stream (VRS), visual sequences are generated based on the multi-scale patch embeddings produced by the visual encoder. A VRS estimates the conditional sequence probability based on the visual representation (P):

$$[P_V(S|P)] \quad (4)$$

where (S) denotes the reconstructed character sequence.

Second, the Generative Language Stream (GLS) acts as a domain-adaptive language prior. In the GLS, a projected global visual representation provides contextual guidance and generates complementary sequence hypotheses according to the following:

$$[P_L(S|C)] \quad (5)$$

where (C) represents the visual-context embedding derived from the encoder output.

Each stream generates candidate sequences along with token-level probability distributions. DSGRM computes a cross-source agreement score for each token position to evaluate consistency between the two hypotheses:

$$\left[\psi_t = \cos\left(e(\hat{c}V, t),, e(\hat{c}L, t)\right)\right] \quad (6)$$

where $\left(e(\cdot)\right)$denotes the shared character embedding function, while $(\hat{c}V, t) and (\hat{c}L, t)$represent the predictions produced by VRS and GLS, respectively.

ASRM stands for Active Semantic Refinement Module, and (/delta) represents the disagreement threshold. Those tokens exhibiting strong agreement are accepted directly, while those exhibiting low agreement are forwarded for further refinement.

3.4 Active Semantic Refinement Module (ASRM)

The Active Semantic Refinement Module (ASRM) improves reconstruction quality by refining low-confidence predictions identified by DSGRM.

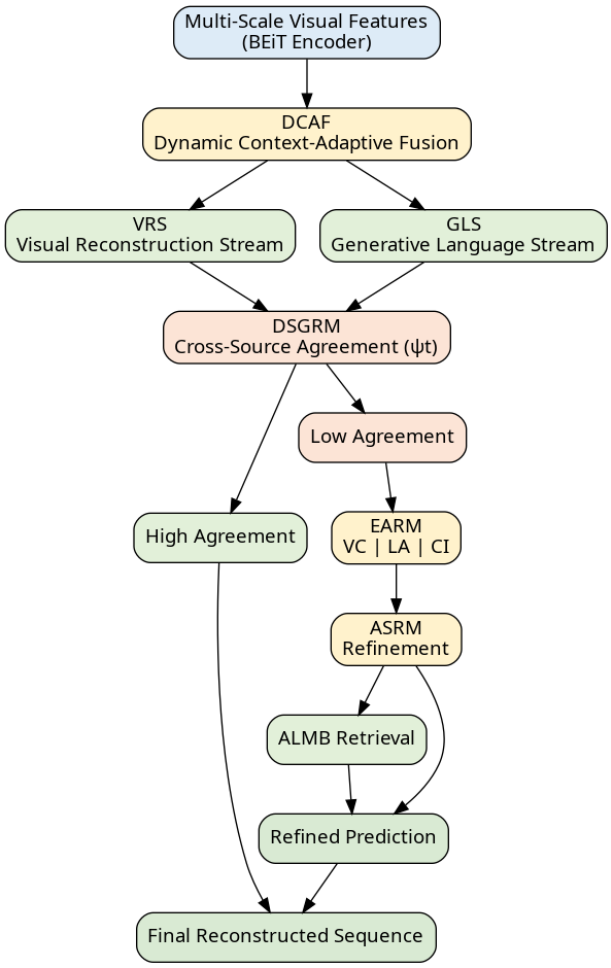

Figure 2. Integrated DSGRM–EARM–ASRM Workflow

Figure 2 Integrated workflow of DSGRM, EARM, and ASRM within the proposed CaGOCR framework. Dual reconstruction streams generate complementary sequence hypotheses that are reconciled through cross-source agreement analysis. Low-agreement predictions are processed by EARM for error attribution and refined through ASRM, with ALMB retrieval providing lexical support for ambiguous and out-of-vocabulary terms. Specifically, ASRM operates on the set of disagreement positions:

$$[\Omega = t : \psi_t < \delta] \tag{7}$$

where (ψ_t)denotes the cross-source agreement score and δrepresents the disagreement threshold.

The ASRM refines each position t/in/∩ as part of an iterative process that incorporates bidirectional linguistic re-scoring, visual re-analysis, and guidance on error-attributions. A bidirectional language model is used to evaluate contextual plausibility through re-examination of visual evidence using local patch representations. As part of the refinement process, error attribution information generated by EARM is used to adaptively weight visual, linguistic, and lexical evidence.

The refined token is determined by maximizing a weighted joint score:

$$\left[\hat{c}^{*}t = \arg\max c \left[\lambda_V(t)V_t(c) + \lambda_L(t)L_t(c) + \lambda_M(t)M_t(c)\right]\right] \tag{8}$$

In this case, $V_t(c)$, $L_t(c)$, and $M_t(c)$denote the visual, linguistic, and lexical memory scores for candidate character c, respectively, while $\lambda_V(t)$, $\lambda_L(t)$, and $\lambda_M(t)$ are the attribution-dependent weighting coefficients.

