

Uni-MuMER: Unified Multi-Task Fine-Tuning of Vision-Language Model for Handwritten Mathematical Expression Recognition

Yu Li* Jin Jiang* Jianhua Zhu Shuai Peng Baole Wei Yuxuan Zhou Liangcai Gao✉
Wangxuan Institute of Computer Technology, Peking University, Beijing, China
liyu@stu.pku.edu.cn jiangjin@stu.pku.edu.cn zhujianhuapku@pku.edu.cn glc@pku.edu.cn

Abstract

Handwritten Mathematical Expression Recognition (HMER) remains a persistent challenge in Optical Character Recognition (OCR) due to the inherent freedom of symbol layouts and variability in handwriting styles. Prior methods have faced performance bottlenecks by proposing isolated architectural modifications, making them difficult to integrate coherently into a unified framework. Meanwhile, recent advances in pretrained vision-language models (VLMs) have demonstrated strong cross-task generalization, offering a promising foundation for developing unified solutions. In this paper, we introduce Uni-MuMER, which fully fine-tunes a VLM for the HMER task without modifying its architecture, effectively injecting domain-specific knowledge into a generalist framework. Our method integrates three data-driven tasks: Tree-Aware Chain-of-Thought (Tree-CoT) for structured spatial reasoning, Error-Driven Learning (EDL) for reducing confusion among visually similar characters, and Symbol Counting (SC) for improving recognition consistency in long expressions. Experiments on the CROHME and HME100K datasets show that Uni-MuMER achieves super state-of-the-art performance, outperforming the best lightweight specialized model SSAN by 16.31% and the top-performing VLM Gemini2.5-flash by 24.42% under zero-shot setting. Our datasets, models, and code are open-sourced at: <https://github.com/BFlameSwift/Uni-MuMER>

1 Introduction

Handwritten Mathematical Expression Recognition (HMER) seeks to translate handwritten expressions into machine-readable markup, supporting document understanding and digital preservation of scientific material. Unlike standard Optical Character Recognition (OCR), HMER involves parsing complex 2D structures [15], ambiguous symbols, and the inherent freedom of symbol layout and variability in handwriting styles [68], requiring not just recognizing individual symbols but also layout reasoning and parsing their complex spatial relationships. Recent approaches have primarily leveraged RNNs [23] or Transformers [48] due

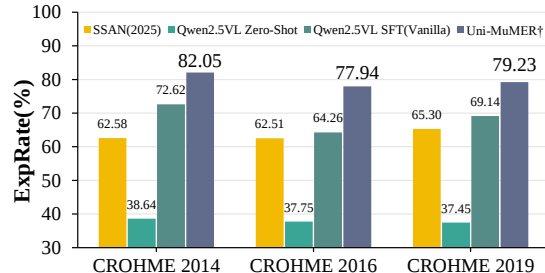


Figure 1: **Performance comparison on CROHME sets (%)**. Our Uni-MuMER[†] achieves significant improvements compared to the previous SOTA (SSAN) and Qwen2.5VL Zero-Shot.

* Equal contribution, ✉Corresponding authors.

to their powerful sequential modeling capabilities. Additional modules, such as tree-structured decoders [64, 52, 70] and relative-position-aware mechanisms [18, 62, 53, 54], have been introduced to further enhance performance. These dominant research directions in HMER have focused on incorporating prior human knowledge into model architectures, particularly through manually designed structural modules or attention mechanisms [63, 65, 66, 67, 30, 18, 4].

Despite the importance of HMER and numerous proposed improvements, the field has seen only marginal progress in recent years. As is shown in Tab. 1, performance on the CROHME datasets [38, 39, 37] has improved by merely 3% (from CoMER [66] to SSAN [62]), underscoring an urgent need for a novel paradigm. The limited progress of these models arose from three constraints: (1) Improvements are isolated and model-specific, making them hard to integrate or scale. (2) Optimizing across multiple auxiliary tasks remains challenging, with many approaches focusing on singular priors rather than adopting a unified, multi-faceted enhancement [7]; and (3) Models trained on single-domain datasets, struggling to generalize to other datasets, lack essential scalability and transferability [68, 46]. Additionally, widely standard metrics like ExpRate fail to capture visual equivalence in $\text{\LaTeX}$ outputs, and the scarcity of diverse data exacerbates these constraints. Recently introduced visual metrics such as CDM and large-scale datasets like Mathwriting offer opportunities for large-scale cross-dataset training and multiple $\text{\LaTeX}$ syntax style evaluation.

Rapid advancements in pretrained vision-language models (VLMs) have significantly enhanced foundational recognition capabilities and generalization across related tasks [8, 25, 28, 31, 32]. While early benchmarks like OCRBench [34] reported poor HMER performance for generalist VLMs, Tab. 1 presents our evaluation of recent open-source and closed-source VLMs, such as Qwen2.5-VL [3], Doubao-1.5-pro [45], Gemini2.5-flash [11], and GPT4o [41]. Our results indicate that these large-scale models have exhibited unexpectedly strong capabilities in handling structured recognition tasks. However, these high-performance closed-source models trained on massive amounts of undisclosed data [1] make it difficult to pinpoint how to systematically improve performance on HMER. Consequently, empowering open-source VLMs to achieve comparable or superior HMER performance remains an open and urgent challenge.

To bridge this gap, we propose Uni-MuMER, a unified multi-task fine-tuning framework for enhancing open-source VLMs in HMER. Unlike previous methods constrained by single-domain datasets or isolated architectural improvements, Uni-MuMER fully exploits existing data resources through data-driven fine-tuning. Another motivation of our work is to unify previously fragmented HMER tasks into a unified framework, shifting the focus from incremental architectural modifications toward generalizable recognition capabilities. Concretely, we introduce specialized training data and learning objectives, employing novel tasks such as Tree-Aware Chain-of-Thought for explicit structural reasoning [51, 57], Error-driven learning to reduce symbol confusion [36], and Symbol Counting [30] to enhance parsing expressions capabilities. Fig. 1 nicely illustrates the jump with Uni-MuMER; it exceeds Qwen-2.5-VL in the zero-shot setting and the vanilla SFT setting across multiple CROHME datasets. Besides, owing to advancements in open-source inference frameworks (e.g., vllm [29]), our method achieves superior inference speeds compared to traditional methods, enhancing its practical applicability. Our contributions are listed below:

- We propose a new unified paradigm for HMER. Unlike prior methods that heavily rely on specialized networks and single-task training, our data-driven multi-task method injects domain knowledge into a generalist VLM, yielding cumulative performance gains.
- We introduce three data-driven tasks: Tree-Chain-of-Thought, Error-Driven Learning, and Symbol Counting. These systematically address the challenges of HMER, which are two-dimensional structure, ambiguous handwriting, and output consistency.
- Our Uni-MuMER method achieves new SOTA results on the CROHME and HME100K datasets. Notably, it attains 79.74% averaged across CROHME datasets (+41.79% over Qwen2.5-VL, +24.42% over Gemini2.5-flash in zero-shot setting, and +16.31% over specialized models SSAN).

2 Related Work

2.1 HMER

Early rule-based approaches attempted to address the challenges of HMER through symbol segmentation, recognition, and syntactic parsing based on handcrafted grammars [6, 60, 10, 2, 22, 47]. These methods struggled with generalization to diverse handwriting styles and complex layout structures.

Sequence-based decoding methods The emergence of deep learning shifted HMER towards end-to-end sequence modeling tasks. Early encoder-decoder architectures based on RNNs, WAP [65] first applied sequence-to-sequence learning to HMER. Subsequent RNN-based models improved visual encoding [63] and robustness to handwriting style [53, 54]. Inspired by the Transformer’s success [48], BTTR [67] introduced the first Transformer-based model for HMER. To enhance alignment during decoding, CoMER [66] introduced a coverage attention mechanism.

