Read the Scribbles: A Contextual HTR for Child Handwriting Recognition Without Linguistic Normalization

Sumi Suresh M S[1*][0000−0001−7135−391X], Sahana Rangasrinivasan[1,2][0000−0003−1210−0124], Abbie Olszewski[3][0000−0001−6220−9676], Srirangaraj Setlur[1][0000−0002−7118−9280], and Venu Govindaraju[1][0000−0002−5318−7409]

[1] Department of Computer Science and Engineering, University at Buffalo, New York, USA
[2] Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri, India
[3] College of Education and Human Development, University of Nevada-Reno, Reno 89557, NV, USA
{sumisure, srangasr, setlur, govind}@buffalo.edu, {aolszewski}@unr.edu

Abstract. Handwriting produced by young school-aged children often exhibits poor letter formation, inconsistent spacing, spelling errors, mirrored letters, and other child-specific writing behaviors that differ substantially from clean adult handwriting used to train most handwriting recognition systems. Effective child handwriting recognition, therefore, requires faithful transcription of the actual text written by the child, without linguistic normalization or autocorrection. Such fidelity is critical for assessing educational difficulties and identifying risk factors associated with specific learning difficulties. However, existing handwriting recognition models often impose strong language models or lack robustness to variations in letter formation in children's handwriting. Moreover, many transformer-based approaches are computationally too expensive for practical deployment in classrooms or other resource-constrained settings. We present ScribbleFormer, a lightweight hybrid handwriting recognition model designed for faithful transcription of child handwriting without linguistic normalization. ScribbleFormer combines a convolutional backbone, a bi-directional LSTM, a compact Transformer encoder, and a CTC-based transcription head to capture local stroke-level details and long-range contextual cues while preserving all child-specific writing behaviors. Experiments using IAM for source-domain pretraining and a child handwriting dataset for target-task evaluation show that the proposed design provides strong transcription fidelity and a favorable accuracy-efficiency trade-off for child handwriting recognition. These results make ScribbleFormer well suited for real-world educational applications that require faithful handwriting transcription.

Keywords: Child Handwriting · Contextual OCR · Educational Assessment · Handwritten Text Recognition · Learning Differences.

1 Introduction

Handwriting is a complex developmental skill that emerges through the interaction of motor control, perceptual processing, and growing orthographic knowledge [5]. During the early elementary school years, children's handwritten text often exhibits substantial visual and linguistic variability, including poor letter formation, inconsistent spacing, nonstandard capitalization, phonetic spellings, mirrored letters, or other child-specific writing behaviors. These irregularities appear across a wide range of developmental contexts, including the early stages of literacy acquisition, and can be indicators of specific learning disabilities (SLDs) such as dyslexia and dysgraphia [3]. Although such characteristics are meaningful indicators of a child's writing process, they pose significant challenges for automatic handwriting transcription systems that are typically designed and evaluated on clean and well-formed adult handwriting.

A key requirement in child handwriting recognition is faithful transcription of the text exactly as written by the child, including spelling errors, mirrored letters, morphological deviations, and other non-standard writing behaviors that may be informative for educational analysis. However, existing handwriting recognition systems often struggle in this setting: models trained on adult handwriting do not generalize well to children's irregular letter forms, while models with strong language-modeling components tend to normalize or autocorrect unconventional sequences, thereby losing important characteristics of the original writing (Fig. 1).

These limitations reflect deeper architectural mismatches between existing handwritten text recognition (HTR) systems and the properties of child handwriting. Furthermore, many state-of-the-art transformer models require substantial computational resources during both training and inference. In practical settings such as classrooms, clinics, and other resource-constrained educational environments, deploying such heavy models that are computationally expensive in terms of memory and time is challenging and underscores the need for recognition systems that balance robustness, accuracy, and efficiency.

To address these challenges, we introduce ScribbleFormer, a lightweight contextual handwriting recognition model that is computationally efficient in memory and time. ScirbbleFormer is designed to capture the irregular and highly variable structure of children's handwriting. It integrates a convolutional backbone, a bidirectional LSTM, and a compact Transformer encoder to capture fine-grained stroke-level cues and long-range contextual dependencies. By retaining a CTC-based transcription head rather than a sequence decoder with strong linguistic constraints, the model preserves the actual written text of the child, including spelling errors and mirrored letters (Fig. 1). This hybrid architecture enables ScribbleFormer to achieve a favorable balance between robustness, fidelity, and computational efficiency compared to other baseline models.