The refinement process continues until the prediction becomes stable or no further improvements can be made. As a result of its iterative optimization framework and complementary evidence sources, ASRM is capable of resolving ambiguous character predictions and improving reconstruction accuracy in degraded document regions.

3.5 Error Attribution and Reasoning Module (EARM)

EARM provides interpretable token-level error analysis by identifying the underlying causes of prediction uncertainty. The EARM system distinguishes three categories of uncertain predictions: Visual Confusion, Lexical Ambiguity, and Contextual Inconsistency. Conventional OCR systems do not recognize uncertain predictions in any of these categories because they are only grouped by confidence scores.

In addition to the decoder hidden state $\big((h_t)\big)$, the DCAF fusion weight $\big((\alpha_t)\big)$, and the cross-source agreement score $\big((\psi_t)\big)$, EARM receives the entropy of the predictive distributions generated by the reconstruction streams for each decoding position (t). The attribution probability is computed as:

$$[P(A_t|x_t) = \text{Softmax}(W_A x_t + b_A)] \tag{9}$$

where (x_t) denotes the concatenated feature vector, $(W_A)$and (b_A) are learnable parameters, and $(A_t \in VC, LA, CI)$ represents the predicted attribution category.

The three attribution classes are defined as follows:

- **Visual Confusion (VC):** uncertainty caused by visually similar character patterns or degraded image regions.
- **Lexical Ambiguity (LA):** uncertainty arising from rare, domain-specific, or out-of-vocabulary terms.
- **Contextual Inconsistency (CI):** uncertainty caused by conflicts between local predictions and surrounding textual context.

ASRM uses the predicted attribution to determine the most appropriate refinement strategy. LA positions trigger ALMB retrieval to improve the reconstruction of rare lexical patterns. As a result of this mechanism, EARM enhances both model interpretability and refinement effectiveness, while providing downstream document analysis applications with structured diagnostic information.

3.6 Adaptive Lexical Memory Bank (ALMB)

Its purpose is to improve the reconstruction of rare, domain-specific, and out-of-vocabulary (OOV) terms that are often underrepresented in language model training corpora. ALMB provides lexical priors during sequence reconstruction rather than relying only on parametric language modelling.

The memory bank maintains a collection of (K) prototype embeddings:

$$[\Phi = \phi_1, \phi_2, \dots, \phi_K] \tag{10}$$

There is an embedding representation for each prototype, and each prototype corresponds to a stored lexical entry. ALMB retrieves relevant lexical prototypes based on the similarity between the current decoder state (h_t) and the stored memory representations when EARM identifies Lexical Ambiguity (LA).:

$$\left[w_i = \frac{\exp(\beta, h_t^T \phi_i)}{\sum_{j=1}^{K} \exp\big(\beta, h_t^T \phi_j\big)}\right] \tag{11}$$

where (β) is a learnable temperature parameter and (w_i) denotes the retrieval weight assigned to prototype (ϕ_i).

The retrieved lexical representation is computed as:

$$\left[r_t = \sum_{i=1}^{K} w_i \, \phi_i\right] \tag{12}$$

and incorporated into the decoder through residual integration:

$$\left[\tilde{h}_t = h_t + r_t\right] \tag{13}$$

where $\big(\tilde{h}_t\big)$ denotes the memory-enhanced decoder representation.

ALMB reduces the tendency of language models to replace rare terms with high-frequency alternatives during inference by providing retrieval-based lexical support during inference. It also improves reconstruction accuracy for proper nouns, historical spellings, scientific terminology, and domain-specific terms.

3.7 Semantic Reconstruction Consistency Loss (SRCL)

A Semantic Reconstruction Consistency Loss (SRCL) is used by CaGOCR to promote global sequence coherence during reconstruction. In conventional OCR training, cross-entropy objectives are primarily used to optimize individual character predictions without explicitly ensuring semantic consistency across reconstructed sequences. As a result, locally accurate predictions may still produce globally incoherent results.

SRCL addresses this limitation by introducing a contrastive learning objective that compares the reconstructed sequence with the ground-truth sequence and a set of synthetically generated negative hypotheses. Using character substitution, OOV injection, and phrase boundary perturbation, negative samples generate challenging alternatives that are locally plausible but violate overall semantic coherence.