Multi-task Learning Beyond sequence modeling, researchers have explored structural decoders and multi-task models to better capture 2D structure. One central line of work focuses on modeling the hierarchical tree structure of expressions. TreeDecoder (TD) [64] and its improved version TDv2 directly predict a tree-structured representation of the expression. SAN [58] introduced syntactic constraints into the decoding process, and TAMER [70] jointly optimizes both sequence and tree decoding within a unified Transformer framework. Various auxiliary tasks have also been proposed to inject domain knowledge: ABM [4] adds a dual-direction decoder loss to enhance context modeling. RLFN [9] fuses a language-model-based module with the recognizer to leverage $\LaTeX$ syntax and context. ICAL [69] proposed an implicit character construction to capture latent symbol-level semantics. To address symbol omission and repetition, CAN [30] incorporated an auxiliary symbol-counting task. UniMERNet [49] adds a Length-Aware Module targeting real-world expressions’ extreme length variance. PosFormer [18] and SSAN [62] explicitly model spatial relationships among symbols to guide the network’s attention.

2.2 Vision-Language Models

Vision-Language Models (VLMs), initially popularized by contrastive learning frameworks like CLIP [43], laid the groundwork for powerful zero-shot multimodal recognition. Donut [27] and LayoutLMv3 [24] extended this capability specifically to OCR tasks, effectively recognizing diverse textual content. High-resolution, document-centric VLMs, notably Monkey [33] and TextMonkey [35], enhanced OCR performance by combining patch-based image encoding with explicit textual supervision. For mathematical OCR, models like Im2LaTeX [12] and Nougat [5] focused explicitly on end-to-end LaTeX reconstruction from printed equations and scientific documents. The recent FERMAT benchmark [40] explicitly evaluated multiple SOTA VLMs on handwritten mathematics recognition, revealing critical gaps between general text OCR and handwritten mathematical recognition. Domain-specific adaptations such as VLPG [20] and HiE-VL [21] propose graph-based or hierarchical adapters to enhance mathematical recognition, although their extensive architectural modifications and modest performance limit broader adoption. In contrast, recent general-purpose frameworks like Qwen2.5-VL [3], equipped with dynamic-resolution ViTs and structured outputs, demonstrate promising potential for adaptation to mathematical OCR tasks.

3 Method

We introduce Uni-MuMER, a unified multi-task fine-tuning framework for enhancing open-source VLMs in HMER. The overall pipeline is shown in Fig. 2. Given an input image of a handwritten expression and a task-specific instruction, the model is trained to produce the corresponding output. Uni-MuMER integrates four tasks: Vanilla HMER, Tree-aware Chain-of-Thought, Error-Driven Learning, and Symbol Counting, into a unified training paradigm. The primary goal is to adapt a general-purpose VLM to the highly structured and domain-specific knowledge of HMER without modifying any original architecture. Subsequent subsections provide additional details for each task.

3.1 Vanilla HMER

Base Model We adopt Qwen2.5-VL-3B [3] as the VLM Backbone. It comprises a ViT-based visual encoder and a transformer-based language decoder, pre-trained to perform various image-to-text tasks, which provide robust visual grounding and structured sequence-generation capabilities, rendering it an effective foundation for end-to-end fine-tuning in HMER tasks.

Vanilla HMER As is shown at the top of Fig. 2, Vanilla HMER involves providing an image of a mathematical expression alongside a textual instruction, prompting the model to directly output the corresponding $\LaTeX$ formatted expression. Traditional lightweight specialized models took images alone as input and generated the expressions, whereas VLMs now exhibit strong generalization capabilities across related tasks. We also utilize models fine-tuned specifically on the Vanilla HMER

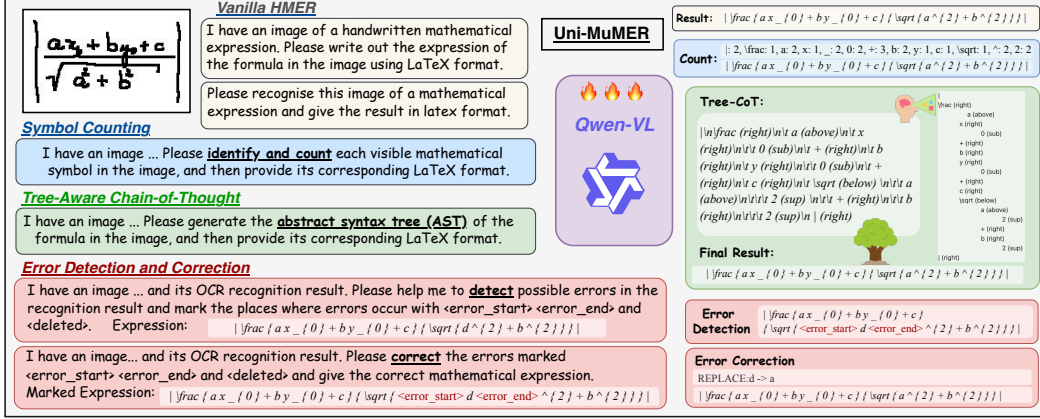


Figure 2: **Overview of our Uni-MuMER Framework:** which augments a VLM base with four distinct integrated tasks: Vanilla HMER, Tree-Aware Chain-of-Thought, Error-Driven Learning, and Symbol Counting to generate robust and accurate $\text{\LaTeX}$ from an input image of a handwritten expression.

task as baselines for subsequent comparison. Through Vanilla HMER, the model acquires essential recognition capabilities and ensures accurate and structured outputs.

3.2 Tree-Aware Chain-of-Thought

We propose a Tree-CoT task that employs parsing Abstract Syntax Trees (ASTs) as intermediate cognitive structures for expression recognition. The model accepts an input comprising an image and an instruction, subsequently generating both a serialized tree and the final recognized $\text{\LaTeX}$ expression. Inspired by ASTs widely utilized in mathematical information retrieval [61, 42] and previous tree-decoding methods in HMER [64, 52, 70], we incorporate explicit tree-structured decoding into the textual output. This explicit intermediate representation enables the model to directly reason the inherent 2D spatial relationships among symbols. Consequently, Tree-CoT guides the model to produce structurally coherent and semantically accurate $\text{\LaTeX}$ outputs by bridging visual input processing and linear textual serialization. Below, we describe in detail the construction.

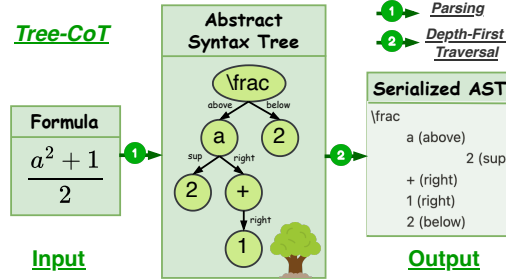


Figure 3: **Illustration of the proposed Tree-CoT construction and serialization procedure.** For input expression, we parse it into an AST, then traverse in depth-first order and serialize it into a structured textual format.

Tree-CoT The construction procedure of Tree-CoT is illustrated in Fig. 3. We first parse the input expression to derive its Abstract Syntax Tree (AST), explicitly capturing hierarchical and spatial relationships among symbols. We subsequently perform a depth-first traversal (DFS) of the AST to sequentially linearize the tree structure. To encode this structured representation, we introduce a specific serialization scheme, using tab-based indentation to reflect tree depth and newline separation to distinguish individual nodes clearly, and the raw text is in Fig. 2. Each serialized line explicitly contains the symbol label and its spatial relation to its immediate parent node, resulting in a coherent textual representation of the AST.

3.3 Error-Driven Learning

Error-Driven Learning (EDL) leverages a learning-from-mistakes paradigm to enhance model accuracy. Its input consists of an error corpus generated by the model itself, structured into two distinct subtasks: error detection and error correction. Inspired by self-improvement strategies and previous

work indicating the suitability of language models for correcting HMER errors [9], we integrate these subtasks explicitly. The subsequent sections detail the generation of the error corpus and the definitions of error detection and correction subtasks.