We evaluate ScribbleFormer using IAM for source-domain pretraining and a private child handwriting dataset collected from Grades K–5 for target-task evaluation. Across these experiments, ScribbleFormer shows strong transcription fidelity together with computational efficiency, making it suitable for child hand-

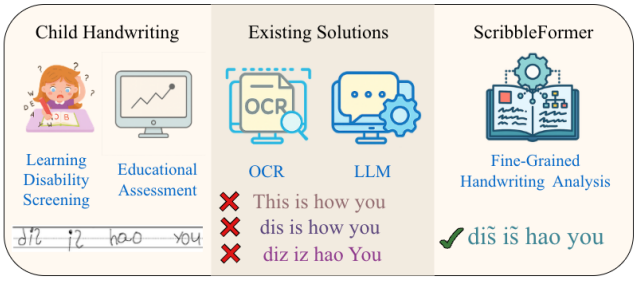


Fig. 1: Overview of the limitations of existing conventional OCR and LLM-based systems on child handwriting and the motivation for ScribbleFormer.

writing recognition in real-world educational settings. The key contributions of this work are as follows:

1. A problem formulation and experimental study focused on faithful, behavior-preserving transcription of child handwriting without linguistic normalization.
2. A practical architectural design to balance local visual modeling, contextual reasoning, transcription fidelity, and computational efficiency for child handwriting.
3. A comprehensive evaluation on child handwriting that analyzes recognition fidelity and efficiency under the constraints of educational use.

2 Related Works

Handwritten text recognition (HTR) has been explored through a variety of architectural paradigms, ranging from sequence-based models to transformer-driven and hybrid designs. We organize prior work into three areas most relevant to this study: (1) sequence-based HTR models, (2) transformer and hybrid OCR architectures, and (3) studies focusing on children's handwriting.

Sequence-Based Handwriting Recognition: CNN–RNN pipelines remain one of the most widely adopted approaches to offline HTR. Architectures that combine convolutional feature extractors with bidirectional LSTMs and CTC decoding [7, 15] have been shown to achieve strong performance on adult handwriting benchmarks. Extensions incorporating multidimensional LSTMs or attention mechanisms [9, 13] further enhance the sequence modeling component. These architectures remain popular due to their alignment-free training and effectiveness on well-structured adult handwriting. Recent studies also explore deeper recurrent stacks and improved feature normalization strategies, reflecting continued interest in strengthening temporal modeling within CTC-based frameworks [4]. However, most sequence-based HTR systems are developed and evaluated on well-structured adult handwriting, where character boundaries and

letter shapes are comparatively stable. To the best of our knowledge, there are no existing HTR systems that have used children's handwriting during training.

Transformer and Hybrid OCR Architectures: Transformer-based models have recently advanced the state of the art in scene text recognition (STR) and HTR by leveraging global self-attention and richer contextual representations [10, 14, 6]. Encoder–decoder architectures enable flexible modeling of long-range dependencies, while hybrid CNN–Transformer designs [11, 2, 18] aim to balance local spatial encoding with contextual reasoning. Despite these advances, many transformer-based OCR models remain computationally heavy, leaving a gap to be addressed for lightweight contextual architectures suitable for practical handwriting recognition. Again, these models lack training on children's handwriting.

Child Handwriting and Learning Disabilities: Handwriting has long been studied as an observable marker of cognitive and motor development, particularly in research on learning disabilities such as dyslexia and dysgraphia [3]. While most computational research focuses on classifying handwriting to detect learning disabilities, relying solely on this metric can lead to misdiagnosis or inaccurate labeling. Instead of serving as a conclusive diagnosis, handwriting should be viewed as important evidence that supports and is considered alongside other information when identifying a potential disability [17]. Handwriting transcription models that can faithfully preserve non-standard spelling and character forms remain largely underexplored. Recent studies have begun investigating OCR-centered approaches for children's handwriting [16, 17]. However, existing recognition systems often tend to enforce linguistic correctness over transcription fidelity due to the intrinsic use of language models in many HTR systems or use data for training that does not include characteristics of children's writing, limiting their ability to preserve spelling errors, mirrored characters, and other child-specific writing behaviors that are critical for educational analysis.

3 Problem Discussion

Child handwriting presents a unique set of challenges because children are still acquiring the necessary motor, perceptual, and linguistic competencies for consistent written text production. As a result, characters produced during this developmental phase can be poorly formed, with character boundaries often ambiguous due to irregular spacing, inconsistent stroke formation, and large intra-class variation in letter shapes. Conventional HTR models, trained primarily on stable adult handwriting, assume a predictable correspondence between discrete labels and visual shapes, which makes them struggle with aligning these unstable patterns seen in children's handwriting. Furthermore, modern OCR systems and foundation models integrate strong language-modeling components to produce fluent output, but this strength becomes a liability when transcribing children's handwriting. These models frequently infer missing words, adjust capitalization, or normalize phonetic spellings, resulting in transcriptions that deviate sharply from the actual written input by correcting non-standard language behaviors.

