UniMERNet: A Universal Network for Real-World Mathematical Expression Recognition

Bin Wang*, Zhuangcheng Gu*,
Chao Xu, Bo Zhang, Botian Shi, Conghui He†

Shanghai AI Laboratory

Abstract. This paper presents the UniMER dataset to provide the first study on Mathematical Expression Recognition (MER) towards complex real-world scenarios. The UniMER dataset consists of a large-scale training set UniMER-1M offering an unprecedented scale and diversity with one million training instances and a meticulously designed test set UniMER-Test that reflects a diverse range of formula distributions prevalent in real-world scenarios. Therefore, the UniMER dataset enables the training of a robust and high-accuracy MER model and comprehensive evaluation of model performance. Moreover, we introduce the Universal Mathematical Expression Recognition Network (UniMERNet), an innovative framework designed to enhance MER in practical scenarios. UniMERNet incorporates a Length-Aware Module to process formulas of varied lengths efficiently, thereby enabling the model to handle complex mathematical expressions with greater accuracy. In addition, UniMERNet employs our UniMER-1M data and image augmentation techniques to improve the model's robustness under different noise conditions. Our extensive experiments demonstrate that UniMERNet outperforms existing MER models, setting a new benchmark in various scenarios and ensuring superior recognition quality in real-world applications. The dataset and model are available at https://github.com/opendatalab/UniMERNet.

Keywords: Mathematical Expression Recognition · Length-Aware Module · Data Augmentation

1 Introduction

Mathematical Expression Recognition (MER), a key task in document analysis, aims to convert image-based mathematical expressions into corresponding markup languages such as LaTeX or Markdown. MER is important in applications such as scientific document extraction, where a robust MER model helps maintain document logical coherence. Unlike typical Optical Character Recognition (OCR) tasks, MER requires a more sophisticated understanding of complex structures, including superscripts, subscripts, and various special symbols.

∗ Equal contribution.
† Corresponding author: heconghui@pjlab.org.cn

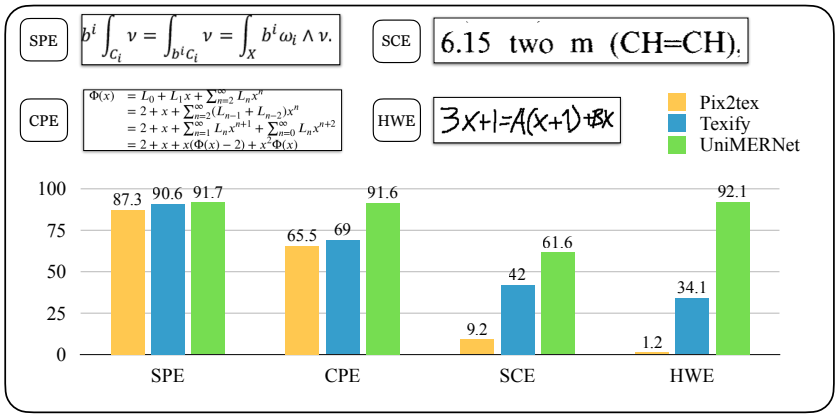

Fig. 1: Performance comparison (BLEU Score) of mainstream models and UniMERNet in recognizing real-world mathematical expressions: Evaluation across Simple Printed Expressions (SPE), Complex Printed Expressions (CPE), Screen-Captured Expressions (SCE), and Handwritten Expressions (HWE).

Existing research has primarily focused on enhancing recognition accuracy for relatively simple rendered expressions [6] and handwritten data [10,18,28,34] through a series of MER algorithms. Some researchers have begun to optimize MER algorithms by scaling up the training data and integrating them with transformer [26] models, ensuring their applicability in diverse scenarios [3, 4, 8, 23]. Other researchers have attempted to directly employ Large Vision-Language Models (LVLMs) for document content extraction, including MER [4, 27]. However, existing MER benchmarks [6, 18] primarily focus on simple printed expressions or handwritten image recognition. Therefore, these models tend to struggle with diverse real-world expressions, such as lengthy equations and noisy scanned document screenshots.

In practice, real-world scenarios require the handling of complex, long expressions and noisy, distorted images from scanned documents or webpage screenshots. To fill this gap, we introduce a comprehensive benchmark, UniMER-Test, which extends the existing test set with longer and real-world scenario expressions. Our benchmark stimulates the progress of MER in robustness and practical usage. As depicted in Fig. 1, we conducted an exhaustive evaluation of state-of-the-art MER methods [3, 23] using our novel benchmark, UniMER-Test. These methods demonstrate remarkable competence in recognizing simple printed expressions. However, their performance noticeably declines when tested with more complex printed expressions, particularly long formulas. The performance degradation becomes even more pronounced when these methods are applied to real-world expressions, such as screen-captured expressions embedded in noisy backgrounds and handwritten expressions. Moreover, large vision-language models such as Nougat [4] and Vary [27], despite their capacity for convenient end-to-end document content extraction, exhibit only mediocre performance in MER.