Error Corpus Generation As is shown in the Fig. 4. Initially, the complete dataset (e.g., CROHME) is randomly partitioned into multiple distinct subsets denoted as $F_1, F_2, \dots, F_n$. Subsequently, we perform cross-validation training, where each VLM is trained on $n - 1$ folds and evaluated on the remaining held-out fold F_i . During this, multiple candidate predictions are generated through multiple sampling for each input image to collect potential model outputs. By comparing these candidate predictions to their corresponding ground-truth labels, we explicitly identify erroneous outputs and compile these erroneous predictions into the final error corpus.

Error Detection and Correction Based on this error corpus, we formulate two related subtasks: error detection and error correction. As shown in the Fig. 2, the error-detection task takes as input the original expression image alongside a potentially erroneous predicted expression, outputting a marked expression wherein errors are explicitly enclosed by `<error_start>` and `<error_end>` tags, and omissions are explicitly indicated by `<deleted>` tags. The subsequent error-correction subtask receives this masked expression as input and produces the correction log and corresponding $\LaTeX$ expression. The scale of the error corpus obtained for each training set used in Sec. 4.1 is approximately equal to that of the original training data.

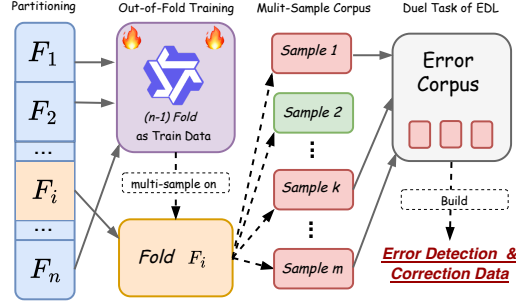


Figure 4: **Illustration of Error Corpus Generation for Error-Driven Learning Task (EDL)**, where erroneous model outputs are collected through cross-fold training and multiple sampling from held-out evaluation subsets.

3.4 Symbol Counting Auxiliary Task

We introduce a Symbol Counting (SC) auxiliary task designed to encourage the model to explicitly account for all symbols appearing in expressions. The input to this task comprises an expression image along with a counting instruction, and the expected output consists of the total count of visible symbols in the expression alongside the corresponding $\LaTeX$ expression. Inspired by the CAN [30], we observe that models frequently produce locally coherent yet globally inconsistent outputs, especially involving repeated symbols. To mitigate this issue, we explicitly integrate symbol-counting information into the textual output representation. This task compels the model to accurately predict symbol counts, thus reducing the likelihood of symbol omissions or hallucinations during final $\LaTeX$ expression generation.

Symbol Counting. We construct training targets that prepend the count string to the actual $\LaTeX$. Specifically, for a given handwritten expression image, the target output becomes `<Count String>\n<Expression string>`. For example: For the expression $\frac{a^2+1}{2}$, the textual count string is: `\frac{1}{2}, a: 1, 2: 2, +: 1`.

4 Experiments

4.1 Experimental Setup

This section details the experimental setup used to evaluate the Uni-MuMER, including datasets, evaluation metrics, training settings, and baselines.

Datasets The CROHME series [37, 38, 39] is the most widely used dataset for HMER. It comprises 8836 training samples, and three test sets contain 986, 1147, and 1199 images for the CROHME 2014, 2016, 2019. Recently, the CROHME 2023 [56] iteration provided an expanded training set of 10979 images and a test set of 2300 images. HME100K [59] is a large-scale, real-world dataset. It consists of 74502 training and 24607 testing images, encompassing various writing styles and conditions. MathWriting [17], released by Google Research in 2024, is the largest HME corpus to date with 230K human-written and 400K synthetically-generated expressions. Finally, Im2Latexv2 [44] builds upon the printed expression dataset Im2Latex-100k [12] and enhances it with improved $\LaTeX$ normalization and 59 diverse rendering styles for more realistic expression images.

Table 1: **Performance comparison on CROHME datasets(%)**. We evaluate our model and four baselines using expression recognition rate (Exp) and ExpRate@CDM (@CDM) on the CROHME 2014, 2016 and 2019 test sets. We use the following notation for reported results: All models are evaluated under zero-shot inference settings, ^a denotes training with data augmentation, ^p denotes reproduced by us, and [†] denotes the use of external data.

Method	CROHME14		CROHME16		CROHME19		Average	
	Exp ↑	@CDM ↑	Exp ↑	@CDM ↑	Exp ↑	@CDM ↑	Exp ↑	@CDM ↑
<i>Closed-Source Models</i>								
GPT-4o[41]	50.61	60.85	46.03	50.65	49.79	57.13	48.81	56.21
Doubao-1.5-pro[45]	50.41	69.17	49.00	64.08	46.87	64.05	48.79	65.77
Gemini2.5-flash [41]	58.01	69.98	52.74	61.34	55.21	65.38	55.32	65.57
<i>Open-Source Models</i>								
Qwen2.5-VL-3B[3]	38.64	44.01	37.75	43.33	37.45	43.04	37.95	43.46
Qwen2.5-VL-7B[3]	55.98	65.51	50.92	56.32	49.62	56.55	52.17	59.46
Qwen2.5-VL-72B[3]	59.74	68.56	54.32	59.37	55.13	61.05	56.40	62.99
<i>Specialized Models</i>								
DenseWAP[63]	50.1	–	47.5	–	–	–	–	–
BTTR[67]	53.96	–	52.31	–	52.96	–	53.08	–
ABM[4]	56.85	–	52.92	–	53.96	–	54.58	–
CAN-ABM[30]	57.26	–	56.15	–	55.96	–	56.46	–
CoMER[66]	58.38	61.66	56.98	63.03	59.12	62.71	58.16	62.47
ICAL	60.63	–	58.79	–	60.51	–	59.98	–
PosFormer[18]	60.45	–	60.94	–	62.22	–	61.20	–
TAMER[70]	61.36	63.28	59.54	62.94	60.13	63.46	60.34	63.23
SSAN[62]	60.85	–	60.02	–	61.83	–	60.90	–
CoMER ^a [66]	59.33	–	59.81	–	62.97	–	60.70	–
CoMER ^{ap†} [66]	46.96	53.55	43.33	48.47	48.25	54.84	46.18	52.29
PosFormer ^a [18]	62.74	65.88	61.03	65.39	64.97	69.30	62.91	66.85
SSAN ^a [62]	62.58	–	62.51	–	65.30	–	63.43	–
UniMERNet ^{a†}	67.4	–	68.4	–	65.4	–	67.07	–
VLPG [†] [20]	60.41	–	60.51	–	62.34	–	61.09	–
HiE-VL ^{a†} [21]	73.30	–	70.70	–	69.3	–	71.13	–
Vanilla (baseline)	72.62	75.66	64.26	67.83	69.14	71.98	68.64	71.82
Uni-MuMER	75.36 ^{+2.74}	79.11 ^{+3.45}	70.79 ^{+6.53}	75.24 ^{+7.41}	73.73 ^{+4.59}	77.23 ^{+5.25}	73.29 ^{+4.65}	77.19 ^{+5.37}
Uni-MuMER[†]	82.05^{+9.43}	85.09^{+9.43}	77.94^{+13.68}	81.08^{+13.25}	79.23^{+10.09}	82.40^{+10.42}	79.74^{+11.10}	82.86^{+11.04}

Evaluation Metrics The standard metric for HMER is **Expression Recognition Rate (ExpRate)**, which measures exact-match accuracy. However, Exprate unfairly penalizes visually correct predictions that differ only in $\LaTeX$ syntax styles. **Character Detection Matching (CDM)** offers a visual evaluation alternative. It renders both prediction and ground-truth into images, comparing them via character detection that considers spatial layout. This yields two metrics: the CDM Score, quantifying visual similarity, and Exprate@CDM, a visually-based accuracy measure. These CDM metrics provide evaluation robustness against variations in $\LaTeX$ syntax.

Training We fine-tune the Qwen2.5-VL-3B model on all datasets with three data-driven tasks simultaneously for a single epoch. Further details can be found in the Appendix B.4.

