

# *Qalam* : A Multimodal LLM for Arabic Optical Character and Handwriting Recognition

Gagan Bhatia El Moatez Billah Nagoudi  
Fakhraddin Alwajih Muhammad Abdul-Mageed  
The University of British Columbia & Invertible AI  
{gagan30@student., muhammad.mageed@}ubc.ca

## Abstract

Arabic Optical Character Recognition (OCR) and Handwriting Recognition (HWR) pose unique challenges due to the cursive and context-sensitive nature of the Arabic script. This study introduces *Qalam*, a novel foundation model designed for Arabic OCR and HWR, built on a SwinV2 encoder and RoBERTa decoder architecture. Our model significantly outperforms existing methods, achieving a Word Error Rate (WER) of just 0.80% in HWR tasks and 1.18% in OCR tasks. We train *Qalam* on a diverse dataset, including over 4.5 million images from Arabic manuscripts and a synthetic dataset comprising 60k image-text pairs. Notably, *Qalam* demonstrates exceptional handling of Arabic diacritics, a critical feature in Arabic scripts. Furthermore, it shows a remarkable ability to process high-resolution inputs, addressing a common limitation in current OCR systems. These advancements underscore *Qalam*'s potential as a leading solution for Arabic script recognition, offering a significant leap in accuracy and efficiency.

## 1 Introduction

Optical Character Recognition (OCR) technology has revolutionized the way we interact with written and printed materials. It enables the conversion of various documents, including scanned paper documents, PDF files, or images captured by a digital camera, into editable and searchable data. The ability of OCR to digitize text has found applications in numerous domains, ranging from banking and healthcare to education and historical research, among others (Singh et al., 2012).

In this work, our focus is on handling Arabic OCR and HWR. Arabic OCR and HWR pose substantial challenges due to several distinctive features of the Arabic script. The Arabic writing system is cursive and context-dependent, a characteristic that complicates the design of robust OCR models. Further intricacies such as diacritical marks,

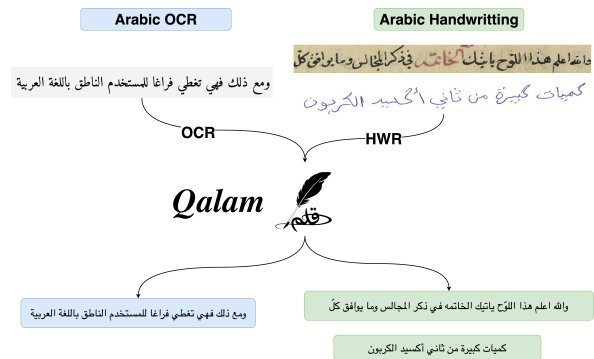


Figure 1: An illustrative overview of *Qalam*'s functioning for Arabic OCR and HWR across diverse text types.

loops, overlaps, ligatures, dots, and multi-context joining contribute to the complexity of the task. In addition to these structural challenges, the vast diversity of Arabic fonts and individual handwriting styles further complicate character recognition and segmentation tasks. Finally, the lack of comprehensive and diverse annotated datasets increases the difficulty of Arabic OCR and HWR tasks. Figure 2 visually represents these challenges.

OCR systems can be classified based on the type of input data they process: handwritten or printed. In this paper, for clarity, we refer to printed text recognition specifically as OCR, although both types broadly fall under the OCR umbrella.

Our objective in this work is to tackle these complex tasks, making several key contributions that collectively underscore the novelty of our research. Concretely, our contributions can be summarized as follows:

- (i) We introduce *Qalam*, a novel OCR and HWR model specifically engineered for Arabic script, establishing a new state-of-the-art performance across diverse datasets.
- (ii) We compile a large and diverse collection of datasets and introduce *Midad* Benchmark for Arabic OCR and HWR. This will serve as a

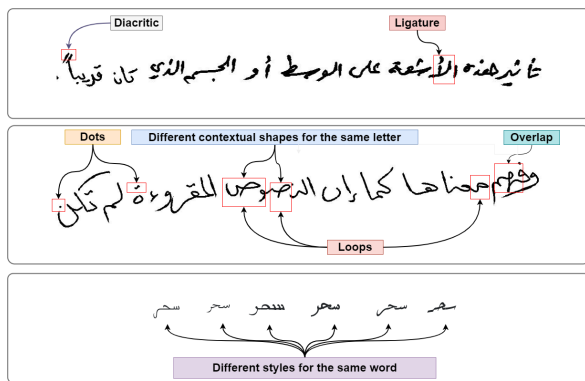


Figure 2: An illustrative depiction highlighting the distinctive characteristics of Arabic script that contribute to its increased complexity compared to other languages.

resource for future research in the community.

- (iii) Our work offers an in-depth analysis of the inherent complexities associated with Arabic script OCR, contributing towards a broader understanding of the challenges in this domain.
- (iv) Last, we provide detailed comparative evaluations against multiple baselines, thus emphasizing the effectiveness of our approach.

The rest of this paper is structured as follows: Section 2 discusses the related work in HWR and OCR. Section 3 introduces *Midad*, our Arabic HWR and OCR benchmark. In Section 4, we present our methodology. Section 5 describes the experiments conducted in this work. Section 6 introduces *Qalam*, the proposed model for Arabic HWR and OCR, and evaluates *Qalam*. We conclude in Section 8 and spell out many limitations related to our work in Section 9.

## 2 Related Works

Traditional Hidden Markov Model (HMM) (Bunke et al., 1995; Park and Lee, 1996; Agazzi and Kuo, 1993) approaches in sequence modeling have been largely surpassed by deep learning techniques that do not require explicit segmentation. Connectionist Temporal Classification (CTC) models (Graves et al., 2006; Graves and Schmidhuber, 2008) and Encoder-Decoder models (Sutskever et al., 2014; Bluche et al., 2017) with attention mechanisms (Bahdanau et al., 2014; Michael et al., 2019) represent two primary deep learning categories. Recent advances include transformer models (Vaswani et al., 2017) and pre-trained models (Devlin et al., 2018).

**Handwriting Recognition (HWR).** HMMs were traditionally used for HWR (Bunke et al., 1995; Park and Lee, 1996), but CTC-based models (Graves et al., 2006; Graves and Schmidhuber, 2008) became more popular due to their accuracy without explicit segmentation. These models often use RNNs and their variations like LSTM (Pham et al., 2014), BLSTM, and MLSTM (Graves and Schmidhuber, 2008; Voigtlaender et al., 2016; Bluche et al., 2017), and combinations with CNNs such as CNN-BLSTM (de Sousa Neto et al., 2020). Recurrence-free models optimized with CTC, using only CNNs, have also been developed (Coquenot et al., 2020).

**Optical Character Recognition (OCR).** OCR, divided into Scene Text Recognition (STR), Scanned Document OCR, and Synthetic Text OCR, has evolved from HMMs to pre-trained transformer models. RNN and CTC-based models (Su and Lu, 2015), combined CNN and BLSTM models (Shi et al., 2016; Breuel, 2017), and Encoder-Decoder architectures with attention mechanisms (Lee and Osindero, 2016; Shi et al., 2018) have been used. Recently, transformer-based models (Li et al., 2021; Kim et al., 2022; Lyu et al., 2022) have gained prominence.