To address these challenges, we first introduce the UniMER dataset, an extensive collection tailored for Mathematical Expression Recognition (MER), featuring 1 million meticulously curated expressions. Designed to both complement and validate the advancements in MER, this dataset includes the comprehensive UniMER-1M training samples and the thorough UniMER-Test sets, aiming to stimulate further research by offering a broader diversity of expressions compared to existing datasets. Together with the dataset, we propose UniMERNet, a model attuned to the sequence lengths of expressions with its additional length-aware module (LAM). UniMERNet leverages context information regarding the length of an expression prior to prediction, allowing it to generate an expression with sequence length matching the visual feature of the input images. This model is further enhanced through image augmentation techniques, significantly boosting its performance in real-world applications. The main contributions of this paper are as follows:

- We introduce **UniMER**, a universal MER dataset, with the training set UniMER-1M and the test set UniMER-Test, which encompasses all types of expressions in practical situations, offering a diverse and comprehensive foundation for MER model development and evaluation.
- We present **UniMERNet**, a novel MER framework based on the encoder-decoder structure, with an innovative sequence Length-Aware Module to improve the model's ability to handle a wider range of expressions than existing models
- Validation of UniMERNet's superior performance through extensive experiments, establishing it as the new benchmark in open-source MER solutions by outperforming existing models in a variety of scenarios.

2 Related Work

2.1 Machine Learning-Based Methods

Several decades ago, researchers began recognizing the uniqueness and significance of Mathematical Expression Recognition (MER), leading to the initiation of corresponding studies. Anderson [1] pioneers an approach to MER in irregular documents, introducing a parsing algorithm for two-dimensional character configurations. Miller & Viola [19] propose an efficient system that integrates specialized character segmentation with the intrinsic grammar of the mathematical layout language. Chan *et al.* [5] develop an online MER system, incorporating an error detection and correction mechanism to handle lexical, syntactic, and semantic errors. INFTY [25] is an integrated OCR system for mathematical documents that achieves high accuracy in character recognition through a series of novel techniques. Despite the innovative works during this stage, the precision of MER remains limited due to the limitations of handcrafted features in traditional machine learning.

2.2 CNN-Based Deep Learning Methods

With the advent and development of deep learning, a series of MER algorithms based on Convolutional Neural Networks (CNN) [9, 24] are proposed. Deng *et al.* [6] introduce an encoder-decoder model for OCR, leveraging a scalable coarse-to-fine attention mechanism. They present a new dataset, IM2LATEX-100K, demonstrating the superiority of their approach over traditional OCR systems in handling non-standard tasks. The WAP [32] model learns mathematical expression grammar and handles symbol segmentation autonomously, with its learned alignments closely mirroring human intuition. The PAL-v2 [28] model employs paired adversarial learning to recognize handwritten mathematical expressions, effectively manage style variations, and show solid performance on the CROHME dataset. Zhang *et al.* [31] propose a tree-structured decoder for image-to-markup tasks, exhibiting superior performance over traditional string decoders in handling complex tree-structured markups, while maintaining flexibility for various sequence-to-sequence problems. Leveraging bi-directional learning, Zhao *et al.* [34] and Bian *et al.* [2] effectively enhance the recognition performance of their encoder-decoder models for MER, thereby promoting advancements in the field of Handwritten Mathematical Expression Recognition (HMER). The CAN [13] enhances HMER by integrating a weakly-supervised counting module into the encoder-decoder model, optimizing HMER and symbol counting concurrently for improved performance. By employing data augmentation strategies, including distortion, decomposition, and scale augmentation techniques, Le *et al.* [10] and Li *et al.* [14] substantially enhance the performance of MER, effectively addressing the issue of data scarcity. Deep learning-based algorithms significantly improve precision compared to traditional machine learning methods. However, current MER focuses more on handwritten formulas, overlooking those in documents and screenshots.

2.3 Transformer-Based Deep Learning Methods

In recent years, with the rapid development of Transformer [26] and large vision-language models [7, 15, 16, 35], researchers have gradually realized that a wealth of knowledge information exists in document data, which is crucial for further enhancing the performance of large models. Therefore, some researchers have begun to study how to extract document information. Donut [8] proposes an end-to-end document information extraction model that can directly convert input document images into structured outputs without relying on OCR. Nougat [4] employs more rule-based auto-generated image-to-markup samples to train an end-to-end transformer-based encoder-decoder model. This model is specifically designed to transcode document pages into markup language. Vary [27] is a fine-grained perception-capable multimodal large model that can be used for document content parsing. However, these methods have not considered the special nature of mathematical expressions compared to general text, and their MER capabilities are relatively weak. Therefore, Pix2tex [3] and Texify [23], based on a large number of rendered mathematical expressions, train encoder-decoder

Table 1: Statistical comparison of the MER dataset. "Max Length" and "Avg Length" mean the maximum length and average string length of the mathematical expression.

Dataset	Type	Train Size	Test Size	Max Len	Avg Len
HME100K	HWE	74,502	24,607	311	24.05
CROHME		8,836	3233	147	22.27
IM2LATEX-100K	SPE	83,883	10,354	440	96.01
Pix2tex		158,480	30,637	2949	93.35
UniMER-1M	Mixed	1,061,791	23,757	7037	79.48

models, which work well on rendered mathematical expressions in documents. However, these methods often fail for complex mathematical expressions, and their recognition results are significantly reduced for noisy mathematical expressions under screenshots. The MER model proposed in this paper takes into account various situations, aiming to build a robust and practical MER model.

3 UniMER Dataset

The UniMER dataset is tailored to tackle the challenge of formula recognition diversity in real-world scenarios. The dataset is divided into two subsets: the UniMER-1M training set and the UniMER-Test testing set. The UniMER-1M training set covers a wide array of mathematical expressions encountered in real-world situations. Meanwhile, UniMER-Test facilitates comprehensive, accurate, multi-dimensional evaluation.