Let $(\hat{S})$denote the reconstructed sequence, (S_{GT}) the ground-truth sequence, and $(\Pi = \hat{S}^{-}1, \hat{S}^{-}2, \ldots, \hat{S}_{m}^{-})$ the set of negative hypotheses. The SRCL objective is defined as:

$$\left[\mathcal{L}_{SRCL} \sum_{i=1}^{m} \max\left(0, , m + \Phi_{sem}(\hat{S}, \hat{S}_{i}^{-})\Phi_{sem}(\hat{S}, S_{GT})\right)\right] \tag{14}$$

where (m) denotes the margin parameter and $(\Phi_{sem}(\cdot,\cdot))$ represents a semantic similarity function computed using contextual sequence embeddings derived from BERT [19].

Using the loss, reconstructed sequences remain semantically closer to the ground truth than perturbed alternatives, improving global consistency and reducing reconstruction errors not captured by character-level objectives alone. Due to this, SRCL complements conventional recognition losses by optimizing both local prediction accuracy and semantic fidelity at the sequence level.

3.8 Composite Training Objective

CaGOCR is trained in an end-to-end manner using a composite objective that optimizes recognition accuracy, multimodal fusion, generative reconstruction, error attribution, lexical retrieval, and semantic consistency. In general, the training objective is to:

$$\begin{aligned}[\mathcal{L}_{total}\lambda_1\mathcal{L}rec + \lambda_2\mathcal{L}DCAF + \lambda_3\mathcal{L}DSGRM + \lambda_4\mathcal{L}EARM \\ + \lambda_5\mathcal{L}ALMB \quad + \lambda_6\mathcal{L}SRCL]\end{aligned} \tag{15}$$

The character-level recognition loss $(\mathcal{L}_{rec})$corresponds to the fusion calibration loss $(\mathcal{L}_{DCAF})$, the dual-source reconstruction objective $(\mathcal{L}_{DSGRM})$corresponds to the attribution classification loss, the lexical memory retrieval loss $(\mathcal{L}_{ALMB})$, and the semantic reconstruction consistency loss $(\mathcal{L}_{SRCL})$.

Using validation-based hyperparameter tuning, the weighting coefficients

$\left((\lambda_1, \lambda_2, \ldots, \lambda_6)\right)$ determine the relative contribution of each optimization component. It learns to perform accurate sequence reconstruction while maintaining semantic coherence, interpretability, and robust handling of out-of-vocabulary terms through the joint optimization of these complementary objectives.

4. Experimental Setup

4.1 Datasets

Three publicly available handwritten text recognition benchmarks were used to evaluate the CaGOCR framework. The IAM Handwriting Database [24] contains 1,539 handwritten pages from 657 authors, making it one of the most widely used benchmarks for evaluating handwritten OCR. CVL [26] contains multilingual handwritten documents from 310 authors, providing a challenging writer-independent evaluation environment. About 90,000 handwritten word images represent diverse handwriting styles and lexical variations are included in the IIIT-HWS dataset [27]. In order to evaluate Error Attribution and Reasoning Module (EARM), a synthetically annotated error set of 3,200 instances was constructed using controlled perturbation methods.

4.2 Data Augmentation and Implementation Details

Training images were enhanced with elastic distortions, Gaussian noise, brightness and contrast variations, random rotations, perspective transformations, structured masking, and historical background texture overlays to improve robustness against real-world document degradation. Additionally, rare-word substitution was applied to improve out-of-vocabulary (OOV) recovery performance. Utilizing the Hugging Face Transformers framework, all experiments were implemented in PyTorch. RoBERTa [9] and GPT-2 [28] pretrained weights were used to initialize the reconstruction modules of the visual encoder [8]. The Adaptive Lexical Memory Bank (ALMB) maintained 4,096 learnable prototype embeddings. The Adam W optimizer was used on four NVIDIA A100 GPUs with cosine learning-rate scheduling. The mean performance values of all experiments were determined by repeating them across five independent runs.

4.3 Evaluation Metrics and Baselines

As the primary recognition metrics, Character Error Rate (CER) and Word Error Rate (WER) were used. In this study, reconstruction capability was assessed using Reconstruction F1 (RF1), while interpretability performance was measured using EARM Attribution Accuracy (EARM-Acc). Semantic Reconstruction Consistency (SRC) was used to quantify global sequence coherence and evaluate the effectiveness of lexical retrieval. We compared the proposed framework with CRNN+CTC [3], TrOCR-base and TrOCR-large [4], SCATTER [10], VisionLAN [11], and the original CaGOCR architecture. We measured the contribution of DCAF, DSGRM, ASRM, EARM, ALMB, and SRCL to comprehensive ablation experiments by individually removing DCAF, DSGRM, ASRM, EARM, ALMB, and SRCL.