**Multimodal Large Language Models (MLLMs).** MLLMs have impacted tasks like Visual Question Answering (VQA) and OCR-related tasks (Zhang et al., 2024; Wadhawan et al., 2024; Shi et al., 2023). Despite their success, challenges in text recognition within images remain due to lower encoder resolution (Liu et al., 2023). Open-source models like LLaVAR (Zhang et al., 2023) and MonkeyText (Liu et al., 2024) are improving text recognition, while closed-source models like GPT-4V (Achiam et al., 2023) and Gemini Pro (Team et al., 2023) contribute to these advancements.

**Arabic HWR and OCR.** Arabic HWR and OCR initially relied on HMMs (Alma’adeed et al., 2002; Prasad et al., 2008). Later, CTC-based models with RNNs and CNNs replaced HMMs. Ahmad et al. (2017) used an MDLSTM model for the KHATT dataset, and Yousef et al. (2020) introduced a pure CNN model optimized with CTC loss. Recent advancements include transformer models (Mostafa et al., 2021; Momeni and BabaAli, 2023), though Arabic MLLMs still face challenges in text recognition within images (Alwajih et al., 2024). For more detailed surveys, see Alrobah and Albahli (2022);

Faizullah et al. (2023) and reviews by Sobhi et al. (2020); Alghyaline (2023). Figure 9 (Appendix A.1) categorizes various model architectures used in these studies.

### 3 MIDAD Benchmark

#### 3.1 Datasets

We utilize a variety of printed and handwritten Arabic OCR datasets in this study, which we have collectively named *MIDAD*. Below is a summary of these datasets.

**MADBase.** A database of 60k training and 10k testing images of Arabic handwritten digits. It’s often used for training and testing CNNs (El-Sawy et al., 2017a).

**AHCD.** Similar to MADBase, the AHCD dataset is used for training and testing CNNs, but contains 16k samples of handwritten Arabic characters (El-Sawy et al., 2017b).

**ADAB.** Consists of 937 Tunisian town and village names in Arabic handwriting (Boubaker et al., 2021). It contains 15k samples in total.

**Alexuw.** This dataset is compiled by Hussein et al. (2014a), includes 25k samples of 109 different Arabic words. This large and diverse dataset allows for the development and testing of segmented letter-based Arabic handwriting recognition algorithms.

**Shotor.** Introduced by Asadi (2020), this is a large-scale open-source dataset, consisting of 120k grayscale images of various Farsi phrases, each rendered in different fonts and sizes. The phrases were sourced from the Ganjoor and Farsi Wikipedia websites (Asadi, 2020).

**PATS-A01.** Introduced by Al-Muhtaseb et al. (2009), it represents the first printed Arabic text set containing 22k line images sourced from classic Arabic literature. Eight different Arabic fonts were used for these images, adding to the variety in this dataset (Al-Muhtaseb et al., 2009).

**IDPL-PFOD.** This synthetic dataset, created by Hosseini et al. (2021), includes 30k TIF images of printed Farsi text lines. Each image has varying background types and levels of blur and distortion to mimic real-life conditions. The text is rendered using popular Farsi typefaces in diverse sizes and styles.

**UPTI.** The UPTI dataset was developed by Sabour and Shafait (2013) and contains 10k synthetic Urdu text lines in the Nastaliq font, providing a valuable resource for testing and training models in a language closely related to Arabic (Jain et al.,

2017).

**OnlineKHATT.** A comprehensive dataset consisting of 10k Arabic text lines, created by 623 authors using various devices. The data come as stored in the online format using InkML files, and we convert it into offline format as images. The dataset provides character-based segments and manually verified ground truths for the written lines (Mahmoud et al., 2018). *MIDAD* provides a broad and varied base for training and validating OCR models for Arabic and other closely related languages.

#### 3.2 Data Splits

In instances where the original datasets came pre-partitioned into standard training (Train), development (Dev), and testing (Test) splits, we opt to maintain these existing divisions to preserve the integrity of the original data structure. In absence of such pre-defined splits, we randomly shuffle each dataset and split it into three 80% Train, 10% Dev, and 10% Test. Comprehensive statistics reflecting the distribution of data splits across all our datasets are in Table 1.

#### 3.3 In the wild Arabic OCR and HWR Datasets

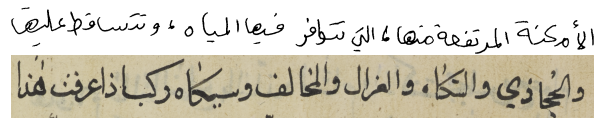


Figure 3: In-the-wild Arabic dataset samples.

**KHATT** The KHATT dataset (Mahmoud et al., 2014), a prominent tool in Arabic handwriting recognition, offers 1,000 handwritten Arabic forms from diverse writers, totalling 4,000 paragraph images—half with similar text and the other half with unique text. Our study concentrated on the unique-text paragraphs providing 6,742 text lines post-segmentation.

**Historical Manuscripts** The dataset contains 120 images from historical Arabic manuscripts, digitized by the British Library and Qatar Foundation (Clausner et al., 2018).

### 4 Methods

In our pursuit to address the challenges of both Arabic HWR and OCR, we employ a Vision Encoder-Decoder (VED) framework that ingeniously brings together transformer-based models

Dataset	Citation	Base	Type	Samples	Language	Dimensions	Sentence Length	Words	Train	Dev	Test
MADBase	El-Sawy et al. (2017a)	Char	HW	70k	Arabic	(28,28)	1	10	48,000	6,000	6,000
AHCD	El-Sawy et al. (2017c)		HW	16k	Arabic	(28,28)	1	10	10,752	1,344	1,344
ADAB	El Abed et al. (2009)		HW	15k	Arabic	(300,80)	1.53	960	12,022	1,503	1,503
Alexuw	Hussein et al. (2014b)	Word	HW	25k	Arabic	(480,232)	1	10,989	20,091	2,512	2,511
Shotor	Asadi (2020)		OCR	120k	Persian	(100,50)	1.08	62,900	96,000	12,000	12,000
PATS01	Al-Muhtaseb et al. (2009)	Line	OCR	22k	Arabic	(1344,80)	16.65	8,248	17,702	2,213	2,213
IDPL-PFOD	sadat Hosseini et al. (2021)		OCR	30k	Persian	(700,50)	15	38,476	24,110	3,014	3,014
OnlineKHATT	Mahmoud et al. (2018)		HW	10k	Arabic	(1089,150)	7.42	22,216	6,798	850	850

Table 1: Summary of Dataset Characteristics: The table presents a comprehensive overview of the datasets used in the study. These datasets collectively offer a diverse range of samples, ensuring the robustness and adaptability of the models evaluated.