Specifically, UniMER-1M comprises 1,061,791 Latex-Image sample pairs, encompassing both short and complex, long formula expressions. A crucial aspect of its construction was the careful balance of different length distributions. This equilibrium is instrumental in enabling models trained on UniMER-1M to improve their overall recognition accuracy and generalization capabilities significantly. UniMER-Test is constructed exclusively for the comprehensive evaluation of MER in real-world scenarios. It offers a detailed evaluation of MER from four dimensions, encompassing 6762 Simple Printed Expressions (SPE), 5,921 Complex Printed Expressions (CPE), 4,774 Screen Capture Expressions (SCE), and 6,332 Handwritten Expressions (HWE). This diverse range of expressions, totaling 23,789 samples, ensures a thorough evaluation of MER's performance.

As shown in Tab. 1, the UniMER dataset presents two notable enhancements over existing datasets. Firstly, the UniMER-1M training set is distinguished by its inclusion of formulas of varied lengths and complex expressions, which are scarcely represented in current datasets, encompassing a total of 1 million instances - a significant increase from the previously largest open-source dataset's 158,480 formulas. Secondly, the UniMER-Test evaluation set encompasses samples from various dimensions, offering a more realistic and comprehensive assessment of model performance in practical applications.

The UniMER-1M dataset includes the following types of formulas:

- SPE. These are formula images rendered from simple LaTeX expressions. They are characterized by uniform font size, clean background, and relatively short formulas, as shown in the first row of Fig. 2.
- CPE. These are formula images rendered from complex, long LaTeX expressions. They are characterized by uniform font size, clean background, and complex, longer formulas, as shown in the second row of Fig. 2.
- SCE. These are screen-captured images of formulas from documents and the web. They are characterized by inconsistent fonts and sizes, background noise, and image deformation, as seen in the third row of Fig. 2.
- HWE. These are collected from referenced handwriting recognition datasets [18, 20, 21, 29]. They are complex and diverse, with varying backgrounds, but are relatively short, as shown in the fourth row of Fig. 2.

3.1 Data Collection Process

Printed Rendered Expressions (SPE, CPE) The assembly of our dataset began with the public dataset provided by Pix2tex [3], which served as the basis for our SPE. Recognizing the limitations of the Pix2tex dataset in terms of volume and complexity, we embarked on a data augmentation process. For the SPE, we expanded the base data by sourcing additional LaTeX expression source codes from platforms such as Arxiv [1], Wikipedia [2], and Physics/Math StackExchange [3]. These codes underwent a regularization process [6] to resolve any LaTeX syntax ambiguities before being compiled into expression PDF files in various fonts using XeLaTeX [4]. Uncompilable expressions were discarded. Subsequently, ImageMagic's conversion function [5] was utilized to transform these images into expressions with multiple DPIs, with data balancing ensuring an even distribution of different lengths.

Following this data expansion pipeline, we sampled 725,246 simple formulas from the augmented data and combined them with the Pix2tex training set to form the SPE training data. The Pix2tex test set is designated as the SPE test data. In contrast, the CPE is derived independently of the Pix2tex dataset. We randomly selected 110,332 complex formulas from the expanded data to create the UniMER-1M-CPE and UniMER-Test-CPE.

Screen-Captured Expressions (SCE) For SCE, we compiled a diverse collection of 1,000 PDF pages, encompassing a spectrum of content types in both Chinese and English, such as books, papers, textbooks, magazines, and newspapers. This diverse collection ensured a broad range of fonts, sizes, and backgrounds for the formulas. We employed two annotators to identify and label the formula boxes in the documents, and automatically capture the content within.

[1] https://arxiv.org/
[2] https://www.wikipedia.org/
[3] https://stackexchange.com
[4] https://www.ctan.org/pkg/xetex
[5] https://imagemagick.org/

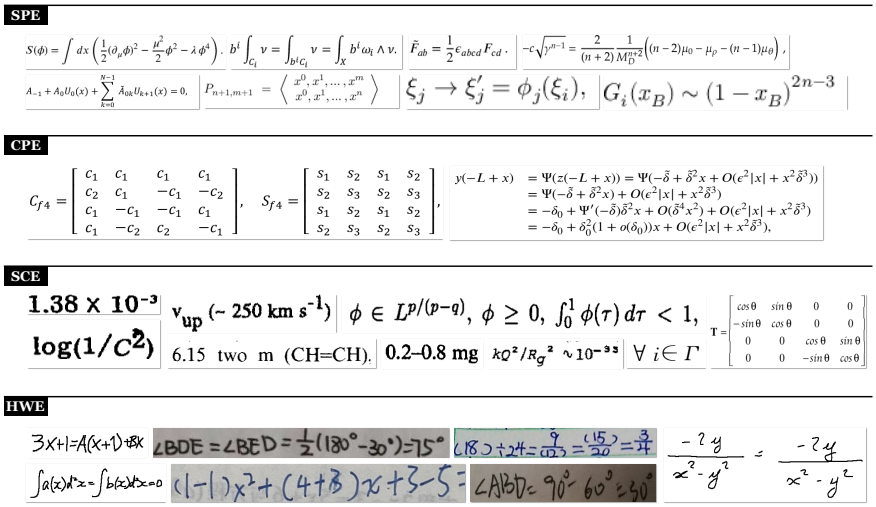

Fig. 2: UniMER-1M: A comprehensive dataset for MER with Simple Printed Expressions (SPE), Complex Printed Expressions (CPE), Screen Capture Expressions (SCE), and Handwritten Expressions (HWE).

This resulted in over 6,000 formula boxes, which were then processed through Mathpix [6] for formula recognition. Manual corrections were applied based on recognition results, and the LaTeX annotations were cross-verified by two annotators. Redundant formulas were identified and removed, yielding 4,744 unique SCEs to be used as the SCE test set.