Baselines We compare our methods against four types of baselines: (i) Closed-source VLMs, including Gemini2.5-flash, etc; (ii) Open-source VLMs, such as the Qwen-2.5VL family. Both closed-source VLMs and open-source VLMs are evaluated under zero-shot inference settings; (iii) Previous SOTA methods on HMER and their data augmentation versions, for example, PosFormer and SSAN, etc; and (iv) A Vanilla(baseline) and a Uni-MuMER model, which is only fine-tuned on the CROHME training set to ensure a fair comparison with previous SOTA methods.

4.2 Main Results

Tab. 1 displays the strong performance of our Uni-MuMER and previous methods on the CROHME dataset. As CDM is a recently proposed metric, we reproduced all reported results involving CDM. For a fair comparison, the Vanilla (baseline) reported in Tab. 1 and Tab. 2 was trained exclusively on either CROHME or HME100K dataset. The Uni-MuMER variant incorporated training samples constructed from the corresponding images for three tasks. Further utilizing external data, UniMuMER[†] was trained on approximately 1.6M samples, comprising data from 386K images across the three tasks and all datasets. Due to space limitations, additional results are presented in Appendix D.

Comparison with Previous SOTA Methods Data augmentation has traditionally improved performance in lightweight models, yet inherent limitations still restrict overall effectiveness. Recently, VLPg [20] and HiE-VL [21], have integrated VLM into HMER. HiE-VL enhances performance through non-end-to-end training, architectural refinements, and extra data. However, both Uni-MuMER and Uni-MuMER[†] (use external data) maintain a notable performance advantage over these VLM-based methods. Notably, we also trained CoMER^{a†} [66] using the same external data (386K images) and observed a significant performance degradation. This finding indicates that the inherent limitations of lightweight models hinder their ability to effectively utilize large and diverse datasets.

Comparison with Large VLMs (Zero-Shot) Next, we investigate the zero-shot performance of both open-source (Qwen-2.5VL family) and closed-source large VLMs on CROHME. Although open-source models exhibit nontrivial results, they tend to lag behind previous SOTA methods when no fine-tuning is involved. Contrastingly, certain closed-source VLMs like Gemini2.5-flash exhibit remarkable accuracy, surpassing lightweight specialized models in a purely zero-shot setting. They still underperform relative to our proposed Uni-MuMER and Uni-MuMER[†] method.

Incremental Gains Building upon the Qwen2.5-VL-3B VLM base, our Vanilla (baseline) model already surpasses most published SOTA methods. After integrating three data-driven tasks, Tree-CoT, EDL, and SC, our method, Uni-MuMER[†] further improves both ExpRate and ExpRate@CDM, as shown in Tab. 1. For a fair comparison, both the Vanilla (baseline) and Uni-MuMER are trained solely on the CROHME training set, without using any external data. Introducing Upon incorporating external data along with these tasks (1.6M samples), our final model, Uni-MuMER[†], achieves even greater performance, notably reaching an ExpRate of 79.74% and ExpRate@CDM of 82.86% on the CROHME Average. Both Uni-MuMER and Uni-MuMER[†] significantly outperform both the Vanilla (baseline) model and prior methods, underscoring the vital role of data-driven tasks and diverse external data in enhancing model accuracy.

Performance on HME100K To focus the evaluation on mathematical expression recognition, following [16], we exclude all test instances containing CJK characters. Detailed impacts are discussed in Appendix C.5. Tab. 2 shows that open-source and closed-source VLMs exhibit modest performance, which is due to the challenging real-world conditions in HME100K dataset. In contrast, several prior lightweight specialized methods exclusively trained on HME100K, such as TAMER, achieve strong performance. Although HiE-VL utilizes external training data, it achieves limited results on HME100K, even underperforming compared to prior models. Our Uni-MuMER achieves SOTA performance. Additionally, by using external data, Uni-MuMER[†] further enhances its performance, clearly surpassing HiE-VL.

Table 2: **Performance on HME100K dataset (%)**, evaluated by Exp, @CDM, and CDM. The [†] denotes the use of external data.

Method	HME100K	
	Exp [†]	@CDM [†]
GPT-4o [41]	22.96	27.02
Gemini2.5-flash [11]	28.14	34.32
Doubao1.5-pro [45]	43.90	55.08
Qwen2.5-VL-3B [3]	44.42	47.41
Qwen2.5-VL-7B [3]	54.57	58.84
Qwen2.5-VL-72B [3]	49.59	55.09
DenseWAP [63]	61.85	-
CoMER [66]	68.12	70.36
TAMER [70]	69.50	71.30
HiE-VL [†] [21]	64.20	-
Vanilla (baseline)	70.15	71.80
Uni-MuMER	71.62 ^{+1.47}	73.40 ^{+1.60}
Uni-MuMER [†]	72.66^{+2.51}	74.30^{+2.50}

5 Analysis

5.1 Ablation Study

In this section, we present an ablation study to quantify the contribution of the three proposed tasks, Tree-CoT, EDL, and SC. As is shown in Tab. 3, each individual task enhances performance compared to the Vanilla(baseline). The Uni-MuMER combined integration of all tasks results in the best overall performance, while removing any module leads to noticeable performance degradation, highlighting their complementary roles and collective importance.

Table 3: **CROHME-Average ablation results (ExpRate %)**.

Tree-CoT	EDL	SC	ExpRate [†]
×	×	×	68.64
✓	×	×	70.85^{+2.21}
×	✓	×	70.30^{+1.66}
×	×	✓	69.86^{+1.22}
×	✓	✓	71.95^{+3.31}
✓	×	✓	71.06^{+2.42}
✓	✓	×	71.76^{+3.12}
✓	✓	✓	73.29^{+4.65}

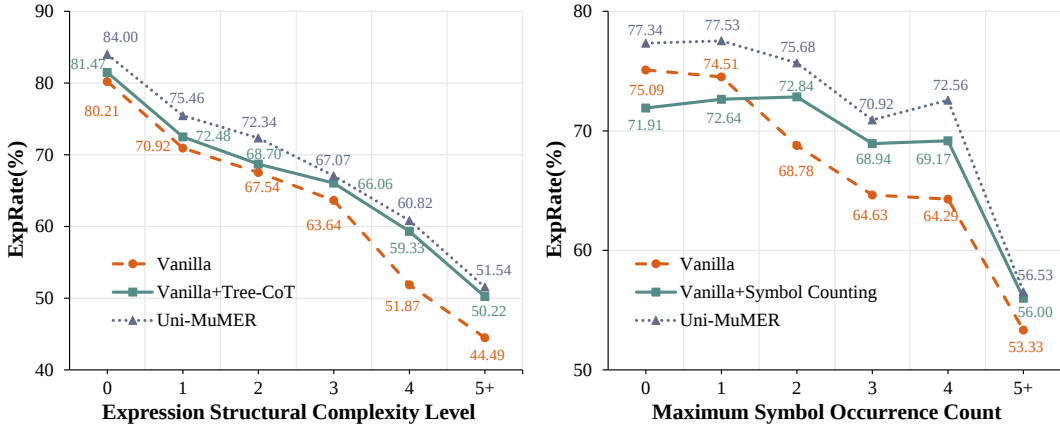
Tree-CoT enhances comprehension of structurally complex expressions Fig. 5a presents a comparative analysis of accuracy across models with varying structural complexities of expressions. The Vanilla model experiences a significant degradation in accuracy as structural complexity increases. Introducing the Tree-CoT strategy notably alleviates this performance decline. Specifically, for structurally complex expressions, Vanilla+Tree-CoT demonstrates an accuracy improvement of approximately 5–6% over the Vanilla baseline, closely matching the performance of the Full model. Conversely, Tree-CoT offers marginal improvements for simpler expressions, suggesting its primary utility is enhancing model robustness and accuracy when handling structurally intricate expressions.

EDL reduces confusion among visually similar characters Tab. 4 enumerates the top five most frequent alphabet-number confusion pairs on CROHME-Average for the Vanilla model and Vanilla+EDL. The Vanilla model frequently misclassifies pairs such as $2 \leftrightarrow z$, resulting in 5.24 misrecognitions per expression in all SUB1 cases. After applying EDL, the number of misclassifications decreases to 3.31. This implies that EDL effectively guides the model in distinguishing fine-grained differences among commonly confused symbols, enhancing symbol-level accuracy and overall recognition performance.