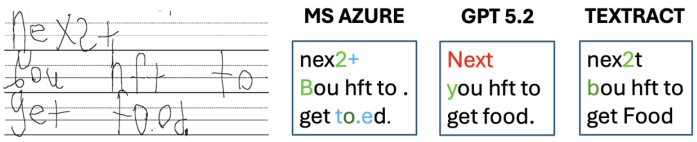


Fig. 2: Comparison of conventional OCR and LLM-based HTR models on a child handwriting sample. Colored highlights indicate normalization and error behaviors: green denotes mirrored-letter correction, blue indicates misrecognition due to poor letter formation, and red denotes spelling-error correction.

These limitations are evident in the qualitative behavior of existing OCR systems on child handwriting. Fig. 2 illustrates a representative failure case in which conventional OCR and LLM-based models systematically normalize or correct child handwriting. In particular, models with strong language models exhibit large variance, suggesting linguistically "correct" alternatives to spelling errors and mirrored letters. Beyond accuracy, many transformer-based architectures further exacerbate the problem by imposing high computational costs, making them impractical for deployment in educational or low-resource settings. Together, these observations highlight the need for a task-driven architectural choice rather than a claim of architectural novelty. The recognition pipeline should tolerate visual ambiguity, minimize reliance on linguistic normalization, and maintain computational efficiency—requirements that directly motivate the design of ScribbleFormer.

4 Methodology

We present ScribbleFormer, a practical hybrid recognition architecture for faithful transcription of child handwriting. As illustrated in Fig. 3, the model combines CNN feature extraction, BiLSTM sequence modeling, Transformer-based contextual reasoning, and CTC decoding to address visual variability, reduce linguistic normalization, and maintain computational efficiency.

4.1 Convolutional Backbone and Sequence Embedding

Given an input grayscale handwritten line image $I \in \mathbb{R}^{H \times W}$, the model first processes it through a lightweight VGG-style convolutional backbone designed to extract local stroke patterns while reducing spatial resolution, with spatial downsampling by a factor of 4 along both dimensions. The backbone consists of four convolutional layers with two max-pooling operations, producing a feature map F.

$$F \in \mathbb{R}^{256 \times H' \times W'}, \qquad H' = H/4,\ W' = W/4,$$

where H' and W' are the downsampled height and width, respectively, and the 256 channels encode local visual features. To convert this 2D feature representation into a 1D sequence suitable for recurrent and Transformer processing,

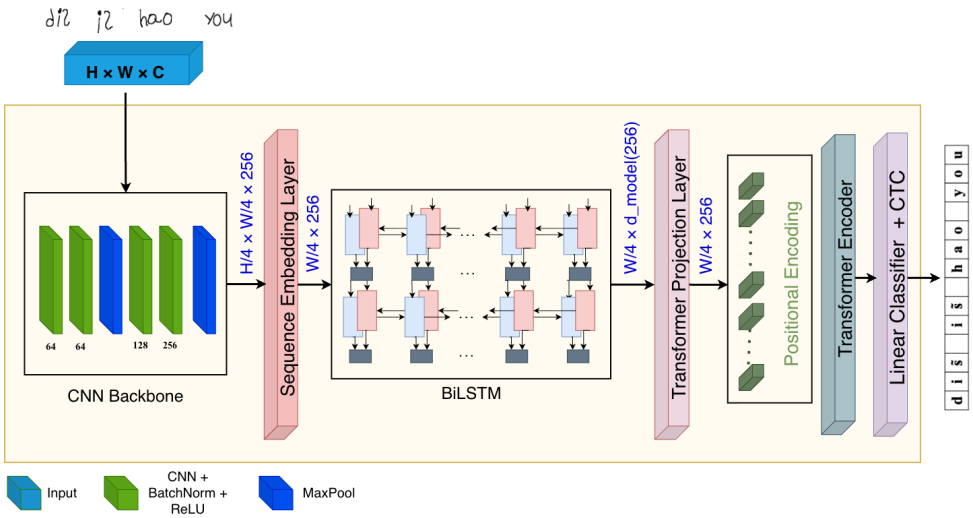


Fig. 3: Architectural diagram of ScribbleFormer

the horizontal axis W' is interpreted as the temporal dimension. For each column t, the channel and height dimensions of F are flattened into a single vector:

$$X_t = \text{Flatten}(F[:, :, t]) \in \mathbb{R}^{256 \cdot H'}, \qquad t = 1, \ldots, W'.$$

Each vector X_t encodes local stroke and shape information at position t.

A linear projection then maps these vectors into a fixed-dimensional embedding space aligned with the downstream BiLSTM and Transformer encoders:

$$E_t = W_{\text{emb}} X_t + b_{\text{emb}} \in \mathbb{R}^{256}.$$

where $\mathbf{W}_{\text{emb}} \in \mathbb{R}^{256 \times (256H')}$ and $\mathbf{b}_{\text{emb}} \in \mathbb{R}^{256}$ denote the learnable embedding weight matrix and bias vector, respectively. This produces a sequence of embeddings $\{E_t\}_{t=1}^{W'}$ that preserves the left-to-right structure of the handwriting and serves as the input to the BiLSTM sequence encoder.