Handwritten Expressions (HWE) For HWE, we leveraged the existing public datasets CROHME [18, 20, 21] and HME100K [29]. CROHME, a widely recognized dataset in the HMER field, originated from the handwritten digit recognition competition and includes 8,836 training expressions and 3,332 test expressions. HME100K, a real-world handwritten expression dataset, comprises 74,502 training and 24,607 test images. Given the high accuracy of these datasets' annotations, we combined them to form our HWE data. Specifically, the HWE training set incorporated 8,836 training formulas from CROHME and 74,502 training formulas from HME100K, totaling 83,338 training samples. The HWE test set comprised 3,332 test set formulas from CROHME and a 3,000 formula samples from the HME100K test set, resulting in a total of 6,332 test formulas.

3.2 Diversified Training Data Sampling

Existing formula datasets, such as HWE (CHROME & HME100K) [18, 20, 21], IM2LATEX [6], and Pix2tex [3] predominantly comprise rendered and handwritten formulas. However, these datasets exhibit limitations in terms of formula

[6] https://mathpix.com/

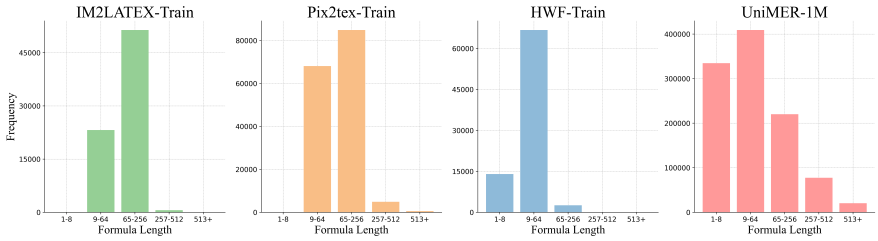

Fig. 3: Formula string length distribution across different datasets.

length and complexity. For instance, Pix2tex mainly contains regular formulas and lacks extremely short or complex long formulas. On the other hand, handwritten formulas are typically short while diverse in handwriting styles, with none exceeding 256 characters. To overcome these limitations, we have enriched our UniMER-1M dataset with a comprehensive range of additional formulas. These formulas have been carefully sampled from diverse sources such as Arxiv and Wikipedia, ensuring a balanced distribution of formula lengths of varying complexity. The length-aware sampling strategy equips the model trained on UniMER-1M to recognize formulas across a broad complexity spectrum, thereby enhancing its applicability. The distribution of IM2LATEX, Pix2tex, HWE, and our UniMER-1M datasets can be seen in Fig. 3.

4 Methods

Mathematical expressions, originating from various sources such as electronic documents, scanned images, screenshots, and photographs, are presented against diverse backgrounds and image representations. These expressions can range from single symbols to highly complex, lengthy formulas. Although many existing algorithms are optimized for specific types of formulas, our study aims to address a broad spectrum of formula recognition problems that arise in real-world scenarios. To this end, we propose **UniMERNet**, a novel architecture capable of processing formulas of all types effectively. UniMERNet, as illustrated in Fig. 4, adopts the transformer-based encoder-decoder architecture [8] as its foundational framework.

During the training phase, each input formula image $\mathbf{I} \in \mathbb{R}^{3\times H_0\times W_0}$ undergoes an image augmentation module, which transforms a singular image representation into a diverse set of images, effectively addressing varied representations of formulas in real-world scenarios. The Swin Transformer [17] encoder then processes the image to generate the feature vector $\mathbf{Z}$, which is fed into two distinct modules. The mBART [12] decoder receives the feature vector $\mathbf{Z}$ and interacts with the output text sequence via a cross-attention mechanism, facilitating the generation of the predicted formula. Simultaneously, the Length-Aware Module utilizes the feature vector to estimate the sequence length corresponding to the original formula image. This length information is encoded and incorporated into the decoder's input, providing additional context guiding the decoder

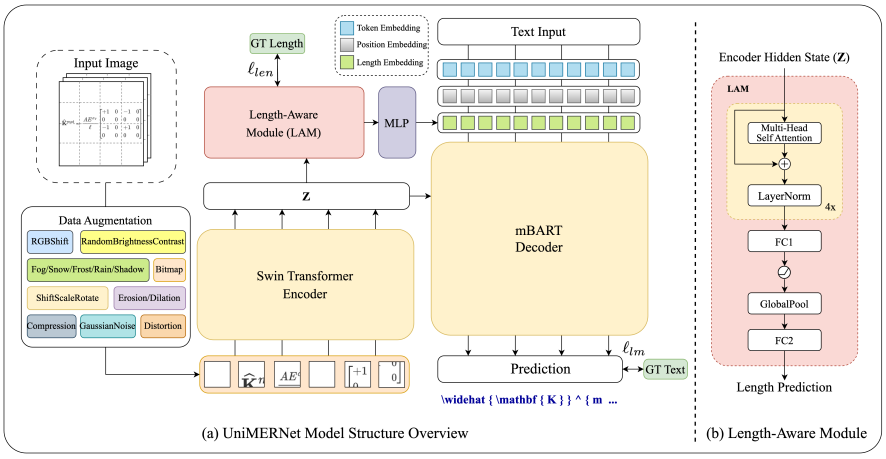

Fig. 4: UniMERNet architecture and detailed view of Length-Aware Module.