Table 4: **Top-5 alphabet-number confusions (SUB1) on CROHME-Average, comparing the Vanilla vs. Vanilla + EDL.**

Type	Top-5 misrecognition pair (%)					Σ
	$2 \leftrightarrow z$	$0 \leftrightarrow o$	$3 \leftrightarrow z$	$1 \leftrightarrow i$	$1 \leftrightarrow n$	
Vanilla	1.40	1.48	0.74	0.89	0.74	5.25
w/ EDL	1.23	1.16	0.15	0.62	0.15	3.31
$\Delta (\downarrow)$	-0.17	-0.32	-0.58	-0.27	-0.59	1.94

SC improves recognition consistency in long expressions Fig. 5b compares model accuracy with increasing symbol repetition. The Vanilla baseline accuracy deteriorates as repetition frequency rises, when symbols repeat five or more times. Incorporating the SC task substantially mitigates this decline, closely matching the Uni-MuMER model’s stable performance under high repetition conditions. However, for simpler expressions with minimal repetition, SC slightly reduces accuracy compared to the Vanilla baseline, suggesting it occasionally diverts attention from primary recognition tasks.



(a) Model performance relative to structural complexity. The Tree-CoT task notably reduces accuracy loss as structural complexity increases. (b) Model performance concerning symbol repetition count. Symbol Counting (SC) task mitigates accuracy degradation caused by frequent symbol repetitions.

Figure 5: **Model accuracy comparison under varying (a) structural complexities and (b) symbol repetition counts.** Both models significantly outperform the Vanilla baseline, approaching Uni-MuMER model performance in challenging scenarios.

5.2 Effects of Data Diversity

In this section, we examine the influence of incrementally scaling training datasets on model performance. As illustrated in Fig. 6, start with the CROHME dataset, which serves as the standard benchmark in HMER. We then sequentially incorporate increasingly diverse datasets, including the updated CROHME-2023 set, the real-scene HME100K dataset, the extensive handwritten corpus MathWriting, and finally the printed-expression Im2LaTeXv2 dataset. At each step, the model is retrained from scratch on the cumulative training data and evaluated across multiple distinct test sets.

To effectively account for variations in $\text{\LaTeX}$ notation styles, we employ ExpRate@CDM as the standardized metric for consistent performance comparison.

Benefits of Enhanced Data Diversity Each incremental expansion of the training sets with more varied handwritten expressions yields consistent gains on all evaluation sets, confirming that performance scales with data diversity.

Out-of-Domain Generalization Accuracy on printed expression images gradually improves with increased handwritten training data, reflecting partial transfer of structural knowledge for expression recognition. Nevertheless, its accuracy improves significantly after adding Im2LaTeXv2, reflecting the value of in-domain printed data for that target.

Leave-One-Out Ablation To further investigate the out-of-domain generalization performance, we conducted an additional ‘‘Leave-One-Out’’ experiment, training Uni-MuMER on N-1 datasets and evaluating it on the held-out N-th dataset. The results below present the expression recognition accuracy in each scenario. For comparison, we include the baseline CoMER and Doubao-1.5-pro under zero-shot settings.

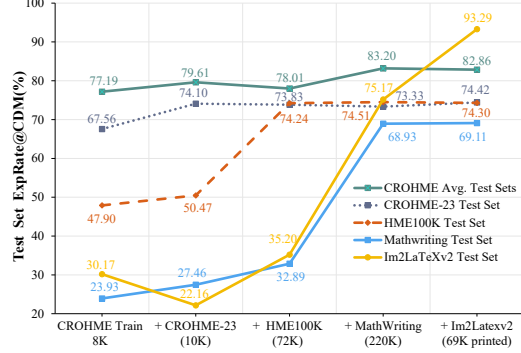


Figure 6: **Impact of incrementally training data scaling on ExpRate@CDM .** Starting from the CROHME training set, we incrementally incorporate four additional datasets. The model is re-trained from scratch and evaluated on five distinct test sets at each expansion step.

Table 5: **Leave-one-out ablation of training datasets. ($\text{ExpRate@CDM}\%$)** We train Uni-MuMER in six configurations: using all data (all five sources), and using all-minus-one for each major dataset. Results are reported on CROHME-Average, CROHME 2023, HME100K, MathWriting, and Im2LaTeXv2. For reference, we also show a specialized model (CoMER) trained with and without CROHME, and a closed-source VLM (Doubao-1.5-pro) for reference.

Model	CROHME-Avg	CROHME 2023	HME100K	MathWriting	Im2LaTeXv2
CoMER [†] (w/o CROHME)	45.99	58.28	39.49	23.23	40.39
CoMER [†]	52.29	59.91	44.63	28.45	53.37
Doubao-1.5-pro	65.77	58.18	55.08	26.34	27.66
Uni-MuMER [†] (w/o CROHME)	78.22	72.19	74.47	68.08	89.18
Uni-MuMER [†] (w/o CROHME-2023)	82.31	71.66	74.68	68.85	89.77
Uni-MuMER [†] (w/o HME100K)	82.30	72.01	52.38	68.94	91.19
Uni-MuMER [†] (w/o MathWriting)	79.58	73.51	74.29	32.55	93.31
Uni-MuMER [†] (w/o Im2LaTeXv2)	83.20	73.33	74.51	68.93	75.17
Uni-MuMER [†]	82.89	74.42	74.30	69.11	93.29

The leave-one-out evaluation confirms that our unified model can generalize to an unseen dataset reasonably well. When the unseen domain is very different, performance drops more markedly, which highlights an area for improvement. HME100K for real-life, low-quality images, MathWriting for densely structured expression, and Im2LaTeXv2 for printed expression images, each of which represents a markedly different and challenging domain.

Crucially, after introducing even modest amounts of domain-relevant data, our unified approach easily achieves top-tier performance. Thus, our results indicate that, compared to existing specialized models and general VLM baselines, our approach provides superior generalization with minimal domain-specific data requirements, highlighting a clear advantage and direction for future improvement.

5.3 Mixing General-Purpose VLM Data

From the perspective of general-purpose LLMs and VLMs, mixing data from different domains can yield substantial performance gains. Although our Uni-MuMER addresses previous limitations by introducing a unified solution, achieving super SOTA performance, it’s necessary to validate whether Uni-MuMER benefits from general domain data beyond HMER-only training and how Uni-MuMER-Data affects the model’s general capabilities. To this end, we introduce Uni-MuMER-LLAVA, fine-tuned using an equal 1:1 mixture of HMER and LLaVA-OneVision data, and compare its performance with Uni-MuMER across benchmarks assessing HMER-specific and general capabilities.

Table 6: **Comparison among Uni-MuMER, Uni-MuMER-LLAVA and Qwen2.5-VL on multi-modal reasoning benchmarks.** Results are reported on MMMU_val, Math-Vision, and MathVista_testmini, covering general multimodal understanding and mathematical reasoning tasks.

Model	MMMU_val	Math-Vision	MathVista_testmini
Uni-MuMER	47.89	24.01	47.8
Uni-MuMER-LLAVA	48.67	24.34	51.1
Qwen2.5-VL-3B-Instruct	52.78	21.38	47.9

Table 7: **Cross-benchmark comparison between Uni-MuMER and Uni-MuMER-LLAVA across benchmarks.** Results are reported on CROHME-Avg, CROHME-2023, HME100K, MathWriting, and Im2LaTeXv2.

Model	CROHME-Avg	CROHME-2023	HME100K	MathWriting	Im2LaTeXv2
Uni-MuMER	82.89	74.42	74.30	69.11	93.29
Uni-MuMER-LLAVA	81.99	69.10	73.48	70.48	93.44

As shown in Tab. 6 and Tab. 7, Uni-MuMER exhibits strong generalization capabilities despite being fine-tuned exclusively on HMER-specific data, achieving performance comparable to Qwen2.5VL-3B. Notably, Uni-MuMER-LLAVA further enhances overall performance, surpassing Uni-MuMER on general tasks like MMMU and Math, while incurring marginal performance decreases on specific HMER test sets. It’s promising to integrate our method and Uni-MuMER-Data to empower VLMs.