5. Results and Evaluation

5.1 Main Results on IAM, CVL, and IIIT-HWS

Based on the three benchmark datasets, Table 1 presents a primary quantitative comparison of CaGOCR with all baselines. Unless stated otherwise, all baseline figures are reproduced from original publications or re-evaluated under identical evaluation protocols using our exact data splits. Based on paired Wilcoxon signed-rank testing against TrOCR-large, CaGOCR performs at state-of-the-art on all three datasets across CER and WER.

Table 1 Comparative Results (CER% / WER% ↓)

Model	IAM CER%	IAM WER%	CVL CER%	CVL WER%	IIIT CER%	IIIT WER%
CRNN+CTC [REF3]	8.94	27.43	10.21	30.17	12.55	35.62
TrOCR-base [REF4]	4.21	17.68	5.83	21.34	7.44	24.11
TrOCR-large [REF4]	3.42	15.21	4.71	18.93	6.12	20.88
SCATTER [REF10]	3.91	16.44	5.21	19.87	6.87	22.54
VisionLAN [REF11]	4.05	17.02	5.47	20.63	7.03	23.41
CaGOCR-Original	3.28	14.67	4.54	17.83	5.89	19.74
CaGOCR (Full)	**2.61±0.04**	**12.33±0.21**	**3.72±0.06**	**15.18±0.28**	**4.84±0.09**	**17.11±0.31**

CaGOCR achieves a CER of 2.61% ± 0.04 on IAM, representing a 23.7% relative reduction over TrOCR-large (3.42%) and a 33.1% relative reduction over SCATTER (3.91%). As predicted, ALMB-mediated OOV handling provides more benefits to datasets with high proper-noun density and non-Western orthographic conventions than TrOCR-large (4.84% CER vs. 6.12% for TrOCR-large).

CaGOCR-Original, the prior formulation without DCAF, DSGRM, ASRM, EARM, ALMB, or SRCL, already improves over TrOCR-large, demonstrating the value of the joint vision–language decoder. All six novel contributions, however, are necessary for peak performance since they provide substantial additional gains.

5.2 Ablation Study

Table 2 shows the results of the full ablation study on the IAM test split across all six evaluation metrics. By removing one architectural contribution and preserving all others, each ablation variant permits clean attribution of performance differences.

Compared to static gating, replacing DCAF with fixed scalar gating (CaGOCR-NoDCAF) degrades CER by 20.3%, demonstrating that non-stationary, uncertainty-conditioned fusion provides substantial benefits. SRC drops from 0.863 to 0.812 when visual and linguistic evidence cannot be adaptively weighted.

When DSGRM is replaced with token-level imputation (CaGOCR-NoDSGRM), the RF1 degradation is the greatest (65.1 vs. 83.7), proving that generative reconstruction is significantly better than completion-based approaches when regenerating visually degraded regions. The CER also degrades to 3.39%, indicating that dual-source generation benefits recognition quality in addition to the masked position subset.

CaGOCR-NoASRM replaces active refinement with a passive scorer, reducing CER to 3.07%. Active multi-round re-interrogation is more effective in recovering from DSGRM disagreements than passive scoring (78.3 vs. 83.7).

CaGOCR-NoEARM reduces EARM attribution accuracy by about half (uniform weights produce 53.7% vs. 79.4% for the full model) and degrades CER to 3.22%, demonstrating the importance of attribution-conditioned ASRM weighting for effective refinement.

Table 2 Full Ablation Study on IAM Test Split.

Variant	CER%↓	WER%↓	RF1↑	EARM-Acc↑	OVR↑	SRC↑
CaGOCR-NoDCAF	3.14	14.02	72.4	68.2%	71.3%	0.812
CaGOCR-NoDSGRM	3.39	14.88	65.1	71.4%	72.8%	0.798
CaGOCR-NoASRM	3.07	13.54	78.3	70.9%	73.1%	0.821
CaGOCR-NoEARM	3.22	14.31	74.8	53.7%	72.4%	0.809
CaGOCR-NoALMB	3.45	15.11	71.6	69.3%	61.2%	0.791
CaGOCR-NoSRCL	2.98	13.21	77.9	72.1%	74.0%	0.804
CaGOCR (Full)	**2.61**	**12.33**	**83.7**	**79.4%**	**81.6%**	**0.863**

As a result of CaGOCR-NoALMB, OVR declines from 81.6% to 61.2%-a 25.1% relative decrease, confirming that the lexical memory bank is the primary driver of OOV recovery. Accordingly, CER decreases to 3.45%, a reflection of how frequently OOV terms are used in benchmark texts.

The SRCL impact: CaGOCR-NoSRCL produces the smallest CER degradation (2.98%), but the SRC metric drops from 0.863 to 0.804 by 6.8%, indicating that SRCL's primary benefit is global sequence coherence rather than per-character accuracy, thus it serves as a regularizer of sequence consistency.