in generating the formula. The decoder combines the feature vector $\mathbf{Z}$, token embedding, position embedding, and length embedding to predict the formula. The loss function includes both the text sequence matching loss and the formula length prediction loss, defined as:

$$\mathcal{L} = \lambda_1 \ell_{lm} + \lambda_2 \ell_{len}, \tag{1}$$

where the ℓ_{lm} and ℓ_{len} correspond to language modeling loss and length loss respectively. For language modeling loss, we adopt a cross-entropy loss which minimise discrepancy between the predicted probability distribution of the next token and the actual distribution observed in the training data. Meanwhile, the length loss, ℓ_{len}, use SmoothL1 Loss to regularize the predicted length of the mathematical expressions, ensuring that the model generates length prediction matching with the visual features from encoder. The definition of two losses are given by:

$$\ell_{lm}(\hat{y}, y) = -\sum_{c=1}^{C} y_{o,c} \log(\hat{y}_{o,c}), \tag{2}$$

$$\ell_{len}(\hat{y}, y) = \begin{cases} 0.5(\hat{y} - y)^2, & \text{if } |\hat{y} - y| < 1 \\ |\hat{y} - y| - 0.5, & \text{otherwise.} \end{cases} \tag{3}$$

Our UniMERNet, anchored by the encoder-decoder Transformer architecture, provides a robust solution to the challenges of formula recognition in diverse real-world scenarios. It incorporates an innovative Length Awareness Module (LAM) that estimates the formula length from the feature vector, enhancing the mBART decoder's ability to generate accurate predictions across a wide range of formula lengths. Complemented by our data expansion and image augmentation, UniMERNet effectively addresses the varied representations of formulas, demonstrating remarkable robustness and versatility. The following sections delve into

the LAM and data augmentation techniques, underscoring their roles in enhancing UniMERNet's performance.

4.1 Length Awareness Module

Predicting real-world mathematical expressions, which range from simple symbols to complex sequences, is challenging in real-world senarios, especially in identifying the precise endpoint for longer expressions. Our solution to this challenge is the Length Awareness Module (LAM) in UniMERNet. LAM estimates the overall formula length and provides this crucial context to the decoder, enhancing formula generation effectiveness significantly. As illustrated in Fig. 4, LAM employs a self-attention mechanism and global average pooling operation to capture the interdependence of characters in long-distance sequences. The input features are derived from the feature vector obtained by the Swin Transformer encoder, with the shape of $B \times T \times D$, where B represents the batch size, T the number of patches (i.e., sequence length), and D the feature dimension of the encoder. Through a series of operations, LAM generates a fixed-length vector, known as the global counting feature, which is then mapped to the prediction space of symbol categories.

To maximize the utility of LAM's output, we introduce a multi-layer perceptron to adjust the dimension from $B \times C$ to $B \times D$, thereby obtaining the Length Embedding. This embedding, in conjunction with the Token Embedding and Positional Embedding, forms the input to the mBART decoder. This strategic design enables the decoder to consider the overall sequence length constraint when generating each token, significantly enhancing the accuracy of LaTeX mathematical formula recognition. In comparison to methods [13] used in prior works for HME tasks, LAM can make accurate estimates of formula lengths from more complex visual information of LaTeX-rendered formulas. Our task uses the complete LaTeX syntax, where the visual features of tokens are more complex, making accurate recognition more challenging. However, through our design, LAM not only adapts to LaTeX input sequences of different lengths but also dynamically adjusts its prediction according to the sequence content, demonstrating superior performance in complex image-to-text recognition tasks such as mathematical formula recognition.

4.2 Data Augmentation for Model Training

UniMERNet's efficacy in MER is significantly enhanced through a comprehensive image augmentation strategy. In the context of enhancing the model's adaptability to the variations between synthesized LaTeX-rendered training images and real-world test images, such as those captured from screens or photographed with inherent noise, a meticulously crafted data augmentation strategy was used during model training to simulate this diversity. This includes, but is not limited to, image dilation, erosion, and weather noise (fog, frost, rain, snow, shadow). Implementing these augmentations, alongside the data expansion strategy used

Table 2: Ablation results on UniMER-Test for models using different augmentations. Here, "HME" refers to a mixed dataset of CHROME and HME100K.

Train Dataset	SPE		CPE		SCE		HWE	
	BLEU ↑	EditDis ↓	BLEU ↑	EditDis ↓	BLEU ↑	EditDis ↓	BLEU ↑	EditDis ↓
Pix2tex	**0.926**	**0.051**	0.790	0.164	0.545	0.373	0.087	0.775
Pix2tex+HWE	0.884	0.064	0.695	0.209	0.490	0.291	0.895	0.070
UniMER-1M	0.917	0.058	**0.916**	**0.060**	**0.616**	**0.229**	**0.921**	**0.055**

in the UniMER-1M training dataset, ensures UniMERNet's stable training, effectively preparing it to tackle the diversity and complexity of real-world mathematical expressions, thereby enhancing its recognition accuracy.

5 Experiments

5.1 Datasets and Evaluation Metrics

We utilize the UniMER-1M dataset to train our model and evaluate its formula recognition performance using the UniMER-Test. Our evaluation relies on BLEU, edit distance, and ExpRate metrics.

BLEU: The BLEU score [22], initially developed for machine translation, quantifies the match of n-grams between candidate and reference sentences. Its application to a similar conversion task of formula recognition provides a robust, quantitative performance measure. **Edit distance:** The edit distance [11] measures the minimum character changes needed to convert one string to another. Its use in formula recognition offers a precise, character-level accuracy assessment, making it a valuable performance metric. **ExpRate:** Expression Recognition Rate (ExpRate) [29] is a widely used metric for handwritten formula recognition, defined as the percentage of predicted mathematical expressions that perfectly match the actual results.