6 Discussion

For Small Models: Larger Model, Better Performance, Faster Speed Previous HMER models were lightweight and relied on task-specific architectural designs (e.g., tree-structured modules). Uni-MuMER builds upon both these task-specific insights and current VLM, taking a significant step forward with a unified and more powerful model. Benefiting from recent advances in efficient inference frameworks, Uni-MuMER not only achieves substantially improved performance but also delivers faster inference compared to smaller models. Further detail is provided in Appendix C.2.

For VLM: A Chain-of-Thought Perspective on Expression Recognition Unlike direct expression recognition, Uni-MuMER adopts a chain-of-thought perspective by formulating three tasks: expression tree construction, error detection and correction, and symbol counting. These tasks guide the model step by step toward the final prediction, enhancing both interpretability and generalization. This approach offers a novel direction for advancing expression recognition with current VLM.

Future Direction: Self-Correction for Expression Recognition Uni-MuMER provides a new perspective for enabling self-correction during inference in expression recognition. By integrating the Tree-CoT mechanism with an explicit error correction task, the model can exhibit self-corrective behavior within CoT process. Moreover, incorporating colloquial, human-like expressions (e.g., "wait, there’s a left parenthesis earlier—this is more likely a right parenthesis, not '2'") can further induce R1-like spontaneous self-correction [19], enhancing the upper limit of the model’s performance.

Limitations Despite the strong results, our method has several limitations. Due to computational resource constraints, a 400k synthetic subset of the MathWriting dataset was unused, potentially limiting the diversity benefits. Additionally, exploring optimal task mixture ratios within Uni-MuMER and optimal domain mixture ratios between HMER-specific data and general-domain data is crucial; however, these experiments were omitted due to significant resource constraints. Investigating these ratios could yield further enhance performance and generalizability.

7 Conclusion

This paper introduces Uni-MuMER, a unified multi-task fine-tuning framework for HMER, establishing a new paradigm by leveraging large-scale VLMs integrating multiple HMER tasks in a unified textual modality. Without architectural modifications, Uni-MuMER integrates Tree-aware Chain-of-Thought, Error-Driven Learning, and Symbol Counting tasks to jointly address structured spatial reasoning, error correction, and long-expression consistency in expression recognition, achieving super SOTA performance. This unified method provides enhanced accuracy, scalability, and faster inference speed, establishing a promising new research direction for future advancements in HMER.

8 Acknowledgement

This work is supported by the projects of Beijing Nova Interdisciplinary Program (20240484647) and National Natural Science Foundation of China (No. 62376012), which is also a research achievement of State Key Laboratory of Multimedia Information Processing and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

References

- [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. [2](#)
- [2] F. Álvaro, J.-A. Sánchez, and J.-M. Benedí. Recognition of on-line handwritten mathematical expressions using 2d stochastic context-free grammars and hidden markov models. *Pattern Recognition Letters*, 35:58–67, 2014. [2](#)
- [3] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-VL technical report, Feb. 2025. [2](#), [3](#), [6](#), [7](#)
- [4] X. Bian, B. Qin, X. Xin, J. Li, X. Su, and Y. Wang. Handwritten mathematical expression recognition via attention aggregation based bi-directional mutual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 113–121, 2022. [2](#), [3](#), [6](#)
- [5] L. Blecher, G. Cucurull, T. Scialom, and R. Stojnic. Nougat: Neural optical understanding for academic documents, 2023. [3](#)
- [6] K.-F. Chan and D.-Y. Yeung. Mathematical expression recognition: a survey. *International Journal on Document Analysis and Recognition*, 3:3–15, 2000. [2](#)
- [7] K. Chen, H. Shen, L. Zhong, and M. Chen. Mmhmmer: Multi-viewer and multi-task for handwritten mathematical expression recognition. *arXiv preprint arXiv:2502.05557*, 2025. [2](#)
- [8] X. Chen, X. Wang, S. Changpinyo, A. J. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, A. Kolesnikov, J. Puigcerver, N. Ding, K. Rong, H. Akbari, G. Mishra, L. Xue, A. V. Thapliyal, J. Bradbury, W. Kuo, M. Seyedhosseini, C. Jia, B. K. Ayan, C. R. Ruiz, A. P. Steiner, A. Angelova, X. Zhai, N. Houlsby, and R. Soricut. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*, Sept. 2022. [2](#)
- [9] Z. Chen, J. Han, C. Yang, and Y. Zhou. Language model is suitable for correction of handwritten mathematical expressions recognition. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4057–4068, Singapore, Dec. 2023. Association for Computational Linguistics. [3](#), [5](#)
- [10] P. A. Chou. Recognition of equations using a two-dimensional stochastic context-free grammar. In *Visual Communications and Image Processing IV*, volume 1199, pages 852–865. SPIE, 1989. [2](#)
- [11] G. DeepMind. Gemini 2.5 Flash: Enhanced multimodal model. <https://blog.google/products/gemini/gemini-2-5-flash-preview/>, Apr. 2025. Public preview announced 17 Apr 2025. [2](#), [7](#)
- [12] Y. Deng, A. Kanervisto, J. Ling, and A. M. Rush. Image-to-markup generation with coarse-to-fine attention. In *International Conference on Machine Learning*, pages 980–989. PMLR, 2017. [3](#), [5](#)
- [13] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, et al. Palm-e: An embodied multimodal language model, Mar. 2023. [32](#)
- [14] A. Fadeeva, P. Schlattner, A. Maksai, M. Collier, E. Kokiopoulou, J. Berent, and C. Musat. Representing online handwriting for recognition in large vision-language models. *arXiv preprint arXiv:2402.15307*, 2024. [32](#)
- [15] S. A. Firdaus and K. Vaidehi. Handwritten mathematical symbol recognition using machine learning techniques: Review. In S. C. Satapathy, K. S. Raju, K. Shyamala, D. R. Krishna, and M. N. Favorskaya, editors, *Advances in Decision Sciences, Image Processing, Security and Computer Vision*, volume 4, pages 658–671. Springer International Publishing, Cham, 2020. [1](#)
- [16] Y. Gao, W. Wang, and H. Ney. Uniblock: Scoring and filtering corpus with unicode block information. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1324–1329, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. [7](#)
- [17] P. Gervais, A. Fadeeva, and A. Maksai. Mathwriting: A dataset for handwritten mathematical expression recognition. *arXiv preprint arXiv:2404.10690*, 2024. [5](#), [25](#)
- [18] T. Guan, C. Lin, W. Shen, and X. Yang. Posformer: recognizing complex handwritten mathematical expression with position forest transformer. In *European Conference on Computer Vision*, pages 130–147. Springer, 2024. [2](#), [3](#), [6](#), [26](#)