Figure 4: A showcase of diverse Arabic script datasets, illustrating the intricate and multifaceted challenges addressed by *Qalam*: (a) Character-based examples, (b) Word-oriented examples, (c) Line-based examples

in a novel manner. This framework leverages the power of transformer-based vision models as encoders, adeptly processing image data, and pairs it with the linguistic sophistication of transformer-based language models as decoders. The result is a synergistic pairing that skillfully transcodes visual information into meaningful textual output, thus overcoming the intricate complexities of Arabic OCR. Furthermore, we extensively analyze various encoder and decoder combinations, investigating their implications for model performance. In VED design, the encoder ingests image data, while the decoder manages ground-truth caption inputs via teacher forcing (Sutskever et al., 2014) during training. We ensure autoregressive training for the next token prediction using causal self-attention in the model. This mechanism restricts a token’s attention

only to its predecessors, maintaining the sequential nature of the input text.

#### 4.1 Encoder Configuration

Our encoder takes images resized to  $(H, W)$ , padded for uniformity, and partitioned into  $N = H * W / P^2$  patches with a fixed size of  $(P, P)$ . These patches are subsequently flattened and linearly projected into  $D$ -dimensional vectors to form patch embeddings. We retain the “[CLS]” token to represent the image and incorporate learnable 1D position embeddings based on absolute positions.

#### 4.2 Decoder Configuration

The decoder in our VED structure consists of layers identical to the encoder, but with an additional *encoder-decoder attention* mechanism. It employs attention masking to prevent the model from peeping into the future during training. Its hidden states are projected linearly to the vocabulary size, and probabilities are computed using softmax. We initialize the cross-attention layer weights randomly when warm-starting from pre-trained transformer-based models.

#### 4.3 Baselines

To provide a comparative analysis, we consider a range of established OCR models as baselines. These models, varying in complexity and technique, include CRNN, Gated-CNN-BiLSTM-CTC, Tesseract, and TrOCR.

**CRNN.** The CRNN model (Puigcerver, 2017) blends convolutional and recurrent layers, and concludes with a linear layer and CTC loss function. It leverages dropout, batch normalization, LeakyReLU, Maxpool, and bidirectional 1D-LSTM layers. Data augmentation through random distortions is used for enhanced robustness.

**Gated-CNN-BiLSTM-CTC.** This architecture (Bluche and Messina, 2017) includes Gated-CNNs for feature extraction. Convolutional layers, gates,

bidirectional LSTM layers, and a linear layer compose the encoder-decoder structure. The model minimizes the CTC objective function using RMSPProp. It also incorporates data augmentation.

**Tesseract.** Introduced by Patel et al. (2012), this is a versatile open-source OCR engine sponsored by Google. It is designed to be a universal text recognition tool, recognizing over 100 languages. Tesseract works by decomposing the input image into lines and words.

**TrOCR.** Introduced by Li et al. (2021), this is a text recognition model that uses pre-trained image and text transformer-based models. Pre-training on large-scale synthetic data, and fine-tuning on human-labeled datasets make TrOCR effective in printed, handwritten, and scene text recognition tasks.

#### 4.4 Evaluation Metrics

Model performance is evaluated based on the Word Error Rate (WER), computed as the normalized Levenshtein distance at the word level. The formula to calculate WER is given in Appendix A.3.

## 5 Experiments

We present our experimental settings, including the selection process for the encoder and decoder from various options. We also discuss the challenges encountered during their selection.

### 5.1 Encoder Selection

The second phase of our experimental design involved selecting the most suitable encoder from available options. We conducted rigorous tests with four different encoders (ViT (Dosovitskiy et al., 2020), DeiT (Touvron et al., 2020), BEiT (Bao et al., 2021), Swin (Liu et al., 2021), and SwinV2 (Liu et al., 2022) Transformers) while keeping the XLM-RoBERTa (Conneau et al., 2019) as the constant decoder. To ensure the uniformity of these experiments, we used the hyperparameters derived from the initial tuning stage.

**Results.** Table 7 displays a performance comparison of several models, including ViT, Swin, BEiT, SwinV2, and DeiT using a constant decoder XLM-R on OCR and HWR tasks. Overall, the DeiT encoder model exhibits superior results and it has a *Midad* score of 19.79, likely due to its adeptness at the discerning text in various forms and orientations in images and recognizing intricate handwritten patterns. However, the SwinV2 encoder model

which has a *Midad* score of 21.60 also shows notable performance, particularly inline-level handwriting recognition on the ‘OnlineKHATT’ dataset, due to its ability to handle variable image inputs Figure 5 explains this visually. The results underline the effectiveness of both DeiT and SwinV2 in handling diverse OCR and HWR tasks across *Midad* datasets.

Cluster	Task	Dataset	E1	E2	E3	E4	E5
HWR	Char	MADBase	7.49	3.39	6.84	3.40	2.74
		AHCD	10.00	8.84	5.95	5.21	3.73
	Word	ADAB	6.92	3.02	8.92	3.05	2.85
		Alexuw	13.92	8.07	7.78	7.92	4.78
Line	OnlineKHATT	75.02	66.39	65.23	64.10	65.03	
OCR	Line	PATS01	52.73	37.74	41.73	33.13	35.24
	Word	Shotor	11.74	8.30	7.73	7.66	7.03
	Line	IDPL-PFOD	46.84	48.43	31.39	30.20	29.42
Overall	Avg.	HWR Score	22.67	17.94	18.94	16.74	15.83
		OCR Score	37.10	31.49	26.95	23.66	23.90
		<i>Midad</i> Score	28.08	23.02	21.95	19.33	18.85

Table 2: Comparative performance analysis using WER (lower is better) of various models across diverse OCR and HWR datasets on the validation set. E1 : ViT, E2 : Swin, E3 : BEiT, E4 : SwinV2, and E5 : DeiT with XLM-R as decoder. The table also provides OCR, HWR, and *Midad* scores to showcase the models’ overall performance in respective tasks.

### 5.2 Decoder Selection

With DeiT established as the optimal encoder, we proceeded to the third phase where we tested the efficacy of different decoders. Maintaining DeiT as a constant encoder, we experimented with five different decoders, assessing their performance on all datasets under the same hyperparameters. We used RoBERTa (Liu et al., 2019), XLM-R (Conneau et al., 2019), ARBERT, MARBERT and MARBERTv2 (Abdul-Mageed et al., 2021).

**Results.** Table 8 contrasts the performance of transformer-based decoders, we have fixed the encoder to DeiT and are experimenting with different decoders on various HWR and OCR tasks. Generally, ARBERT stands out with the lowest WER across most tasks and datasets, showing its strong deciphering ability for diverse texts and handwriting. It acquires a *Midad* score of 12.06 WER for HWR and 18.83 for OCR, with an overall (on both tasks) *Midad* of 17.02 WER. Marbert<sub>v2</sub>, however, excels specifically on the OCR ‘IDPL-PFOD’ which is a Persian line-level task, implying its potential suitability for specific OCR scenarios with dataset which have different languages that use the Arabic Script. These findings highlight ARBERT’s overall dominance, while suggesting the usefulness

of MARBERT<sub>v2</sub> for particular tasks, emphasizing the importance of task-specific model selection.