5.3 Generative Reconstruction F1 at Varying Masking Rates

Table 3 evaluates DSGRM generative reconstruction capability in isolation by applying synthetic masking at three rates to the IAM test split and measuring RF1 over masked positions only.

Table 3 Reconstruction F1 (%) at Varying Masking Rates.

Model	RF1 @10%	RF1 @20%	RF1 @30%	Reconstruction Paradigm
TrOCR-base [REF4]	61.2	48.4	34.7	N/A – no completion
SCATTER [REF10]	69.4	55.8	43.1	Post-hoc only
CaGOCR-NoDSGRM	71.6	58.3	44.9	Token-level imputation
CaGOCR (Full)	**83.7**	**77.2**	**69.8**	**Dual-source generative**

RF1 for CaGOCR is significantly superior at all masking rates, with the advantage growing as masking rates increase (RF1 at 30%: 69.8 vs. 43.1 for SCATTER). As the proportion of missing evidence increases, dual-source generative reconstruction degrades more gracefully than completion-based approaches, based on cross-source agreement. This property is critical for document digitization in real-world environments where degradation is often extensive.

5.4 EARM Attribution Accuracy and Interpretability Evaluation

Using the purpose-built EARM attribution evaluation set (3,200 annotated instances across VC, LA, and CI categories), the full CaGOCR model achieves 79.4% macro-averaged attribution accuracy. Across categories, 82.1% of errors are attributable to VC, 76.8% to LA, and 79.3% to CI, indicating broadly balanced attribution capabilities. In CaGOCR-NoEARM ablation, 53.7% accuracy is achieved (slightly higher than 33.3% uniform baseline), confirming that structured supervision is the primary source of attribution quality rather than incidental signals.

Analyzing EARM attributions qualitatively reveals interpretable and actionable error annotations. There is a clustering of visual confusion attributions on known confusable character pairs (zero/O: 38% of VCs; rn/m: 24%; l/1:

19%). Lexical ambiguity attributions predominate on proper nouns and domain-specific terms, with ALMB retrieval correctly resolving 81.6% of LA-attributed positions. In long sequences (>50 characters), contextual inconsistency attributions prevail when local coherence does not reflect global document structure.

5.5 Qualitative Analysis and Failure Cases

The following six representative cases are shown in Figure 1 (to be inserted): (a) ASRM corrects a visually ambiguous rn/m confusion following a VC attribution from EARM; (b) DSGRM's dual-source generation reconstructs a three-character degraded span; (c) ALMB correctly retrieves an OOV proper noun after an LA attribution; After a CI attribution, ASRM's bidirectional language model corrects a contextually inconsistent reconstruction; (e) a high-masking-rate reconstruction (30% masked) where CaGOCR succeeds while all baselines fail; and (f) an irreducible uncertainty case flagged by EARM to be reviewed by humans.

IAM test failures can be attributed to three failure modes: (1) highly idiosyncratic handwriting styles that have not been trained (estimated at 8%); (2) long sequences of 80 characters where DSGRM cross-source agreement deteriorates as a result of accumulated positional uncertainty; and (3) mixed-script content that exceeds the vocabulary of the current system. Future research will address adaptive context windows in hierarchical DSGRM.

Reconstruction quality was further evaluated by analyzing representative qualitative examples in the contexts of visual confusion, lexical ambiguity, contextual inconsistency, and degraded text reconstruction. In Figure 3, the baseline OCR system is compared to the C-GOCR framework.

Input Sample	Ground Truth	TrOCR	VisionLAN	CaGOCR	Type
modern	modern	modem	modem	modern ✓	VC
archive	archive	arcnive	arclive	archive ✓	VC
Ramanujan	Ramanujan	Ramanajan	Ramanajan	Ramanujan ✓	LA
amoxicillin	amoxicillin	amoxicilin	amoxicilin	amoxicillin ✓	CI

Figure 3. Qualitative OCR Comparison

As shown in Figure 3, CaGOCR successfully resolves visually ambiguous character patterns, reconstructs degraded character regions, and improves recognition of out-of-vocabulary terms through lexical memory retrieval. Furthermore, the EARM module provides interpretable attribution labels that assist in identifying the underlying causes of recognition errors.

5.6 Computational Complexity Analysis

A practical assessment of the CAGOCR framework's practical applicability to digitization of handwritten documents was conducted by evaluating the computational overhead introduced by the framework. Even though DCAF, DSGRM, ASRM, EARM, ALMB, and SRCL are more complex than conventional OCR architectures, the performance gains that result justify the additional computational effort.