4.2 BiLSTM Sequence Encoder

The embedded feature sequence $\{E_t\}_{t=1}^{W'}$ is then passed through a bidirectional LSTM (BiLSTM) to model local and mid-range temporal dependencies along the handwriting line. Formally, at each time step t, the BiLSTM produces a hidden representation

$$H_t = \text{BiLSTM}(E_t), \qquad t = 1, \ldots, W'.$$

where $H_t \in \mathbb{R}^d$ denotes the concatenated forward and backward hidden states. This recurrent encoding stabilizes the sequential representation by integrating short-range contextual information before global context modeling.

To ensure compatibility with the subsequent Transformer encoder, the BiLSTM outputs are linearly projected into a fixed-dimensional feature space:

$$Z_t = \mathbf{W}_{\text{proj}} H_t + \mathbf{b}_{\text{proj}} \in \mathbb{R}^{256}.$$

yielding the adapted sequence $\{Z_t\}_{t=1}^{W'}$ used for contextual self-attention.

4.3 Transformer Encoder (TE) and Contextual Modeling

While the BiLSTM captures local and mid-range dependencies along the handwriting sequence, it remains limited in modeling long-range contextual relationships across an entire text line and in disambiguating visually similar letters based on the surrounding characters. To address this, ScribbleFormer employs a multi-layer Transformer encoder (TE) that enables global information exchange between all temporal positions.

Before entering the Transformer, we add positional encodings to the BiLSTM outputs $\{Z_t\}_{t=1}^{W'}$, to preserve sequential order information:

$$\hat{Z}_t = Z_t + \text{PE}_t,$$

where PE_t denotes a fixed sinusoidal positional encoding. The resulting sequence is then processed by a stack of Transformer encoder layers:

$$U_t = \text{TE}(\hat{Z}_t), \qquad t = 1, \ldots, W'.$$

The output representations $\{U_t\}$ integrate both local stroke information and long-range contextual structure, forming a robust basis for subsequent character prediction.

4.4 Character Vocabulary and CTC Transcription

The context-aware representations $\{U_t\}$ produced by the Transformer encoder provide a temporally aligned sequence of character-level features. These features are converted into an actual-text transcription using (i) an extended character vocabulary and (ii) a CTC-based transcription strategy.

Extended Character Vocabulary: Child handwriting often exhibits non-canonical character forms, including left–right and top–bottom mirrored letters and their combinations, which are commonly normalized or misrecognized by conventional OCR vocabularies. To preserve these writing behaviors, ScribbleFormer employs an extended character-level vocabulary that augments standard alphanumeric and punctuation symbols with dedicated tokens representing mirrored counterparts, following the encoding strategy introduced in E-TrOCR [16]. Each mirrored form is treated as a distinct symbol rather than mapped to its visually similar canonical counterpart. For example, a base character such as 'a' is represented as ($\tilde{a}$) for the left–right mirrored form, ($\hat{a}$) for the top–bottom mirrored form, and ($\ddot{a}$) for the combined mirrored form. This explicit representation enables the model to distinguish visually similar but semantically distinct

forms and ensures transcription fidelity to the original spelling, capitalization, and letter-formation patterns in the child's handwriting.

CTC Transcription: Transformer outputs $\{U_t\}$ are mapped to frame-wise character logits over the character vocabulary through a linear classifier $\{O_t\}$

$$O_t = w_c U_t + b_c \in \mathbb{R}^C,$$

where C denotes the size of the extended character vocabulary and w_c, b_c are learnable parameters. The logits are converted to log-probabilities and optimized using Connectionist Temporal Classification (CTC) loss:

$$\mathcal{L}_{\text{CTC}} = \text{CTC}\big(\log P_{1:T}, \mathbf{y}\big),$$

where $\log P_{1:T}$ represents the frame-wise log-probabilities and $\mathbf{y}$ is the ground-truth character sequence. CTC enables alignment-free training between variable-length input sequences and target transcriptions while preserving temporal order. During training, CTC loss serves as the sole optimization objective, while recognition accuracy is evaluated separately using CER and WER metrics.

At inference time, the actual-text prediction is obtained using greedy CTC decoding, which merges consecutive identical predictions corresponding to the same character instance and removes blank labels, while preserving genuine repeated characters present in the handwriting. Because decoding operates directly over the extended character set, the output retains child-specific spelling irregularities and mirrored letter forms without imposing linguistic normalization.

5 Experimental Setup

We conducted a comprehensive set of experiments to assess the effectiveness, robustness, and generalization ability of the proposed ScribbleFormer model for actual child handwriting recognition. Our evaluation examines (i) the model's ability to generalize from adult handwriting to child handwriting, (ii) its ability to preserve exact spellings and writing patterns produced by children, and (iii) its performance relative to classical, sequence-based, and transformer-based OCR architectures. All experiments were conducted on a high-performance computing server equipped with a 1× NVIDIA A100 GPU (80 GB), running CUDA 12.2. The model was implemented in PyTorch.