Cluster	Task	Dataset	D1	D2	D3	D4	D5
HWR	Char	MADBase	13.03	3.74	6.03	2.15	<b>0.52</b>
		AHCD	14.49	4.73	9.6	2.81	<b>1.85</b>
	Word	ADAB	9.51	3.85	6.57	3.03	<b>1.13</b>
		Alexuw	13.92	5.78	9.74	5.43	<b>1.93</b>
	Line	OnlineKHATT	75.26	66.03	67.59	58.73	<b>54.84</b>
OCR	Line	PATS01	54.03	31.71	45.61	27.58	<b>22.92</b>
	Word	Shotor	11.39	8.03	7.04	5.51	<b>3.62</b>
	Line	IDPL-PFOD	64.6	30.42	36.15	<b>27.92</b>	28.45
HWR+OCR		HWR Score	28.30	16.83	19.91	14.43	<b>12.05</b>
		OCR Score	43.34	23.39	29.60	20.34	<b>18.33</b>
	Avg.	<b>Midad Score</b>	34.74	19.29	23.54	16.64	<b>14.41</b>

Table 3: Performance comparison using WER (lower is better) of various transformer-based decoder models on HWR and OCR tasks for different datasets on the validation set. **D1**: RoBERTa, **D2**: XLM-R, **D3**: MARBERT, **D4**: MARBERT<sub>v2</sub>, and **D5**: ARBERT with DeiT as constant encoder. Tasks are categorized into character-level (Char), word-level (Word), and line-level (Line) recognition.

### 5.3 Error Analysis

We carry out a manual error analysis on our validation set, using our top model, to identify challenges with our DeiT and ARBERT models. To this end, we randomly select 200 images from the validation set, divided among two experienced annotators (authors). Annotators reviewed these images, tested the model, and noted its performance. From this small-scale heuristic analysis, we identify two significant challenges our models faced:

**Encoder-Related Issues.** Despite its effectiveness across tasks, DeiT has limitations with scalability and high-resolution inputs. It struggles with larger input sizes, impacting performance in scenarios like detailed document analysis or high-resolution character recognition. As shown in Figure 5, DeiT, after undergoing unsupervised training with Masked Imaging Modeling (MIM) (Xie et al., 2022), may fail to reconstruct high-resolution image content due to inefficient feature processing. Conversely, SwinV2 exhibits superior handling of high-resolution inputs.

**Decoder-Related Issues.** Regarding the decoder, ARBERT has shown limitations in handling diacritics as these are not included in its vocabulary. Diacritics play a crucial role in many languages, including Arabic and Persian, as they can significantly change word meanings. The inability of ARBERT to recognize and process these diacritics may lead to incorrect word recognition and, subsequently, incorrect text interpretation. For instance,

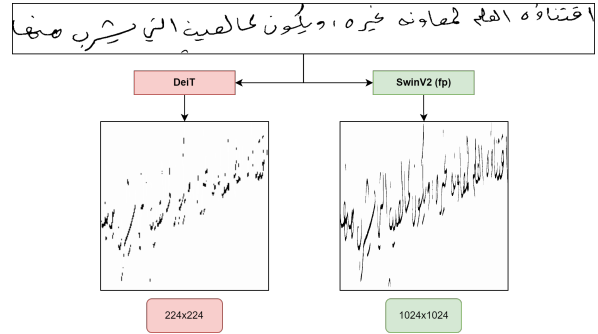


Figure 5: Example of a high-resolution image that may present processing difficulties for DeiT. The images are reconstructions produced by both DeiT and SwinV2 following MIM training.

as shown in Figure 6, ARBERT might interpret the word incorrectly due to its inability to handle diacritics.



Figure 6: Example of a word with diacritics that ARBERT may struggle to interpret correctly.

## 6 Building Qalam

To further optimize the performance of our Vision Encoder-Decoder framework, we introduce additional pre-training strategies for the encoder and the decoder, including the SwinV2 and RoBERTa models. This is undertaken to capitalize on their specific architectural strengths in handling large image sizes and next-token prediction tasks.

**Encoder Upgrades.** To enhance the model’s performance, we further pre-train the Swin Transformer v2 encoder. We choose this model due to its proficiency in handling high-resolution inputs and capturing rich spatial information. The training strategy involved augmenting the input image size and utilizing a robust dataset comprising 4.5M images extracted from Arabic manuscripts and books. The Masked Language Modeling approach is employed during training, promoting the encoder’s adaptability to real-world OCR challenges and the complexities inherent in Arabic script.

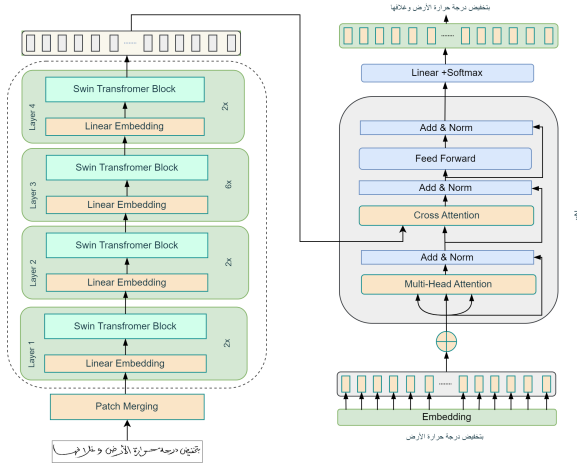


Figure 7: The architecture of *Qalam* model, comprising the SwinV2 encoder for image representation and the Roberta decoder for sequence prediction, indicating the flow of data during the OCR and HWR tasks

**Decoder Upgrades.** On the decoder side, we further pre-train the the RoBERTa model using the Masked Language Modeling approach (Devlin et al., 2018). We select RoBERTa as the decoder for *Qalam* for its superior performance in next-token prediction tasks. A substantial and diverse training dataset provided the model with various language patterns, including Arabic Wikipedia and AraC4 (Abdul-Mageed et al., 2023). The model was also modified to handle longer sequences during training, improving its comprehension and generation of complex sentence structures. We used a sentence-piece tokenizer that also can understand diacritics, which is crucial for processing the Arabic language, thereby enhancing the decoder’s effectiveness in Arabic OCR tasks. Figure 7 shows our final architecture.

**Synthetic Data.** This data comprises 60k image-text pairs and features more than 28 diverse fonts, comprehensively representing potential inputs. *Qalam* is subsequently fine-tuned using this synthetic data in a supervised manner. Figure 10 shows a few examples of this dataset. We generate this dataset using text files from Hindawi Arabic books (Hindawi). We subsequently augment it into different fonts and organize it into PDF files of size 760x640. The model was then fine-tuned on this dataset to teach it about the Arabic script.

**Data Augmentation.** We supplement each training sample from the line-based datasets with four additional synthesized samples, as illustrated in Figure 8. Then, we fine-tune *Qalam* using this syn-

thetic data in a supervised manner. This strategy facilitates the model’s adaptability to various text styles and complexities in Arabic scripts. Here, we take the gold labels of the dataset’s training split and augment it with four different randomly chosen fonts.

**Training Procedure.** The training process for *Qalam* utilizes specific hyperparameters to optimize performance. Table 6 provides a comprehensive overview of these settings. The model is trained using the Adam optimizer with a  $5 \times 10^{-5}$  learning rate. We employ a cosine learning rate scheduler over 50 epochs. The training process uses a batch size of 8, with gradient accumulation steps of eight, resulting in a total effective batch size of 64. For reproducibility, we set the random seed to 42. The evaluation batch size is also set to eight. These carefully chosen hyperparameters are instrumental in achieving optimal performance for the *Qalam* model.