5.2 Implementation Details

The proposed model, UniMERNet, uses PyTorch with a maximum sequence length set to 1024. Training is conducted on a single GPU equipped with CUDA. Specifically, we utilize an NVIDIA A100 with 80GB of memory. During the training phase, we employ four such GPUs with a batch size of 64. The learning rate schedule is linear warmup cosine, with an initial learning rate of 1×10^{-4}, a minimum learning rate of 1×10^{-8}, and a warmup learning rate of 1×10^{-5}. Weight decay is set to 0.05. The total iteration is set to 180,000. The loss weight λ_1 and λ_2 are set to 1 and 0.5 by our default settings.

5.3 Ablation Study

UniMER-1M The diversity and quantity of training data are crucial for a model's accurate recognition of various formula types. As shown in Tab. 2,

Table 3: Ablation results on UniMER-Test with models using different augmentations.

Train Dataset	Augment	SPE		CPE		SCE		HWE	
		BLEU ↑	EditDis ↓	BLEU ↑	EditDis ↓	BLEU ↑	EditDis ↓	BLEU ↑	EditDis ↓
Pix2tex	✗	0.925	0.051	0.779	0.174	0.520	0.373	0.087	0.004
	✔	**0.926**	**0.051**	0.790	0.164	0.545	0.373	0.087	0.775
UniMER-1M	✗	0.916	0.059	0.907	0.063	0.559	0.252	0.912	0.056
	✔	0.917	0.058	**0.916**	**0.060**	**0.616**	**0.229**	**0.921**	**0.055**

Table 4: Ablation of Length-Aware Module on UniMER-Test.

LAM	SPE		CPE		SCE		HWE	
	BLEU ↑	EditDis ↓	BLEU ↑	EditDis ↓	BLEU ↑	EditDis ↓	BLEU ↑	EditDis ↓
✗	0.918	0.056	0.893	0.065	0.610	0.227	0.920	0.055
✔	0.917	0.058	**0.916**	**0.060**	0.616	0.229	0.921	0.055

UniMERNet, when trained solely on the Pix2tex dataset, performs well on the SPE subset (BLEU score of 0.926), but poorly on CPE, SCE, and HWE subsets. The simplicity of Pix2tex leads to overfitting on SPE and difficulty recognizing complex and handwritten formulas. However, when trained on both Pix2tex and HWE datasets, performance on the HWE subset improves significantly (BLEU score of 0.895), but there's a slight decline on SPE, CPE, and SCE. Importantly, when trained with our proposed UniMER-1M dataset, UniMERNet excels across all subsets. Compared to training on Pix2tex+HWE, the BLEU score improves by 0.221 on the CPE subset, and the edit distance decreases from 0.209 to 0.060. On the SCE subset, the BLEU score improves by 0.126, and the edit distance decreases from 0.291 to 0.229. For the HWE subset, the BLEU score improves by 0.026, and the edit distance decreases to 0.055.

Data Augmentation In real-world formula recognition tasks, we frequently deal with noisy images, such as those from scanned documents or photographs. To address this, we've incorporated an image augmentation module in our approach. This module simulates a variety of image alterations that may occur in real-world scenarios using diverse image augmentation techniques during training. As demonstrated in Tab. 3, when training solely with the Pix2tex dataset, the addition of image augmentation results in consistent improvements across all evaluation subsets. This is particularly noticeable on the SCE subset, where the BLEU score improves by 2.50%. When we train with the UniMER-1M dataset, we observe similar trends. The improvement on the SCE subset is even more pronounced, with the BLEU score increasing from 0.559 to 0.616 and the edit distance reducing from 0.252 to 0.229.

Length-Aware Module The LAM we propose in this paper, when integrated into UniMERNet, significantly enhances the accuracy and stability of formula detection across varying complexities. Using the UniMER-1M dataset, we trained

Table 5: Comparison with SOTA methods on UniMER-Test .

Method	SPE		CPE		SCE		HWE	
	BLEU ↑	EditDis ↓	BLEU ↑	EditDis ↓	BLEU ↑	EditDis ↓	BLEU ↑	EditDis ↓
Pix2tex [3]	0.873	0.088	0.655	0.408	0.092	0.817	0.012	0.920
Texify [23]	0.906	0.061	0.690	0.230	0.420	0.390	0.341	0.522
UniMERNet	**0.917**	**0.058**	**0.916**	**0.060**	**0.616**	**0.229**	**0.921**	**0.055**

Table 6: Comparison with SOTA methods on handwriting expression dataset. ExpRate (Perfect Match), ≤1, ≤2, denoting the proportions of expressions accurately transcribed or requiring up to 1 or 2 modifications, respectively.