5.1 Datasets

IAM Handwriting Dataset (Source-Domain Pretraining Only): The IAM dataset serves as the pretraining corpus for all models. It contains over 13,000 line-level English handwriting samples written by 657 adult writers, collected under controlled conditions. The dataset primarily consists of clean, well-formed adult handwriting with consistent spacing and minimal spelling variations. We use the standard train/validation/test split.

Child Handwriting Dataset (Target Task and Main Evaluation): Our private corpus[1] consists of 500 handwriting samples collected from elementary school students in Grades K–5 using an e-ink tablet [17]. These samples were segmented into 2,437 line-level instances, which were split in a writer-independent manner into 70% training, 10% validation, and 20% testing sets. In contrast to IAM, this dataset exhibits substantial visual and linguistic variability: incomplete strokes, mirrored letters, inconsistent spacing, phonetic spellings, nonstandard capitalization, and developmental errors typical of early writers. All samples are manually transcribed, preserving the exact text written by the child (verified with the child during data collection), without correction or normalization, ensuring that the evaluation measures how accurately models can reproduce authentic child writing. The target task of this paper is child handwriting recognition; IAM is used only to initialize models before adaptation and is not the primary basis for the paper's claims.

5.2 Baseline Models

We compare ScribbleFormer against representative baselines from three model families: classical recurrent HTR models, sequence-based scene text recognizers, and transformer-based OCR architectures. The classical baselines include RNN+CTC and CRNN [8, 19]; the scene-text models include ASTER, MASTER, and SRN [20, 12, 21]; and the transformer-based OCR models include ViTSTR, TrOCR, and E-TrOCR [1, 10, 16]. All baselines were trained using official implementations under identical preprocessing and training settings. These comparisons are intended to contextualize ScribbleFormer across representative model families and to examine how different design choices behave under child-handwriting conditions.

5.3 Training Strategy

We adopt a two-stage optimization pipeline. All baselines without publicly released pretrained checkpoints (RNN+CTC, CRNN, ASTER, MASTER, SRN, ViTSTR, E-TrOCR, and the proposed ScribbleFormer) are first trained from scratch on the IAM dataset and then fine-tuned on the child handwriting dataset. TrOCR is initialized from the publicly released `trocr-base` checkpoint, which is already pre-trained on large-scale printed and handwritten corpora; therefore, we only fine-tune it on the child handwriting data and do not perform additional pretraining on IAM.

Stage 1: Pretraining on IAM. ScribbleFormer and all trainable baselines are optimized for 100 epochs with a batch size of 32 using the AdamW optimizer with a learning rate of 3×10^{-4}, weight decay of 1×10^{-4}, and gradient clipping at 1.0. CTC loss is used for all CTC-based architectures, and the checkpoint

[1] The data in this dataset was collected under a protocol approved by the Institutional Review Board (IRB) at the University of Nevada, Reno. We are exploring options to release a de-identified version of the corpus publicly.

with the lowest validation Character Error Rate (CER) on IAM is retained as the initialization for the second stage.

Stage 2: Fine-tuning on Child Handwriting. The best IAM checkpoint is then loaded, and the full network is unfrozen for end-to-end fine-tuning on the child handwriting dataset. Fine-tuning is performed for 100 epochs with a batch size of 32 using the AdamW optimizer, but with a reduced learning rate of 1×10^{-4} while keeping the weight decay and gradient clipping unchanged. The model with the lowest validation CER on the child validation set is used for final testing in the experiments.

5.4 Evaluation Metrics

We evaluate ScribbleFormer along two dimensions: transcription quality and computational efficiency. Transcription quality is measured using Character Error Rate (CER) and Word Error Rate (WER), where lower values indicate better recognition performance. For efficiency, we report parameter count, epoch time, training throughput, peak GPU memory during training & inference, and inference latency.