لم يتوصل حتى الآن إلى تسمية الجزيرة قبل

(a) Real sample from the training data.



(b) Augmented samples for the above sample.

Figure 8: Data augmentation for the training data.

## 6.1 Performance Evaluation of *Qalam*

This section encapsulates the performance evaluation of various models, emphasizing our proposed model, *Qalam*, alongside baseline models and alternative architectures. The results are collated in Table 4, highlighting the exemplary performance of *Qalam* across diverse datasets.

On the Handwriting Recognition (HWR) front, *Qalam* exhibits remarkable performance. Specifically, in the MADBase and AHCD datasets, *Qalam* can recognize all test samples without errors. Meanwhile, in word-based datasets such as ADAB



Cluster	Task	Dataset	M1	M2	M3	M4	M5	M6	SOTA	SOTA References	
HWR	Char	MADBase	124.05	02.20	08.00	03.77	00.49	00.43	00.57	de Sousa (2018)	<b>0.004</b>
		AHCD	109.00	23.40	05.43	03.43	01.00	00.90	01.58	de Sousa (2018)	<b>0.003</b>
	Word	ADAB	66.00	02.60	12.58	07.43	00.99	00.99	01.01	Maalej et al. (2016)	<b>0.01</b>
		Alexuw	101.80	05.00	15.74	03.23	01.00	00.83	07.84	Hussein et al. (2014c)	<b>0.01</b>
		OnlineKHATT	62.00	64.32	25.90	15.78	45.00	21.54	12.24	Alwajih et al. (2021)	<b>3.95</b>
OCR	Line	PATS01	63.00	30.23	32.43	26.65	18.00	05.23	n/a	-	<b>1.90</b>
	Word	Shotor	124.00	05.40	08.32	09.01	02.03	02.02	n/a	-	<b>0.12</b>
	Line	IDPL-PFOD	36.00	43.01	29.80	32.44	20.00	04.54	n/a	-	<b>1.53</b>
Overall	Avg.	HWR <sub>Score</sub>	92.57	19.50	13.53	06.73	09.70	04.94	04.65	-	<b>0.80</b>
		OCR <sub>Score</sub>	74.33	26.21	23.52	22.70	13.34	03.93	n/a	-	<b>1.18</b>
		<i>Midad</i> <sub>Score</sub>	85.73	22.02	17.28	12.72	11.06	04.56	04.65	-	<b>0.94</b>

Table 4: Comparative performance analysis of various models across diverse OCR and HWR datasets. The models are: **M1**: Tesseract, **M2**: TrOCR Base, **M3**: CRNN+CTC, **M4**: CNN+BiLSTM+CTC, **M5**: DeIT+ARBERT, **M6**: SwinV2(fp) + ARBERT, **M7**: SOTA, and *Qalam*. The table lists WER (Lower is better) achieved by each model on different Arabic and Persian datasets classified by base, type, and language. The table also provides OCR, HWR, and *MIDAD* scores to showcase the models’ overall performance in respective tasks. *MIDAD* Score is the average WER of across all tasks and datasets

and Alexuw, *Qalam* achieves an equally impressive WER of just 0.01

When assessing line recognition tasks, especially on the OnlineKHATT dataset, *Qalam* continues to excel, recording a WER of 3.95%. This excellence extends to OCR tasks as well, where *Qalam* upholds a low WER on various datasets, including 1.90% on PATS01, 0.12% on Shotor, and 1.53% on IDPL-PFOD. These results also show that our model can perform well on Persian datasets.

In summary, *Qalam* delivers unparalleled performance across the board. The average scores indicate its superiority at a WER of 0.80% for HWR tasks and 1.18% for OCR tasks. Further, it achieves a unique evaluation metric, the *Midad*<sub>Score</sub>, of 0.94%. This collective evidence positions *Qalam* as a leading solution in both the HWR and OCR domains, reflecting its robustness and adaptability to diverse textual challenges.

## 6.2 Performance Evaluation of *Qalam* in the wild Arabic data


Datasets	SOTA Ref	SOTA	
KHATT	(Momeni and BabaAli, 2023)	18.45	<b>10.43</b>
Historical Manuscripts	(Clausner et al., 2018)	<b>21.9</b>	30.88

Table 5: Zero Shot Evaluation of *Qalam* on In the wild Arabic OCR datasets.

Table 5 showcases the comparative performance of the zero-shot evaluation of the *Qalam* system on "in the wild" Arabic OCR datasets in terms of Character Error Rate (CER) as these datasets are more complex and are completely unseen by the

model. The table contrasts the results of the state-of-the-art (SoTA) references with the outcomes achieved using the *Qalam* system. We observe the following:

- **KHATT**: The SoTA result, as referenced by Momeni and BabaAli 2023, yields a CER of 18.45. In comparison, the *Qalam* system significantly improved, achieving a CER of 10.43.
- **Historical Manuscripts**: The performance on this dataset provides a different scenario. The referenced SoTA result from Clausner et al. 2018 reports a CER of 21.9. However, the *Qalam* system shows a higher CER of 30.88 in this context.

In summary, while the *Qalam* model exhibits superior performance on the KHATT dataset, its performance on the Historical Manuscripts dataset is less competitive. The Historical Manuscripts dataset is mainly out-of-domain data, but the performance is still competitive.

## 7 Discussion

The exemplary performance of *Qalam* across Arabic and Persian OCR and HWR tasks (Table 4) highlights its potential. Despite the diversity in the OnlineKHATT dataset, *Qalam* achieves a relatively low residual error of 4%, indicating scope for improved handling of diverse writing styles. The superior performance of *Qalam* over CTC-based models like CRNN+CTC and CNN+BiLSTM+CTC, which emphasizes the transformative potential



of transformer-based models when supplemented with substantial training datasets.

The stark disparity between *Qalam* and TrOCR, despite TrOCR’s strength, underscores the limitation of models pretrained on English data when applied to different scripts, emphasizing the need for script-specific training. The results also highlight the limitations of task-specific models like Tesseract, which excels in line-based recognition but underperforms in character or word-based tasks. Finally, the exceptional performance of *Qalam* can be attributed to the synergy between SwinV2 encoder and RoBERTa decoder, effectively tackling OCR and HWR complexities. Examples of the Demo can be found in the Appendix A.6.

## 8 Conclusion

We introduced *Qalam*, a foundation model for Arabic OCR and HWR. *Qalam* establishes a new standard in Arabic OCR and HWR tasks. The robustness of its architecture, comprising the SwinV2 encoder and RoBERTa decoder, outperforms previous state-of-the-art systems. Our study demonstrates that Arabic script’s unique challenges can be effectively addressed by leveraging the strengths of transformer-based models. The performance of *Qalam* on the *Midad* benchmark validates the scalability and flexibility of our approach, suggesting its potential application to OCR and HWR tasks in other complex scripts. Moving forward, *Qalam* offers a compelling basis for further innovation in Arabic OCR and HWR systems, contributing to the advancement of this critical area of research.