- [19] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 10
- [20] H.-Y. Guo, C. Wang, F. Yin, X.-H. Li, and C.-L. Liu. Vision–language pre-training for graph-based handwritten mathematical expression recognition. *Pattern Recognition*, 162:111346, 2025. 3, 6, 7
- [21] H.-Y. Guo, F. Yin, J. Xu, and C.-L. Liu. HiE-VL: A large vision-language model with hierarchical adapter for handwritten mathematical expression recognition. In *Icassp 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (Icassp)*, pages 1–5, Hyderabad, India, Apr. 2025. IEEE. 3, 6, 7, 25
- [22] J. Ha, R. M. Haralick, and I. T. Phillips. Understanding mathematical expressions from document images. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 2, pages 956–959. IEEE, 1995. 2
- [23] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982. 1
- [24] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM international conference on multimedia*, pages 4083–4091, 2022. 3
- [25] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2
- [26] N. Jiang, S. Liang, C. Wang, J. Wang, and L. Tan. Latte: Improving latex recognition for tables and formulae with iterative refinement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4030–4038, 2025. 26
- [27] G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022. 3
- [28] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023. 2
- [29] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023. 2
- [30] B. Li, Y. Yuan, D. Liang, X. Liu, Z. Ji, J. Bai, W. Liu, and X. Bai. When counting meets hmer: counting-aware network for handwritten mathematical expression recognition. In *European Conference on Computer Vision*, pages 197–214. Springer, 2022. 2, 3, 5, 6
- [31] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2
- [32] J. Li, D. Li, C. Xiong, and S. Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900. PMLR, June 2022. 2
- [33] Z. Li, B. Yang, Q. Liu, Z. Ma, S. Zhang, J. Yang, Y. Sun, Y. Liu, and X. Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26763–26773, 2024. 3
- [34] Y. Liu, Z. Li, M. Huang, B. Yang, W. Yu, C. Li, X.-C. Yin, C.-L. Liu, L. Jin, and X. Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024. 2
- [35] Y. Liu, B. Yang, Q. Liu, Z. Li, Z. Ma, S. Zhang, and X. Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024. 3
- [36] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegrefe, U. Alon, N. Dziri, S. Prabhume, Y. Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023. 2
- [37] M. Mahdavi, R. Zanibbi, H. Mouchere, C. Viard-Gaudin, and U. Garain. ICDAR 2019 CROHME+ TFD: Competition on recognition of handwritten mathematical expressions and typeset formula detection. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1533–1538. IEEE, 2019. 2, 5
- [38] H. Mouchere, C. Viard-Gaudin, R. Zanibbi, and U. Garain. Icfhr 2014 competition on recognition of on-line handwritten mathematical expressions (crohme 2014). In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 791–796, 2014. 2, 5
- [39] H. Mouchère, C. Viard-Gaudin, R. Zanibbi, and U. Garain. Icfhr2016 crohme: Competition on recognition of online handwritten mathematical expressions. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 607–612, 2016. 2, 5

- [40] O. Nath, H. Bathina, M. S. U. R. Khan, and M. M. Khapra. Can vision-language models evaluate handwritten math? *arXiv preprint arXiv:2501.07244*, 2025. 3
- [41] OpenAI. GPT-4o: Multimodal generative transformer. <https://openai.com/index/hello-gpt-4o/>, 2024. Flagship model unveiled 13 May 2024. 2, 6, 7
- [42] S. Peng, L. Gao, K. Yuan, and Z. Tang. Image to latex with graph neural network for mathematical formula recognition. In *Document Analysis and Recognition-ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 648–663. Springer, 2021. 4
- [43] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, July 2021. 3
- [44] F. M. Schmitt-Koopmann, E. M. Huang, H.-P. Hutter, T. Stadelmann, and A. Darvishy. MathNet: A data-centric approach for printed mathematical expression recognition. *IEEE Access*, 12:76963–76974, 2024. 5, 25, 34
- [45] B. D. Team. Doubao 1.5 Vision Pro: Multimodal large language model. https://seed.bytedance.com/en/special/doubao_1_5_pro, Jan. 2025. Official release reported 22 Jan 2025. 2, 6, 7
- [46] T.-N. Truong, C. T. Nguyen, R. Zanibbi, H. Mouchère, and M. Nakagawa. A survey on handwritten mathematical expression recognition: The rise of encoder-decoder and GNN models. *Pattern Recognition*, 153:110531, Sept. 2024. 2
- [47] H. M. Twaakyondo and M. Okamoto. Structure analysis and recognition of mathematical expressions. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 430–437. IEEE, 1995. 2
- [48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 3
- [49] B. Wang, Z. Gu, G. Liang, C. Xu, B. Zhang, B. Shi, and C. He. Unimernet: A universal network for real-world mathematical expression recognition. *arXiv preprint arXiv:2404.15254*, 2024. 3
- [50] B. Wang, F. Wu, L. Ouyang, Z. Gu, R. Zhang, R. Xia, B. Shi, B. Zhang, and C. He. Image over text: Transforming formula recognition evaluation with character detection matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19681–19690, 2025. 35
- [51] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2
- [52] C. Wu, J. Du, Y. Li, J. Zhang, C. Yang, B. Ren, and Y. Hu. TDv2: A Novel Tree-Structured Decoder for Offline Mathematical Expression Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2694–2702, June 2022. 2, 4
- [53] J.-W. Wu, F. Yin, Y.-M. Zhang, X.-Y. Zhang, and C.-L. Liu. Image-to-markup generation via paired adversarial learning. In *Machine learning and knowledge discovery in databases: European conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pages 18–34. Springer, 2019. 2, 3
- [54] J.-W. Wu, F. Yin, Y.-M. Zhang, X.-Y. Zhang, and C.-L. Liu. Handwritten mathematical expression recognition via paired adversarial learning. *International Journal of Computer Vision*, 128:2386–2401, 2020. 2, 3
- [55] Y. Xie and H. Mouchère. Tst: Tree structured transformer for handwritten mathematical expression recognition. In *International Conference on Document Analysis and Recognition*, pages 236–252. Springer, 2025. 33
- [56] Y. Xie, H. Mouchère, F. Simistira Liwicki, S. Rakesh, R. Saini, M. Nakagawa, C. T. Nguyen, and T.-N. Truong. Icdar 2023 crohme: Competition on recognition of handwritten mathematical expressions. In *International Conference on Document Analysis and Recognition*, pages 553–565. Springer, 2023. 5, 25
- [57] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023. 2
- [58] Y. Yuan, X. Liu, W. Dikubab, H. Liu, Z. Ji, Z. Wu, and X. Bai. Syntax-aware network for handwritten mathematical expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4553–4562, 2022. 3
- [59] Y. Yuan, X. Liu, W. Dikubab, H. Liu, Z. Ji, Z. Wu, and X. Bai. Syntax-aware network for handwritten mathematical expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4553–4562, 2022. 5
- [60] R. Zanibbi, D. Blostein, and J. R. Cordy. Recognizing mathematical expressions using tree transformation. *IEEE Transactions on pattern analysis and machine intelligence*, 24(11):1455–1467, 2002. 2
- [61] R. Zanibbi, H. Mouchere, and C. Viard-Gaudin. Evaluating structural pattern recognition for handwritten math via primitive label graphs. In *Document Recognition and Retrieval XX*, volume 8658, pages 411–421. SPIE, 2013. 4

- [62] H. Zhang, X. Su, X. Zhou, and G. Gao. SSAN: A symbol spatial-aware network for handwritten mathematical expression recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(21):22398–22406, Apr. 2025. 2, 3, 6
- [63] J. Zhang, J. Du, and L. Dai. Multi-scale attention with dense encoder for handwritten mathematical expression recognition. In *2018 24th international conference on pattern recognition (ICPR)*, pages 2245–2250, 2018. 2, 3, 6, 7
- [64] J. Zhang, J. Du, Y. Yang, Y.-Z. Song, S. Wei, and L. Dai. A tree-structured decoder for image-to-markup generation. In *ICML*, page In Press, 2020. 2, 3, 4
- [65] J. Zhang, J. Du, S. Zhang, D. Liu, Y. Hu, J. Hu, S. Wei, and L. Dai. Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition. *Pattern Recognition*, 71:196–206, 2017. 2, 3
- [66] W. Zhao and L. Gao. Comer: Modeling coverage for transformer-based handwritten mathematical expression recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 392–408. Springer, 2022. 2, 3, 6, 7
- [67] W. Zhao, L. Gao, Z. Yan, S. Peng, L. Du, and Z. Zhang. Handwritten mathematical expression recognition with bidirectionally trained transformer. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 570–584. Springer, 2021. 2, 3, 6
- [68] D. Zhelezniakov, V. Zaytsev, and O. Radyvonenko. Online handwritten mathematical expression recognition and applications: A survey. *IEEE Access*, 9:38352–38373, 2021. 1, 2
- [69] J. Zhu, L. Gao, and W. Zhao. Ical: Implicit character-aided learning for enhanced handwritten mathematical expression recognition. *arXiv preprint arXiv:2405.09032*, 2024. 3
- [70] J. Zhu, W. Zhao, Y. Li, X. Hu, and L. Gao. Tamer: Tree-aware transformer for handwritten mathematical expression recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10950–10958, 2025. 2, 3, 4, 6, 7