6 Results and Discussions

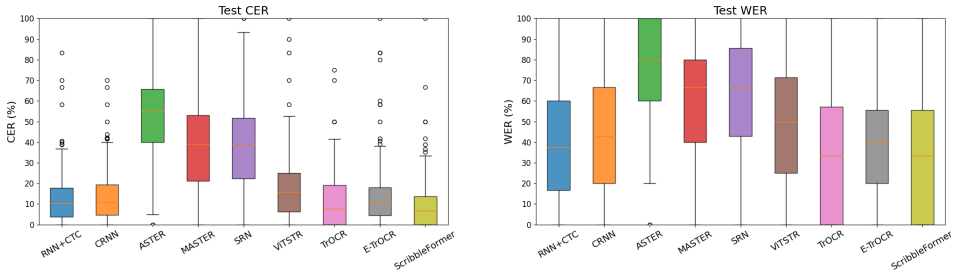
6.1 Quantitative Results

Pretraining Performance on IAM: Table 1 reports the validation and test performance of all models on the IAM dataset. Models originally developed for scene text recognition exhibit notably lower accuracy, suggesting that their spatial rectification and word-level attention mechanisms do not transfer effectively to line-level handwriting. Classical baselines (RNN+CTC, CRNN) offer more stable transcription performance, with CRNN outperforming other non-transformer models. Transformer-based architectures further reduce both CER and WER, benefiting from their stronger sequence modeling capacity. Since we initialize TrOCR from the publicly released `trocr-base` checkpoint, which is already pretrained on large-scale handwriting corpora, it does not undergo IAM pretraining in our pipeline. The extended E-TrOCR model, which is specifically adapted for child handwriting, performs competitively on the IAM dataset. However, ScribbleFormer provides competitive source-domain pretraining performance, indicating that the architecture can learn stable line-level handwriting representations before adaptation to child handwriting.

Fine-tuning Performance on Child Handwriting: Table 1 summarizes model performance after domain-specific fine-tuning on the child handwriting dataset. While RNN+CTC and CRNN adapt reasonably well to this domain shift, scene-text models continue to lag behind, indicating that the distortions present in child handwriting differ fundamentally from geometric irregularities commonly addressed in scene-text recognition. Transformer-based models exhibit stronger robustness, and E-TrOCR in particular benefits from its extended

Table 1: Recognition results for pretraining on IAM and target-task evaluation on the child handwriting dataset.

Model	**IAM Dataset**				**Child Handwriting Dataset**			
	Validation		Test		Validation		Test	
	CER ↓	WER ↓	CER ↓	WER ↓	CER ↓	WER ↓	CER ↓	WER ↓
Classical recurrent NN baselines								
RNN+CTC	20.35	50.54	12.27	39.52	13.56	41.24	13.51	39.31
CRNN	12.74	40.51	9.27	33.09	13.10	41.31	13.21	40.95
Sequence-based scene text recognizers								
ASTER	73.59	97.80	61.84	81.40	62.18	85.25	56.50	80.23
MASTER	57.98	83.74	44.84	67.25	45.86	71.00	40.68	68.39
SRN	57.44	81.38	45.25	66.21	46.68	71.54	41.54	66.03
Transformer-driven OCR models								
ViTSTR	29.16	66.14	21.96	57.45	19.30	51.35	18.37	50.33
TrOCR	–	–	–	–	17.34	41.20	14.88	36.95
E-TrOCR	18.27	46.41	10.97	35.62	14.06	40.97	13.90	40.12
ScribbleFormer	**10.17**	**34.22**	**7.60**	**28.57**	**10.92**	**35.89**	**10.87**	**35.25**



(a) (b)

Fig. 4: Box-plot comparison of recognition errors on the child handwriting test set across all baseline models. (a) Distribution of CER, (b) Distribution of WER

token representations and prior specialization toward child handwriting, achieving comparatively low CER and WER among the baselines. ScribbleFormer provides the strongest overall balance among the compared models, combining lower recognition error with behavior-preserving transcription and moderate computational cost. These results support the use of the proposed hybrid design as a practical architecture choice for child handwriting recognition. Fig. 4 illustrates the distribution of recognition errors across all child handwriting test samples, highlighting how different models vary in robustness under diverse writing conditions.

Computational Efficiency: Table 2 presents a detailed comparison of computational efficiency in terms of memory and time on the child handwriting dataset. Large encoder-decoder transformer architectures, such as TrOCR and E-TrOCR, incur substantial overhead in terms of parameter count, training memory, and inference latency. This makes them computationally expensive

Table 2: Computational efficiency comparison across models on the child handwriting dataset.

Model	Params (M)	Epoch time (s)	Throughput (img/s)	Peak GPU Mem (GB)		Latency (ms/line)
				Train	Infer	
Classical recurrent NN baselines						
RNN+CTC	1.222	2.230	874.160	1.270	0.596	3.437
CRNN	2.267	3.886	501.567	4.925	1.394	3.768
Sequence-based scene text recognizers						
ASTER	15.074	2.642	737.599	0.566	0.336	3.393
SRN	15.666	2.687	725.270	0.596	0.341	3.406
MASTER	17.245	2.716	717.642	0.650	0.349	13.385
Transformer-driven OCR models						
ViTSTR	4.767	0.912	2137.608	0.451	0.140	2.944
TrOCR	333.920	55.576	35.069	17.078	5.780	39.550
E-TrOCR	87.160	42.448	45.914	11.910	2.388	11.376
ScribbleFormer	4.470	4.695	415.091	4.970	1.440	3.680

and potentially impractical for real-time or resource-constrained deployments. In contrast, lightweight models based on recurrent neural networks (RNN) or smaller Vision Transformers (ViT) achieve excellent throughput and low memory consumption. However, they consistently fail to achieve competitive recognition accuracy on challenging child handwriting.