## 9 Limitations

Given the limited availability of HWR and OCR datasets for Arabic, particularly for handwriting with diacritics - a feature often omitted in everyday writing - certain challenges arise. The most notable of these is the prevalence of code-switching and dialectal in real-world writing, both in OCR and HWR contexts, which the current model may struggle to address. Complex tasks such as Scene Text Recognition (STR), multiline, and full-page recognition also show significant limitations to the capabilities of *Qalam*.

Furthermore, *Qalam* has been specifically designed for Arabic OCR and HWR tasks. As a result, its performance has not been assessed on other

scripts or languages. Therefore, its effectiveness in these contexts may not be optimal without further modifications and fine-tuning. This should be considered when attempting to generalize *Qalam*’s capabilities beyond Arabic OCR and HWR tasks.

## Acknowledgments

We acknowledge support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 435-2018-0576; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,<sup>1</sup> and UBC ARC-Sockeye.

## References

- Muhammad Abdul-Mageed, Abdelrahim Elmadany, Alcides Inciarte, Md Tawkat Islam Khondaker, et al. 2023. Jasmine: Arabic gpt models for few-shot learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16721–16744.
- Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. *Arbert & marbert: Deep bidirectional transformers for arabic*.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Oscar E Agazzi and Shyh-shiaw Kuo. 1993. Hidden markov model based optical character recognition in the presence of deterministic transformations. *Pattern recognition*, 26(12):1813–1826.
- Riaz Ahmad, Saeeda Naz, M Zeshan Afzal, S Faisal Rashid, Marcus Liwicki, and Andreas Dengel. 2017. Khatt: A deep learning benchmark on arabic script. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 7, pages 10–14. IEEE.
- Husni Al-Muhtaseb, Sabri Mahmoud, and Rami Qahwaji. 2009. Automatic arabic text image optical character recognition method.
- Salah Alghyaline. 2023. Arabic optical character recognition: A review. *CMES-Computer Modeling in Engineering & Sciences*, 135(3).
- Somaya Alma’adeed, Colin Higgins, and Dave Elliman. 2002. Recognition of off-line handwritten arabic words using hidden markov model approach. In

<sup>1</sup><https://alliancecan.ca>

- 2002 *International Conference on Pattern Recognition*, volume 3, pages 481–484. IEEE.
- Naseem Alrobah and Saleh Albahli. 2022. Arabic handwritten recognition using deep learning: A survey. *Arabian Journal for Science and Engineering*, 47(8):9943–9963.
- Fakhraddin Alwajih, Eman Badr, Sherif Abdou, and Aly Fahmy. 2021. [Deeponkhatt: An end-to-end arabic online handwriting recognition system](#). *International Journal of Pattern Recognition and Artificial Intelligence*, 35.
- Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. 2024. Peacock: A family of arabic multimodal large language models and benchmarks. *arXiv preprint arXiv:2403.01031*.
- Amir Abbas Asadi. 2020. [Shotor dataset](#).
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.
- Killian Barrere, Yann Soullard, Aurélie Lemaitre, and Bertrand Coüasnon. 2022. A light transformer-based architecture for handwritten text recognition. In *International Workshop on Document Analysis Systems*, pages 275–290. Springer.
- Théodore Bluche, Jérôme Louradour, and Ronaldo Messina. 2017. Scan, attend and read: End-to-end handwritten paragraph recognition with mdlstm attention. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 1050–1055. IEEE.
- Théodore Bluche and Ronaldo Messina. 2017. Gated convolutional recurrent neural networks for multilingual handwriting recognition. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 646–651. IEEE.
- Houcine Boubaker, Abdelkarim Elbaati, Najiba Tagougui, Haikal El Abed, Monji Kherallah, Volker Märgner, and Adel M. Alimi. 2021. [Adab database](#).
- Thomas M Breuel. 2017. High performance text recognition using a hybrid convolutional-lstm implementation. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 11–16. IEEE.
- Horst Bunke, Markus Roth, and Ernst Günter Schukat-Talamazzini. 1995. Off-line cursive handwriting recognition using hidden markov models. *Pattern recognition*, 28(9):1399–1413.
- Christian Clausner, Apostolos Antonopoulos, Nora Mcgregor, and Daniel Wilson-Nunn. 2018. Icfhr 2018 competition on recognition of historical arabic scientific manuscripts–rasm2018. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 471–476. IEEE.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Unsupervised cross-lingual representation learning at scale](#). *CoRR*, abs/1911.02116.
- Denis Coquenat, Clément Chatelain, and Thierry Paquet. 2020. Recurrence-free unconstrained handwritten text recognition using gated fully convolutional network. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 19–24. IEEE.
- Denis Coquenat, Clément Chatelain, and Thierry Paquet. 2022. End-to-end handwritten paragraph text recognition using a vertical attention network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):508–524.
- Iam Palatnik de Sousa. 2018. Convolutional ensembles for arabic handwritten character and digit recognition. *PeerJ Computer Science*, 4.
- Arthur Flor de Sousa Neto, Byron Leite Dantas Bezerra, Alejandro Héctor Toselli, and Estanislau Baptista Lima. 2020. Htr-flor: A deep learning system for offline handwritten text recognition. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 54–61. IEEE.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Patrick Doetsch, Albert Zeyer, and Hermann Ney. 2016. Bidirectional decoder networks for attention-based end-to-end offline handwriting recognition. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 361–366. IEEE.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Haikal El Abed, Volker Märgner, Monji Kherallah, and Adel M Alimi. 2009. Icdar 2009 online arabic handwriting recognition competition. In *2009 10th International Conference on Document Analysis and Recognition*, pages 1388–1392. IEEE.
- Ahmed El-Sawy, Hazem EL-Bakry, and Mohamed Loey. 2017a. Cnn for handwritten arabic digits recognition based on lenet-5. pages 566–575.