Method	CROHME2014			CROHME2016			CROHME2019			HME100K		
	ExpRate	≤1	≤2	ExpRate	≤1	≤2	ExpRate	≤1	≤2	ExpRate	≤1	≤2
DWAP [30]	50.1	-	-	47.5	-	-	-	-	-	-	-	-
DWAP-MSA [30]	52.8	68.1	72.0	50.1	63.8	67.4	-	-	-	-	-	-
BTTR [34]	54.0	66.2	70.3	52.3	63.9	68.6	53.0	66.0	69.1	-	-	-
CoMER [33]	59.3	71.7	75.7	59.8	74.4	80.3	63.0	77.4	81.4	-	-	-
CAN-DWAP [13]	65.6	77.4	83.4	62.5	74.6	82.5	63.2	78.1	82.5	67.3	82.9	89.2
CAN-ABM [13]	65.9	78.0	84.2	63.1	75.9	82.7	64.5	78.7	83.0	**68.1**	83.2	**89.9**
SAN [29]	63.1	75.8	82.0	61.5	73.3	81.4	62.1	74.5	81.0	67.1	-	-
UniMERNet	**67.4**	**80.9**	**89.2**	**68.4**	**82.3**	**89.2**	**65.4**	**80.7**	**87.6**	68.0	**84.1**	89.6

two models—one with LAM and one without—and evaluated their performance on the UniMER-Test. The evaluation results are displayed in Tab. 4. As can be observed, our UniMERNet model, even without the LAM module, achieves exceptional results overall, owing to the comprehensive UniMERNet, image augmentation during training, and the encoder-decoder architecture. The BLEU scores on SPE, CPE, and HWE are nearly or above 0.90; even for noisy CPE, the score is 0.60. With the integration of the LAM, UniMERNet's performance on the complex long formula combination CPE significantly improves by 2% while maintaining stable performance on other subsets.

5.4 Comparison with State-of-the-Art

Comparison with Printed Expression Recognition Methods To more intuitively measure the formula recognition performance of UniMERNet, we made a full comparison with the current state-of-the-art (SOTA) methods that are specifically designed for printed-type formulas such as Pix2tex and Texify. As can be seen from Tab. 5, for simple printed formulas SPE, our model is significantly higher than the Pix2tex [3] and Texify [23] models in terms of both BLEU and edit distance. On CPE and SCE, our results far exceed other two methods, with BLEU scores improving by 0.226, 0.196 compare to previous SOTA.

Comparison with Handwritten Expression Recognition Methods To fairly evaluate our model's ability to recognize handwritten formulas, we compared it with some of the most mainstream handwritten recognition models. As shown in Tab. 6, although UniMERNet was not specifically optimized for handwritten formulas, it surpasses the SOTA handwritten recognition model across

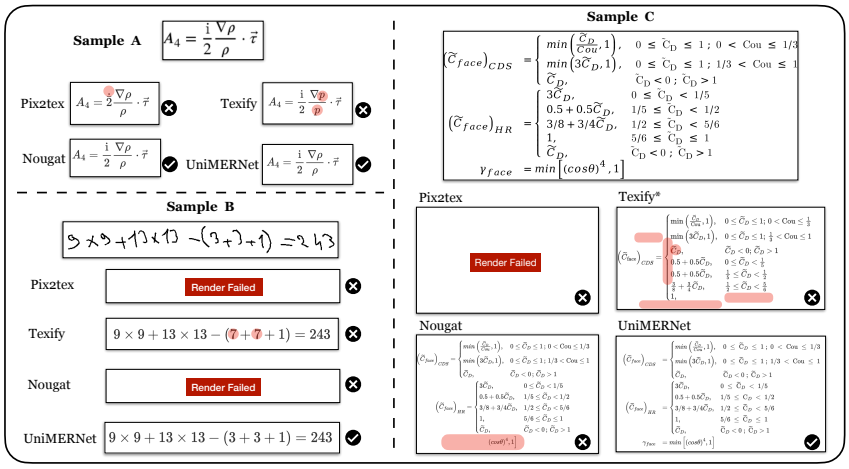

Fig. 5: Comparative Visualization of Recognition Results Using Different Methods.

all CROHME test sets. On the CROHME 2014, CROHME 2016, CHOME 2019, and HME 100K test sets, UniMERNet achieves a completely accurate formula recognition rate (without any modifications) of 67.4%, 68.4%, and 65.4%, outperforming the SOTA results on each dataset by 1.5%, 5.3%, and 1.1% respectively. Furthermore, on the HME100K dataset, our model achieves comparable high-quality results to the specifically designed SOTA model.

Qualitative Comparisons As shown in Fig. 5, we selected three representative samples from the UniMER-Test set to thoroughly compare the performance between Pix2tex [3], Texify [23], Nougat [4], and UniMERNet. It's important to highlight that Nougat, being primarily designed for full-page recognition, tends to underperform with isolated formulas; thus, we prepared the test images by integrating random text with the formulas to adapt to Nougat's inference capabilities. Notably, while the other models exhibit certain shortcomings in handling these test samples, our model consistently delivers robust and accurate recognition results.

6 Conclusion

This research presents the UniMER-1M dataset and UniMER-Test, substantial contributions to the Mathematical Expression Recognition (MER) field, offering unprecedented scale and diversity. Our novel UniMERNet, validated through rigorous experimentation, sets a new benchmark in MER. The model's robustness and adaptability to varied lengths and noise levels underscore its practical value. Our resources are publicly available, aiming to catalyze further advancements in MER. Future work includes refining UniMERNet and exploring its fusion with LVLMs for holistic document understanding. This study represents a pivotal stride in MER, providing a robust foundation for future research.