The ScribbleFormer architecture preserves a favorable balance between accuracy and efficiency. It maintains a compact parameter size with moderate memory usage during both training and inference, while achieving inference latency comparable to classical recurrent baseline models. At the same time, ScribbleFormer delivers higher recognition accuracy than lightweight baselines while maintaining substantially lower computational overhead than large-scale transformer architectures. The importance of these efficiency results is not only lower computational cost, but also the feasibility of deploying behavior-preserving child-handwriting recognition in educational or assistive settings where model size, memory use, and latency matter.

Ablation Study: To assess the contribution of the sequential modeling components in ScribbleFormer, we compare four architectural variants: CNN+CTC, CNN+BiLSTM+CTC, CNN+Transformer Encoder+CTC, and the full ScribbleFormer (CNN+BiLSTM+Transformer Encoder+CTC). As shown in Table 3, CNN+CTC alone performs poorly, indicating that local visual features are insufficient for robust handwriting transcription. Adding either BiLSTM or Transformer based contextual modeling leads to substantial improvements. The full model achieves the lowest CER and WER on both IAM and the child handwriting dataset, showing that combining local sequential modeling with global contextual reasoning yields the most effective design.

Table 3: Component-Wise Ablation of ScribbleFormer on IAM and Child Handwriting

Architecture	**IAM Dataset**				**Child Handwriting Dataset**			
	Validation		Test		Validation		Test	
	CER ↓	WER ↓	CER ↓	WER ↓	CER ↓	WER ↓	CER ↓	WER ↓
CNN+CTC	33.76	84.4	28.98	80.57	28.98	81.17	27.45	71.77
CNN+BiLSTM+CTC	13.34	41.68	9.40	33.60	13.14	41.36	13.00	41.30
CNN+TE+CTC	14.88	49.42	10.96	39.54	11.35	38.73	11.84	38.81
ScribbleFormer	**10.17**	**34.22**	**7.60**	**28.57**	**10.92**	**35.89**	**10.87**	**35.25**

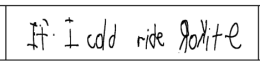
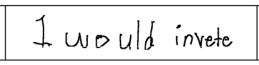
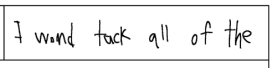
Model			
Classical recurrent NN baselines			
RNN+CTC	If I cold ride RoKite	I would invete	I wnd tack all of the
CRNN	If I cold ride RoKite	I wsuld invete	I wnd tack all of the
Sequence-based scene text recognizers			
ASTER	If I cold ride RoKite	I umed mint	I would talk of a the
MASTER	If the cold the car is	I surea wil innd	the car of the car of
SRN	If I cold ride RoKite	I was wit in the	I would tet all of the
Transformer-driven OCR models			
ViTSTR	If I cold ride RoKite	I wo uld imete	I wand tact all of the
TrOCR	If I add ride Roxite	I would investe	I wond tack gll of the
E-TrOCR	If I cold ride ŘoǨite	I ua uld invete	waond tack all of the
ScribbleFormer	If I cold ride ŘoǨite	I wo uld invete	I wond tack all of the

Fig. 5: Qualitative comparison on three child handwriting samples. ScribbleFormer best preserves the actual written text.

6.2 Qualitative Results

Fig. 5 presents predictions from all models on three representative child handwriting samples. The first example contains homophonic variants of spellings (e.g., "*RoKite*") and inconsistent stroke formation; many baselines attempt lexical correction and drift away from the written form, whereas E-TrOCR and ScribbleFormer most faithfully reproduce the written text by the child. The second and third examples exhibit poor letter formation and ambiguous character boundaries. Scene-text and transformer-based models drift semantically or produce partially incoherent outputs. In contrast, ScribbleFormer consistently maintains the written structure and character-level fidelity across all samples, producing the most stable transcription among the evaluated models.

Overall, the qualitative results reinforce the quantitative findings: while most baselines struggle with faithful transcription of the child's handwriting, ScribbleFormer and E-TrOCR better replicate the actual text written by children. Most importantly, ScribbleFormer achieves this fidelity with substantially lower computational cost than E-TrOCR, making it suitable for real-time applications.

7 Conclusion

Handwriting produced by young school-aged children exhibits a wide range of child-specific writing behaviors that pose significant challenges for handwriting recognition models. Recognizing such handwriting requires models that can handle visual and linguistic irregularities to faithfully preserve critical characteristics in a child's writing. This work studies child handwriting recognition as a behavior-preserving transcription problem rather than a standard language-normalized HTR task. Within this setting, ScribbleFormer serves as a practical hybrid architecture that combines a CNN backbone, a bidirectional LSTM encoder, a compact Transformer, and CTC decoding to preserve child-specific writing behaviors while remaining computationally efficient. This design enables effective modeling of both local stroke-level details and long-range visual context without the influence of strong linguistic normalization. Experimental evaluations using IAM for source-domain pretraining and the child handwriting dataset for target-task evaluation show that ScribbleFormer provides strong transcription fidelity and competitive recognition performance compared to recurrent, scene-text, and transformer-based baselines. In addition, ScribbleFormer exhibits a favorable balance between accuracy and computational efficiency, with substantially lower model complexity and inference cost than large transformer-based approaches. These results indicate that ScribbleFormer is well-suited for real-world educational settings, where computational resources are limited and preserving child-specific writing behaviors is essential. Future work will focus on improving robustness and extending the framework toward broader educational analytics applications.