- Ahmed El-Sawy, Mohamed Loey, and Hazem El-Bakry. 2017b. Arabic handwritten characters recognition using convolutional neural network. *WSEAS Transactions on Computer Research*, 5(1):11–19.
- Ahmed El-Sawy, Mohamed Loey, and Hazem El-Bakry. 2017c. Arabic handwritten characters recognition using convolutional neural network. *WSEAS Transactions on Computer Research*, 5(1):11–19.
- Safiullah Faizullah, Muhammad Sohaib Ayub, Sajid Hussain, and Muhammad Asad Khan. 2023. A survey of ocr in arabic language: Applications, techniques, and challenges. *Applied Sciences*, 13(7):4584.
- Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.
- Alex Graves and Jürgen Schmidhuber. 2008. Offline handwriting recognition with multidimensional recurrent neural networks. *Advances in neural information processing systems*, 21.
- Hindawi. [\[link\]](#).
- Fatemeh sadat Hosseini, Shima Kashef, Elham Shabaninia, and Hossein Nezamabadi-pour. 2021. [Idpl-pfod: An image dataset of printed farsi text for ocr research](#). pages 22–31.
- Mohamed E. Hussein, Marwan Torki, Ahmed Elsallamy, and Mahmoud Fayyaz. 2014a. [Alexu-word: A new dataset for isolated-word closed-vocabulary offline arabic handwriting recognition](#).
- Mohamed E Hussein, Marwan Torki, Ahmed Elsallamy, and Mahmoud Fayyaz. 2014b. Alexu-word: a new dataset for isolated-word closed-vocabulary offline arabic handwriting recognition. *arXiv preprint arXiv:1411.4670*.
- Mohamed E. Hussein, Marwan Torki, Ahmed Elsallamy, and Mahmoud Fayyaz. 2014c. [Alexu-word: A new dataset for isolated-word closed-vocabulary offline arabic handwriting recognition](#).
- Mohit Jain, Minesh Mathew, and C.V. Jawahar. 2017. [Unconstrained ocr for urdu using deep cnn-rnn hybrid networks](#). pages 747–752.
- Lei Kang, Pau Riba, Marçal Rusiñol, Alicia Fornés, and Mauricio Villegas. 2022. Pay attention to what you read: non-recurrent handwritten text-line recognition. *Pattern Recognition*, 129:108766.
- Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 498–517. Springer.
- Chen-Yu Lee and Simon Osindero. 2016. Recursive recurrent nets with attention modeling for ocr in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2231–2239.
- Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. Trocr: Transformer-based optical character recognition with pre-trained models. *arXiv preprint arXiv:2109.10282*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#).
- Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. 2023. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.
- Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*.
- Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. 2022. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Pengyuan Lyu, Chengquan Zhang, Shanshan Liu, Meina Qiao, Yangliu Xu, Liang Wu, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. 2022. Maskocr: Text recognition with masked encoder-decoder pre-training. *arXiv preprint arXiv:2206.00311*.
- Rania Maalej, Najiba Tagougui, and Monji Kherallah. 2016. [Online arabic handwriting recognition with dropout applied in deep recurrent neural networks](#). In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 417–421.
- Sabri Mahmoud, Irfan Ahmad, Wasfi Al-Khatib, Mohammad Alshayeb, Mohammad Parvez, Volker Märgner, and Gernot Fink. 2014. [Khatt: An open arabic offline handwritten text database](#). *Pattern Recognition*, 47:1096–1112.
- Sabri A Mahmoud, Hamzah Luqman, Baligh M Al-Helali, Galal BinMakhashen, and Mohammad Tanvir Parvez. 2018. Online-khatt: an open-vocabulary database for arabic online-text processing. *The Open Cybernetics & Systemics Journal*, 12(1).

- Johannes Michael, Roger Labahn, Tobias Grüning, and Jochen Zöllner. 2019. Evaluating sequence-to-sequence models for handwritten text recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1286–1293. IEEE.
- Saleh Momeni and Bagher BabaAli. 2023. A transformer-based approach for arabic offline handwritten text recognition. *arXiv preprint arXiv:2307.15045*.
- Aly Mostafa, Omar Mohamed, Ali Ashraf, Ahmed El-behery, Salma Jamal, Ghada Khoriba, and Amr S Ghoneim. 2021. Ocformer: A transformer-based model for arabic handwritten text recognition. In *2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pages 182–186. IEEE.
- Hee-Seon Park and Seong-Whan Lee. 1996. Off-line recognition of large-set handwritten characters with multiple hidden markov models. *Pattern Recognition*, 29(2):231–244.
- Chirag Patel, Atul Patel, and Dharmendra Patel. 2012. Optical character recognition by open source ocr tool tesseract: A case study. *International Journal of Computer Applications*, 55(10):50–56.
- Vu Pham, Théodore Bluche, Christopher Kermorvant, and Jérôme Louradour. 2014. Dropout improves recurrent neural networks for handwriting recognition. In *2014 14th international conference on frontiers in handwriting recognition*, pages 285–290. IEEE.
- Rohit Prasad, Shirin Saleem, Matin Kamali, Ralf Meier, and Prem Natarajan. 2008. Improvements in hidden markov model based arabic ocr. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE.
- Joan Puigcerver. 2017. Are multidimensional recurrent layers really necessary for handwritten text recognition? 01:67–72.
- Nazly Sabbour and Faisal Shafait. 2013. A segmentation-free approach to arabic and urdu ocr.
- Fatemeh sadat Hosseini, Shima Kashef, Elham Shabania, and Hossein Nezamabadi-pour. 2021. Idpl-pfod: An image dataset of printed farsi text for ocr research. In *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*, pages 22–31.
- Baoguang Shi, Xiang Bai, and Cong Yao. 2016. 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*, 39(11):2298–2304.
- Baoguang Shi, Mingkun Yang, Xinggong Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2018. Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2035–2048.
- Yongxin Shi, Dezhi Peng, Wenhui Liao, Zening Lin, Xinhong Chen, Chongyu Liu, Yuyi Zhang, and Lianwen Jin. 2023. Exploring ocr capabilities of gpt-4v (ision): A quantitative and in-depth evaluation. *arXiv preprint arXiv:2310.16809*.
- Amarjot Singh, Ketan Bacchuwar, and Akshay Bhasin. 2012. A survey of ocr applications. *International Journal of Machine Learning and Computing*, 2(3):314.
- Mohamed Sobhi, Yasser Hifny, and Saleh Mesbah Elkafas. 2020. Arabic optical character recognition using attention based encoder-decoder architecture. In *2020 2nd International Conference on Artificial Intelligence, Robotics and Control*, pages 1–5.
- Bolan Su and Shijian Lu. 2015. Accurate scene text recognition based on recurrent neural network. In *Computer Vision—ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part I 12*, pages 35–48. Springer.
- Jorge Sueiras, Victoria Ruiz, Angel Sanchez, and Jose F Velez. 2018. Offline continuous handwriting recognition using sequence to sequence neural networks. *Neurocomputing*, 289:119–128.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2020. Training data-efficient image transformers & distillation through attention.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Paul Voigtlaender, Patrick Doetsch, and Hermann Ney. 2016. Handwriting recognition with large multidimensional long short-term memory recurrent neural networks. In *2016 15th international conference on frontiers in handwriting recognition (ICFHR)*, pages 228–233. IEEE.
- Rohan Wadhawan, Hritik Bansal, Kai-Wei Chang, and Nanyun Peng. 2024. Contextual: Evaluating context-sensitive text-rich visual reasoning in large multimodal models. *arXiv preprint arXiv:2401.13311*.

- Christoph Wick, Jochen Zöllner, and Tobias Grüning. 2022. Rescoring sequence-to-sequence models for text line recognition with ctc-prefixes. In *Document Analysis Systems: 15th IAPR International Workshop, DAS 2022, La Rochelle, France, May 22–25, 2022, Proceedings*, pages 260–274. Springer.
- Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663.
- Mohamed Yousef, Khaled F Hussain, and Usama S Mohammed. 2020. Accurate, data-efficient, unconstrained text recognition with convolutional neural networks. *Pattern Recognition*, 108:107482.
- Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

## A Appendices

We provide an addition organized as follows:

- Model architectures in Section A.1.
- Datasets Details Section A.2.
- WER equation Section A.3.
- Hyperparameter Table Section A.4.
- Synthetic Data A.5
- *Qalam* demo Section A.6.
- Test Results A.7

### A.1 Model architectures in the literature

In this section, we provide an illustrative Figure 9 of various model architectures used in the literature.