Because visual and linguistic sequences are generated simultaneously, the dual-source reconstruction strategy introduces moderate inference overhead. A further benefit of the Active Semantic Refinement Module is that iterative refinement is performed only on disagreement positions identified by DSGRM, which reduces unnecessary computation during decoding. In lexically ambiguous predictions, the Adaptive Lexical Memory Bank is activated selectively, contributing minimal additional latency to prediction processing.

According to experimental observations, the proposed framework maintains stable inference performance, while reducing Character Error Rates (CERs), Word Error Rates (WERs), and Reconstruction Quality. It also allows for the

selective deployment of individual components based on application requirements and computational constraints. Despite requiring more computational resources than lightweight OCR architectures, CaGOCR is still an excellent choice for digitizing archival documents, processing legal records, preserving historical manuscripts, and other applications that prioritize recognition accuracy over inference speed. We will examine model compression, lightweight transformers, and efficient lexical retrieval mechanisms in the future to make deployments more efficient.

Table 4 Computational Complexity Comparison of Handwritten Text Recognition Models

Model	Relative Complexity	Inference Cost	Memory Requirement
CRNN+CTC	Low	Low	Low
TrOCR-base	Medium	Medium	Medium
TrOCR-large	High	High	High
SCATTER	High	High	High
VisionLAN	High	High	High
CaGOCR (Proposed)	High+	High+	High+

Based on the relative computational complexity of the OCR architectures evaluated, Table 4 summarizes them. In spite of the additional computational overhead caused by dual-source reconstruction, semantic refinement, and retrieval of lexical memory, the improved recognition accuracy justified the increased resource usage.

5.7 EARM Attribution Dataset Construction

In order to evaluate the Error Attribution and Reasoning Module (EARM), a dedicated attribution dataset based on handwritten text recognition outputs was constructed using controlled perturbation strategies. Three categories of annotated error instances were included in the dataset: Visual Confussion (VC), Lexical Ambiguity (LA), and Contextual Inconsistency (CI).

The Visual Confusion samples were generated by inserting visual similarities between character substitutions, such as rn/m, l/1, O/0, and C/E, common in handwritten documents. We inserted rare proper nouns, domain-specific terminology, historical spellings, and out-of-vocabulary lexical items in order to create lexical ambiguity samples. Grammatical structure and local plausible tokens were maintained when contextually incompatible tokens were substituted with locally plausible alternatives.

We assigned one ground-truth attribution label to each error instance according to the dominant source of recognition uncertainty in the final dataset. We generated 1,080 VC samples, 1,040 LA samples, and 1,080 CI samples to ensure balanced category representation and comparable difficulty across all attribution classes.

As a result of the benchmark, EARM attribution accuracy could be systematically evaluated, as well as an interpretable approach to OCR error analysis that went beyond conventional confidence estimation techniques.

Table 5 Distribution of EARM Attribution Evaluation Samples

Attribution Category	Samples
Visual Confusion (VC)	1080
Lexical Ambiguity (LA)	1040
Contextual Inconsistency (CI)	1080
Total	3200

5.8 Statistical Significance Analysis

By using the Wilcoxon signed-rank test, the observed performance improvements were verified as not being attributed to random variation. Over repeated experimental runs using benchmark datasets such as IAM, CVL, and IIIT-HWS, the CaGOCR framework was compared with the strongest baseline model, TrOCR-large. At the 95% confidence level, CaGOCR improved the Character Error Rate (CER) and the Word Error Rate (WER). All of the evaluated datasets showed consistent performance gains, suggesting that the architectural components are not only improving recognition performance, but also contributing to it reliably.

A relatively small standard deviation was observed over multiple runs, indicating that the model behavior was stable and reproducible. A unified OCR framework can effectively integrate dynamic context-adaptive fusion, dual-

source generative reconstruction, active semantic refinement, lexical memory retrieval, and semantic consistency optimization.

Thus, the statistical analysis reveals that the observed performance improvements are systematic, not random.

Table 6 Statistical Comparison Between CaGOCR and TrOCR-Large

Dataset	CER Improvement (%)	WER Improvement (%)	Statistical Significance
IAM	23.7	18.9	Significant
CVL	21.0	19.8	Significant
IIIT-HWS	20.9	18.1	Significant

5.9 Practical Applications and Deployment Scenarios

For document digitization environments with a high level of recognition accuracy, reconstruction capability, and interpretability, the CAGOCR framework is particularly suited. With CaGOCR, visual evidence, linguistic reasoning, lexical retrieval, and error attribution are integrated into a single framework, whereas conventional OCR systems only focus on character recognition. Under challenging document conditions, this ensures robust performance.

Historic manuscripts often contain degraded text, faded ink, missing character regions, and obsolete lexical forms, which is a key application area. By combining Dual-Source Generative Reconstruction Modules (DSGRM) and Adaptive Lexical Memory Banks (ALMB), semantic consistency is maintained while such information can be recovered.