Acknowledgments This material is based upon work supported by the National AI Research Institutes program of the U.S. National Science Foundation and the Institute of Education Sciences, U.S. Department of Education, under Award #2229873 (National AI Institute for Exceptional Education). This work was also supported in part by a Google Foundation Research Grant. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, the U.S. Department of Education, or Google.

References

1. Atienza, R.: Vision transformer for fast and efficient scene text recognition. In: International conference on document analysis and recognition. pp. 319–334. Springer (2021)
2. Baena, R., Kalleli, S., Aubry, M.: General detection-based text line recognition. Advances in Neural Information Processing Systems **37**, 42388–42404 (2025)
3. Baggett, M., Diamond, L.L., Olszewski, A.: Dysgraphia and dyslexia indicators: Analyzing children's writing. Intervention in School and Clinic **59**, 319–330 (2024)
4. Chaudhary, K., Bali, R.: Easter2. 0: Improving convolutional models for handwritten text recognition. arXiv preprint arXiv:2205.14879 (2022)

5. Chung, P.J., Patel, D.R., Nizami, I.: Disorder of written expression and dysgraphia: definition, diagnosis, and management. Translational pediatrics **9**(1), S46 (2020)
6. Fujitake, M.: Dtrocr: Decoder-only transformer for optical character recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 8025–8035 (2024)
7. Garrido-Munoz, C., Rios-Vila, A., Calvo-Zaragoza, J.: Handwritten text recognition: A survey. arXiv preprint arXiv:2502.08417 (2025)
8. Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., Schmidhuber, J.: A novel connectionist system for unconstrained handwriting recognition. IEEE transactions on pattern analysis and machine intelligence **31**(5), 855–868 (2008)
9. Kang, L., Toledo, J.I., Riba, P., Villegas, M., Fornés, A., Rusinol, M.: Convolve, attend and spell: An attention-based sequence-to-sequence model for handwritten word recognition. In: German Conference on Pattern Recognition. pp. 459–472. Springer (2018)
10. Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F.: Trocr: Transformer-based optical character recognition with pre-trained models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 13094–13102 (2023)
11. Li, Y., Chen, D., Tang, T., Shen, X.: Htr-vt: Handwritten text recognition with vision transformer. Pattern Recognition **158**, 110967 (2025)
12. Lu, N., Yu, W., Qi, X., Chen, Y., Gong, P., Xiao, R., Bai, X.: Master: Multi-aspect non-local network for scene text recognition. Pattern Recognition **117**, 107980 (2021)
13. Michael, J., Labahn, R., Grüning, T., Zöllner, J.: Evaluating sequence-to-sequence models for handwritten text recognition. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1286–1293. IEEE (2019)
14. Parres, D., Anitei, D., Paredes, R.: Handwritten document recognition using pre-trained vision transformers. In: International Conference on Document Analysis and Recognition. pp. 173–190. Springer (2024)
15. Puigcerver, J.: Are multidimensional recurrent layers really necessary for handwritten text recognition? In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR). vol. 1, pp. 67–72. IEEE (2017)
16. Rangasrinivasan, S., MS, S.S., Setlur, S., Jayaraman, B., Govindaraju, V.: From scribbles to text: A novel transformer-based recognition model for child handwriting. In: International Conference on Document Analysis and Recognition. pp. 115–131. Springer (2025)
17. Rangasrinivasan, S., Suresh, S., Olszewski, A., Setlur, S., Jayaraman, B., Govindaraju, V.: Ai-enhanced child handwriting analysis: A framework for the early screening of dyslexia and dysgraphia. SN Computer Science **6**(5), 1–26 (2025)
18. Retsinas, G., Nikolaidou, K., Sfikas, G.: Enhancing crnn htr architectures with transformer blocks. In: International Conference on Document Analysis and Recognition. pp. 425–440. Springer (2024)
19. Shi, B., Bai, X., Yao, C.: 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 (2016)
20. Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., Bai, X.: Aster: An attentional scene text recognizer with flexible rectification. IEEE transactions on pattern analysis and machine intelligence **41**(9), 2035–2048 (2018)
21. Yu, D., Li, X., Zhang, C., Han, J., Liu, J., Ding, E.: Towards accurate scene text recognition with semantic reasoning networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 12110–12119 (2020)