### A.2 Dataset Details

Table 1 provides additional statistics *Midad*.

### A.3 WER Equation

In this section, we present the equation used for the calculation of WER:

$$WER = \frac{(S + D + I)}{N} = \frac{(S + D + I)}{(S + D + C)} \quad (1)$$

where:

$S$  : number of substitutions,

$D$  : number of deletions,

$I$  : number of insertions,

$C$  : number of correct words,

$N$  : number of words in the reference.

### A.4 Hyperparameter Table

In this section, we present Table 6, detailing the hyperparameters employed in our study.

Hyperparameter	Value
Learning Rate	$5 \times 10^{-5}$
Train Batch Size	8
Eval Batch Size	8
Seed	42
Gradient Accumulation Steps	8
Total Train Batch Size	64
Optimizer	Adam (betas=(0.9,0.999), epsilon= $1 \times 10^{-8}$ )
LR Scheduler Type	Cosine
Num Epochs	50

Table 6: Summary of hyperparameters used for the training process.

### A.5 Synthetic Data

#### A.6 *Qalam* Demo

In addition to the computational experiments, we also developed a practical demonstration that accepts two types of inputs: handwriting and images. The handwriting input facilitates our model’s HWR capabilities, allowing users to test the system’s performance in real-time. Simultaneously, the image input caters to OCR tasks, enabling users to upload images of Arabic scripts and observe the model’s interpretation. One of the noteworthy features of our model is its capacity to handle complex diacritics, a characteristic intrinsic to Arabic scripts. Arabic diacritics are essential in the language, affecting word meanings and pronunciations. However, their tiny size and positioning above or below the line of text make them challenging for many OCR systems. As evidenced by the demonstration, our model exhibits robust performance in recognizing and interpreting these diacritics. The proficiency of our model isn’t limited to diacritics; it extends to handling various types of Arabic texts. Whether it be different fonts, styles, or levels of complexity, our system’s adaptability makes it a potent tool for Arabic script recognition. The demonstration provides a tangible testament to these capabilities, illustrating how the advancements in our model translate into practical, real-world applications. Additionally, Figure 10 displays screenshots of some synthetic samples used in our study.

### A.7 Test Results

Cluster	Task	Dataset	E1	E2	E3	E4	E5
HWR	Char	MADBase	5.99	4.89	5.34	4.90	<b>4.24</b>
		AHCD	8.49	10.34	7.45	6.71	<b>5.23</b>
	Word	ADAB	5.42	4.52	9.42	4.55	<b>4.35</b>
		Alexuw	12.42	9.57	9.28	9.42	<b>6.28</b>
Line	OnlineKHATT	73.52	67.89	66.73	<b>65.60</b>	66.53	
OCR	Line	PATS01	51.23	39.24	43.23	40.76	<b>32.21</b>
	Word	Shotor	10.24	9.80	9.23	9.16	<b>8.53</b>
	Line	IDPL-PFOD	45.34	49.93	32.89	31.70	<b>30.92</b>
Overall	Avg.	HWR Score	21.17	19.44	19.65	21.99	<b>17.33</b>
		OCR Score	35.61	32.99	28.45	27.21	<b>23.89</b>
		<i>Midad</i> Score	26.58	24.52	22.95	21.60	<b>19.79</b>

Table 7: Comparative performance analysis of various models across diverse OCR and HWR datasets. E1 : ViT, E2 : Swin, E3 : BeiT, E4 : SwinV2, and E5 :DeiT with XLM-R as decoder. The table also provides OCR, HWR, and *Midad* scores to showcase the models’ overall performance in respective tasks.

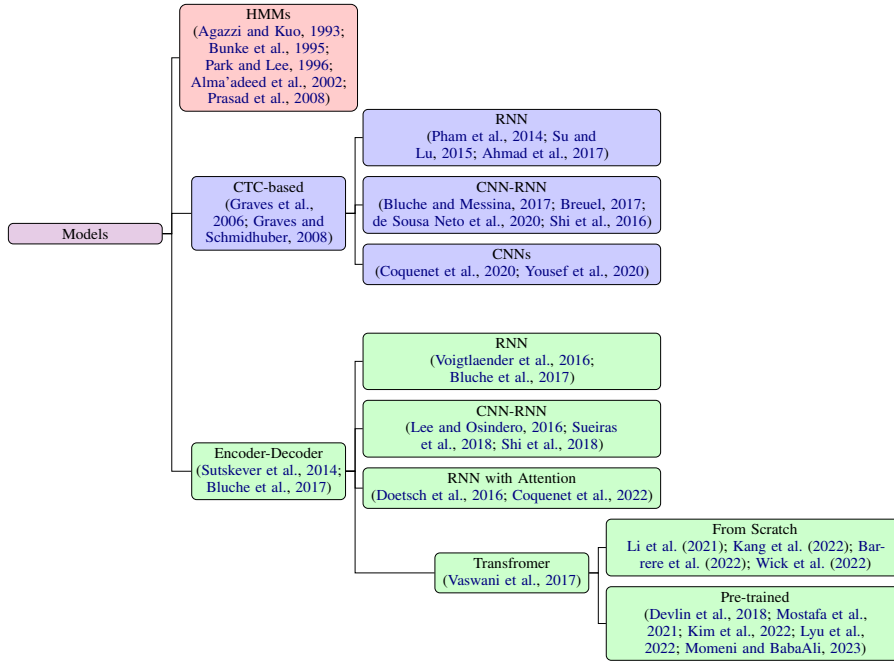


Figure 9: A categorization of diverse model architectures leveraged in the literature, providing an overarching view of the methodological landscape in OCR and HWR.



Figure 10: Qalam Demo samples.

Cluster	Task	Dataset	D1	D2	D3	D4	D5
HWR	Char	MADBase	12.53	4.24	5.53	2.65	<b>0.53</b>
		AHCD	13.99	5.23	9.10	2.31	<b>1.35</b>
	Word	ADAB	9.01	4.35	6.07	3.53	<b>1.63</b>
		Alexuw	13.42	6.28	9.24	5.93	<b>1.43</b>
	Line	OnlineKHATT	74.76	66.53	67.09	59.23	<b>55.34</b>
OCR	Line	PATS01	53.53	32.21	45.11	28.08	<b>23.42</b>
	Word	Shotor	10.89	8.53	6.54	6.01	<b>4.12</b>
	Line	IDPL-PFOD	64.10	30.92	35.65	<b>28.42</b>	28.95
HWR+OCR	Avg.	HWR Score	24.74	17.33	19.41	14.73	<b>12.06</b>
		OCR Score	42.84	23.89	29.10	20.84	<b>18.83</b>
		<b>Mudad Score</b>	31.53	19.79	23.04	17.02	<b>14.60</b>

Table 8: Performance comparison of various transformer-based decoder models on HWR and OCR tasks for different datasets. D1 : RoBERTa, D2 : XLM-R, D3 : MARBERT, D4 : MARBERT<sub>v2</sub>, and D5 : ARBERT with DeiT as constant encoder. Tasks are categorized into character-level (Char), word-level (Word), and line-level (Line) recognition.