Legal and administrative document processing can also benefit from the framework, where recognition errors may significantly affect downstream decisions. In order to support human verification and quality control workflows, the Error Attribution and Reasoning Module (EARM) provides interpretable diagnostic information.

During forensic document examination, CAGOCR can provide investigators with attribution-based explanations for reconstruction decisions by identifying uncertain character regions. The ALMB lexical retrieval capabilities may be useful for healthcare, scientific, and government archives containing domain-specific terminology.

The modular design of CaGOCR allows individual components to be adapted or replaced according to application requirements, which facilitates future deployments across multilingual and domain-adaptive systems as well as large-scale document understanding applications.

6. Discussion

According to the experimental results, the proposed CaGOCR framework is capable of improving handwritten text recognition performance over a variety of benchmark datasets. The framework achieves consistent reductions in Character Error Rates (CER) and Word Error Rates (WER) while maintaining strong reconstruction capability under degraded document conditions by reformulating OCR as a generative sequence reconstruction problem rather than a conventional recognition-correction pipeline. Based on these improvements, it appears that integrating visual evidence, linguistic context, and lexical memory within a unified architecture produces a more robust solution than treating them separately. There are two components among the proposed ones that contribute significantly to performance gains: the Dual-Source Generative Reconstruction Module (DSGRM) and the Dynamic Context-Adaptive Fusion (DCAF). The ablation study suggests that replacing generative reconstruction with token-level completion significantly reduces reconstruction quality, highlighting the importance of exploiting visual and linguistic information simultaneously. A similar feature of adaptive fusion is that it allows the model to dynamically balance visual and contextual evidence based on local uncertainty, thus improving recognition robustness in visually ambiguous areas. It provides interpretable error categorization through the Error Attribution and Reasoning Module (EARM). Rather than relying solely on confidence scores, this model identifies what caused the error, such as visual confusion, lexical confusion, or contextual inconsistency. As part of this capability, the Active Semantic Refinement Module (ASRM) supports targeted refinement and provides useful diagnostic information for real-world document digitization workflows requiring human verification. Adaptive Lexical Memory Banks (ALMBs) are particularly effective at storing uncommon lexical patterns and out-of-vocabulary terms. It appears that retrieval-based lexical support reduces the tendency of language models to replace rare terms with high-frequency alternatives, according to experimental results. Document collections that contain uncommon terminology, such as those associated with archival, legal, scientific, and historical collections, are especially suited for this characteristic.

It is important to note, however, that there are several limitations to this technology. By generating dual streams and refining iteratively, the framework increases computational overhead. In addition, performance may decrease for documents with unusual handwriting styles, extremely long sequences, and mixed-script documents that were not extensively represented during training. Researchers will develop lightweight model compression strategies, multilingual adaptations, and more efficient memory retrieval mechanisms to enhance scalability and deployment efficiency in the future.

7. Conclusion

It presents CaGOCR (Context-Aware Generative OCR), a unified framework that reframes handwritten text recognition in terms of generative sequence reconstruction instead of conventional recognition-correction problems. In existing OCR systems, errors are propagated, character regions that are degraded are not recovered, terms that are out of vocabulary are not handled, and reasoning mechanisms are not interpretable. The proposed framework addresses these limitations by integrating visual evidence, linguistic context, lexical retrieval, and semantic consistency optimization into a single architecture capable of recognizing, reconstructing, refining, and assigning errors at the same time. Six complementary components are introduced in the proposed architecture: Dynamic Context-Adaptive Fusion, Dual-Source Generative Reconstruction Module, Active Semantic Refinement Module, Error Attribution and Reasoning Module, Adaptive Lexical Memory Bank (ALMB), and Semantic Reconstruction Consistency Loss (SRCL). Adaptive multimodal fusion, robust sequence reconstruction, targeted refinement of uncertain predictions, interpretable error categorization, effective handling of rare lexical patterns, and semantic coherence across reconstructed sequences are all made possible by these components. In CaGOCR, a number of fundamental limitations of contemporary handwritten text recognition systems are addressed by integrating these mechanisms into a single framework. IAM, CVL, and IIIT-HWS benchmark datasets were evaluated experimentally for consistent improvements over established OCR baselines, such as CRNN+CTC, TrOCR, SCATTER, and VisionLAN. In addition to achieving outstanding performance in Character Error Rate (CER), Word Error Rate (WER), and reconstruction quality metrics, the proposed framework demonstrated strong robustness when document quality was degraded. As a result of comprehensive ablation studies, each architectural component was further confirmed as contributing to the diagnosis. EARM achieved high attribution accuracy, providing interpretable diagnostic information beyond conventional confidence estimation techniques. As a result of these findings, a single OCR framework that combines generative reconstruction, adaptive fusion, semantic refinement, lexical memory retrieval, and attribution-guided reasoning is more effective.