References

1. Anderson, R.H.: Syntax-directed recognition of hand-printed two-dimensional mathematics. In: Symposium on interactive systems for experimental applied mathematics: Proceedings of the Association for Computing Machinery Inc. Symposium. pp. 436–459 (1967) 3
2. Bian, X., Qin, B., Xin, X., Li, J., Su, X., Wang, Y.: Handwritten mathematical expression recognition via attention aggregation based bi-directional mutual learning. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 36, pp. 113–121 (2022) 4
3. Blecher, L.: pix2tex - latex ocr. https://github.com/lukas-blecher/LaTeX-OCR (2022), accessed: 2024-2-29 2, 4, 6, 7, 13, 14
4. Blecher, L., Cucurull, G., Scialom, T., Stojnic, R.: Nougat: Neural optical understanding for academic documents. arXiv.org **2308.13418** (2023) 2, 4, 14
5. Chan, K., Yeung, D.: Error detection, error correction and performance evaluation in on-line mathematical expression recognition (Jan 1999) 3
6. Deng, Y., Kanervisto, A., Ling, J., Rush, A.M.: Image-to-markup generation with coarse-to-fine attention. In: International Conference on Machine Learning (ICML). pp. 980–989. PMLR (2017) 2, 4, 6, 7
7. Dong, X., Zhang, P., Zang, Y., Cao, Y., Wang, B., Ouyang, L., Wei, X., Zhang, S., Duan, H., Cao, M., et al.: Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. arXiv.org **2401.16420** (2024) 4
8. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Ocr-free document understanding transformer. In: European Conference on Computer Vision (ECCV). pp. 498–517. Springer (2022) 2, 4, 8
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. NeurIPS **25** (2012) 4
10. Le, A.D., Indurkhya, B., Nakagawa, M.: Pattern generation strategies for improving recognition of handwritten mathematical expressions. Pattern Recognition Letters **128**, 255–262 (2019) 2, 4
11. Levenshtein, V.I., et al.: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady. vol. 10, pp. 707–710. Soviet Union (1966) 11
12. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv.org (2019) 8
13. Li, B., Yuan, Y., Liang, D., Liu, X., Ji, Z., Bai, J., Liu, W., Bai, X.: When counting meets hmer: counting-aware network for handwritten mathematical expression recognition. In: European Conference on Computer Vision (ECCV). pp. 197–214. Springer (2022) 4, 10, 13
14. Li, Z., Jin, L., Lai, S., Zhu, Y.: Improving attention-based handwritten mathematical expression recognition with scale augmentation and drop attention. In: International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 175–180. IEEE (2020) 4
15. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv.org **2310.03744** (2023) 4
16. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. vol. 36 (2024) 4
17. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (2021) 8

18. Mahdavi, M., Zanibbi, R., Mouchere, H., Viard-Gaudin, C., Garain, U.: Icdar 2019 crohme+ tfd: Competition on recognition of handwritten mathematical expressions and typeset formula detection. In: International Conference on Document Analysis and Recognition (ICDAR). pp. 1533–1538. IEEE (2019) 2, 6, 7
19. Miller, E., Viola, P.: Ambiguity and constraint in mathematical expression recognition. National Conference on Artificial Intelligence (NCAI) (Jul 1998) 3
20. Mouchere, H., Viard-Gaudin, C., Zanibbi, R., Garain, U.: Icfhr 2014 competition on recognition of on-line handwritten mathematical expressions (crohme 2014). In: International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 791–796. IEEE (2014) 6, 7
21. Mouchère, H., Viard-Gaudin, C., Zanibbi, R., Garain, U.: Icfhr2016 crohme: Competition on recognition of online handwritten mathematical expressions. In: International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 607–612. IEEE (2016) 6, 7
22. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002) 11
23. Paruchuri, V.: Texify. https://github.com/VikParuchuri/texify (2023), accessed: 2024-2-29 2, 4, 13, 14
24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR. Computational and Biological Learning Society (2015) 4
25. Suzuki, M., Tamari, F., Fukuda, R., Uchida, S., Kanahori, T.: Infty: an integrated ocr system for mathematical documents. In: Proceedings of the 2003 ACM symposium on Document engineering. pp. 95–104 (2003) 3
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017) 2, 4
27. Wei, H., Kong, L., Chen, J., Zhao, L., Ge, Z., Yang, J., Sun, J., Han, C., Zhang, X.: Vary: Scaling up the vision vocabulary for large vision-language models. arXiv.org **2312.06109** (2023) 2, 4
28. Wu, J.W., Yin, F., Zhang, Y.M., Zhang, X.Y., Liu, C.L.: Handwritten mathematical expression recognition via paired adversarial learning. International Journal of Computer Vision (IJCV) **128**, 2386–2401 (2020) 2, 4
29. Yuan, Y., Liu, X., Dikubab, W., Liu, H., Ji, Z., Wu, Z., Bai, X.: Syntax-aware network for handwritten mathematical expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4553–4562 (2022) 6, 7, 11, 13
30. Zhang, J., Du, J., Dai, L.: Multi-scale attention with dense encoder for handwritten mathematical expression recognition. In: Proc. of the International Conf. on Pattern Recognition (ICPR). pp. 2245–2250. IEEE (2018) 13
31. Zhang, J., Du, J., Yang, Y., Song, Y.Z., Wei, S., Dai, L.: A tree-structured decoder for image-to-markup generation. In: International Conference on Machine Learning (ICML). pp. 11076–11085. PMLR (2020) 4
32. Zhang, J., Du, J., Zhang, S., Liu, D., Hu, Y., Hu, J., Wei, S., Dai, L.: Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition. Pattern Recognition **71**, 196–206 (2017) 4
33. Zhao, W., Gao, L.: Comer: Modeling coverage for transformer-based handwritten mathematical expression recognition. In: European Conference on Computer Vision (ECCV). pp. 392–408. Springer (2022) 13

34. Zhao, W., Gao, L., Yan, Z., Peng, S., Du, L., Zhang, Z.: Handwritten mathematical expression recognition with bidirectionally trained transformer. In: International Conference on Document Analysis and Recognition (ICDAR). pp. 570–584. Springer (2021) 2, 4, 13
35. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv.org **2304.10592** (2023) 4