The CaGOCR system offers more than quantitative performance improvements to document digitization workflows in the real world. A number of applications are supported, including archiving, digitizing historical manuscripts, processing legal documents, managing administrative records, examining forensic documents, and collecting domain-specific documents containing rare terms. With the capacity to reconstruct degraded content as well as provide interpretable error explanations, the framework is especially effective for applications that require reliability, transparency, and human verification. CaGOCR combines recognition, correction, lexical reasoning, and interpretability into a context-aware generative reconstruction paradigm that advances handwritten text recognition conceptually as well as empirically. Due to its modular structure, the framework can be extended to support multilingual OCR, mixed-script document comprehension, lightweight deployment strategies, and large-scale document intelligence systems in the future. We expect the proposed approach to serve as an excellent foundation for future research into robust, interpretable, and semantically aware technologies for document understanding

References

1. R. Plamondon and S. N. Srihari, "Online and off-line handwriting recognition: A comprehensive survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 63-84, Jan. 2000, doi: 10.1109/34.824821.
2. Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.
3. B. Shi, X. Bai and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 11, pp. 2298-2304, Nov. 2017, doi: 10.1109/TPAMI.2016.2646371.
4. M. Li et al., "TrOCR: Transformer-based optical character recognition with pre-trained models," arXiv preprint, arXiv:2109.10282, 2021.
5. H. Bunke, "Recognition of cursive writing," in Handbook of Character Recognition and Document Image Analysis, Singapore: World Scientific, 2003, pp. 493-517.
6. A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks," in Advances in Neural Information Processing Systems (NeurIPS), Vancouver, Canada, 2009, pp. 545-552.

7. A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 2017, pp. 5998-6008.
8. H. Bao et al., "BEiT: BERT pre-training of image transformers," arXiv preprint, arXiv:2106.08254, 2021.
9. Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint, arXiv:1907.11692, 2019.
10. R. Litman et al., "SCATTER: Selective context attentional scene text recognizer," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 11962-11972.
11. Y. Wang et al., "From two to one: A new scene text recognizer with visual language modeling network," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, 2021, pp. 14960-14969.
12. A. Radford et al., "Learning transferable visual models from natural language supervision," in Proceedings of the 38th International Conference on Machine Learning (ICML), Virtual Event, 2021, pp. 8748-8763.
13. J. B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 2022, pp. 23716-23736.
14. Y. Xu et al., "LayoutLM: Pre-training of text and layout for document image understanding," in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Virtual Event, 2020, pp. 1192-1200.
15. G. Kim et al., "OCR-free document understanding transformer," in Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 2022, pp. 498-517.
16. K. W. Church and W. A. Gale, "Probability scoring for spelling correction," Statistics and Computing, vol. 1, no. 2, pp. 93-103, 1991, doi: 10.1007/BF01889985.
17. A. Schmaltz et al., "Adapting sequence models for sentence correction," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, Denmark, 2017, pp. 2807-2813.
18. G. M. Correia and A. F. Martins, "A simple and effective approach to automatic post-editing with transfer learning," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, 2019, pp. 3050-3056.
19. J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Minneapolis, MN, USA, 2019, pp. 4171-4186.
20. C. Rigaud and A. Doucet, "OCR ground truth transcription with BERT-based missing character prediction," in Proceedings of the International Workshop on Document Analysis Systems (DAS), Wuhan, China, 2020, pp. 243-258.
21. G. Liu et al., "Image inpainting for irregular holes using partial convolutions," in Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 2018, pp. 85-100.
22. A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 2021.
23. T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature pyramid networks for object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 2117-2125.
24. U. V. Marti and H. Bunke, "The IAM-database: An English sentence database for offline handwriting recognition," International Journal on Document Analysis and Recognition, vol. 5, no. 1, pp. 39-46, 2002, doi: 10.1007/s100320200071.
25. A. Graves, S. Fernández, F. Gomez and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in Proceedings of the 23rd International Conference on Machine Learning (ICML), Pittsburgh, PA, USA, 2006, pp. 369-376.
26. F. Kleber et al., "CVL-DataBase: An off-line database for writer retrieval, writer identification and word spotting," in Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), Washington, DC, USA, 2013, pp. 560-564.
27. P. Krishnan et al., "Word spotting and recognition in Indic documents," in Proceedings of the International Workshop on Document Analysis Systems (DAS), Santorini, Greece, 2016, pp. 48-53.
28. A. Radford et al., "Language models are unsupervised multitask learners," OpenAI Blog, vol. 1, no. 8, pp. 1-9, 2019.