9 Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: The abstract and Introduction enumerate three contributions, which are unified multi-task fine-tuning, three data-driven tasks, and new SOTA results, all verified experimentally in Tab. 1 and Tab. 2.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: A dedicated “Limitations” paragraph in Sec. 6, which discusses domain-specific scope, compute constraints, and unused synthetic data.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The work is empirical and introduces no formal theorems or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have taken steps to ensure the reproducibility of our result. Sec. 4.1 detail datasets, and Appendix introduce how to preprocess datasets and hyper-parameters.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We have not included the actual code or data with the submission. However, we provide thorough instructions and descriptions in the paper and the appendix to reproduce all main experiments. We will publicly release our code, model checkpoints, and detailed usage instructions in the future.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All data are public, and provide the split training sets and test sets. Hyperparameters and optimizer are introduced in Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We did not report explicit error bars, confidence intervals, or statistical significance tests for our experimental results. Due to the high cost of fine-tuning large VLMs, each configuration was run once; we therefore report point estimates without confidence intervals.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix introduces training used 4 x A800-80 GB GPUs, 80 GB VRAM each, single-epoch on 1.6M train samples, which cost 96 hours, and describes that training was run for one epoch on a combined dataset of around 1.6M images.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We have read the NeurIPS Code of Ethics and confirm that our research adheres to it. Our work involves improving a handwriting recognition model using public, licensable datasets and does not present unethical content or methodology.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: We did not include a dedicated broader societal impacts section because our work focuses on a specialized technical application with limited direct societal implications. The primary impact of our research is positive, improving tools for digitizing and understanding handwritten mathematics, which can benefit education and scientific document processing.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The released fine-tuned model targets mathematical OCR and poses minimal dual-use risk; no additional safeguards are required.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Sec. 4.1 cites CROHME, HME100K, MathWriting, and Im2LaTeXv2 with their original papers and open licences; code bases (e.g., Qwen2.5-VL) are Apache-2.0.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not introduce new proprietary assets in this paper. Our contributions lie in the unified fine-tuning approach and tasks, applied to existing datasets and an existing open-source model.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This work involves only publicly available handwriting datasets; no human subjects were recruited.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: IRB approval is not applicable because the study does not collect new human-subject data.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We clearly describe our use of a large-scale pre-trained model in the paper. In particular, Sec. 3, explains that our method is built on the Qwen2.5-VL-3B vision-language model, which we fine-tune for HMER. The role of this LLM/VLM in our approach is explicitly documented as a core component of the methodology.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

Appendix

Table of Contents

A Prompt Design for Multi-Task Training	23
B Expreimental Details	25
B.1 Uni-MuMER-Data Preprocess	25
B.2 Evaluation Datasets and Metrics	26
B.3 Evaluation Prompting Protocol	27
B.4 Training Detail	27
C Extended Analysis	28
C.1 Effects of Data Diversity Compared with Specialized HMER Model	28
C.2 Inference Efficiency	28
C.3 Alternative explorations format	28
C.4 Hallucination for HMER	30
C.5 CJK influence for HMER	31
D More Experimental Results	32
D.1 Results on Handwritten Mathematical Expression	32
D.2 Results on Printed Mathematical Expressions	33
E Case Study	35

A Prompt Design for Multi-Task Training

We design distinct prompts for each of the four training tasks in Uni-MuMER. Each prompt guides the model on a specific sub-task, and the model’s expected output format is tailored to that task. Each prompt consists of a fixed system prompt: *You are a helpful assistant*. Below we list the prompts used for Vanilla HMER, Tree-CoT, Error-Driven Learning, and Symbol Counting (SC), along with example prompts and the form of expected outputs. It is a detailed expansion of the content in Fig. 2.

Example A.1: Prompt for Vanilla HMER

Prompt: I have an image of a handwritten mathematical expression. Please write out the expression of the formula in the image using $\text{\LaTeX}$ format.

Answer:

$$\left| \frac{ax_0 + by_0 + c}{\sqrt{a^2 + b^2}} \right|$$

Example A.2: Prompt for Tree-CoT

Prompt: I have an image of a handwritten mathematical expression. Please generate the **abstract syntax tree (AST)** of the formula in the image, and then provide its corresponding $\text{\LaTeX}$ format.

Answer:

```
|
\frac (right)
  a (above)
    x (right)
      0 (sub)
    + (right)
      b (right)
        y (right)
          0 (sub)
        + (right)
          c (right)
        \sqrt (below)
          a (above)
            2 (sup)
          + (right)
            b (right)
              2 (sup)
        | (right)
\frac { a x _ { 0 } + b y _ { 0 } + c } { \sqrt { a ^ { 2 } + b ^ { 2 } } } |
```

Example A.3: Prompt for Error Detection

Prompt: I have an image of a handwritten mathematical expression and its OCR recognition result. Please help me to **detect** possible errors in the recognition result and mark the places where errors occur with `<error_start>` `<error_end>` and `<deleted>`.

erroneous formula: $\left| \frac{ax_0 + by_0 + c}{\sqrt{d^2 + b^2}} \right|$

Marked formula:

Answer:

$\left| \frac{ax_0 + by_0 + c}{\sqrt{\text{\textless error_start\textgreater} d \text{\textless error_end\textgreater}^2 + b^2}} \right|$

Example A.4: Prompt for Error Correction

Prompt: I have an image of a handwritten mathematical expression and a predicted formula with error tags, **correct** the formula by modifying the parts marked with `<error_start>` and `<error_end>` and inserting content where `<deleted>` are present. Output the modifications in a single string using the following format: REPLACE:old -> new for errors to be replaced. INSERT:content after the position for missing content. DELETE:to_delete for parts to be removed.

Marked formula: $\left| \frac{ax_0 + by_0 + c}{\sqrt{\text{\textless error_start\textgreater} d \text{\textless error_end\textgreater}^2 + b^2}} \right|$

Correction log:

Answer:

REPLACE:d -> a

$\left| \frac{ax_0 + by_0 + c}{\sqrt{a^2 + b^2}} \right|$

Example A.5: Prompt for Symbol Counting

Prompt: I have an image of a handwritten mathematical expression. Please **identify and count** each distinct visible mathematical symbol in the image, and then provide its corresponding $\LaTeX$ format.

Answer:

l: 2, \frac: 1, a: 2, x: 1, _: 2, 0: 2, +: 3, b: 2, y: 1, c: 1, \sqrt: 1, ^: 2, 2: 2
|\frac { a x _ { 0 } + b y _ { 0 } + c } { \sqrt { a ^ { 2 } + b ^ { 2 } } } |

These carefully designed prompts ensure that during fine-tuning, Uni-MuMER learns not only to recognize the expressions but also to reason about structure, identify and correct errors, and pay attention to every symbol. By training on this multi-task prompt set, the model acquires a richer understanding of handwritten math, which cumulatively improves its accuracy and robustness.

B Experimental Details

B.1 Uni-MuMER-Data Preprocess

We perform comprehensive preprocessing on all datasets to prepare the training and evaluation data for Uni-MuMER. This involves cleaning noisy data, tokenizing $\LaTeX$ strings into standardized tokens, and normalizing different $\LaTeX$ notations to a consistent style. $\LaTeX$ expressions can often be written in syntactically different ways while conveying the same meaning, so unifying these representations is crucial for effective cross-dataset training.

Data Cleaning First, we exclude all instances containing CJK characters. For example, the HME100K dataset occasionally contains Chinese characters in annotations; these were removed to avoid confusing the model. Additionally, we remove characters or symbols that lack semantic value and negatively impact recognition accuracy, such as `\underline{\quad}` and empty brace pairs (`{}`), which do not contribute to the math content.

Data Tokenization Since the $\LaTeX$ expressions provided by datasets like CROHME 2023 [56] and Mathwriting [17] are not inherently tokenized. To train a sequence model effectively, we convert each expression into a standardized sequence of $\LaTeX$ tokens. Specifically, we adopt the tokenization procedure from the `preprocess_formulas` script of the open-source `im2markup` repository, leveraging the KaTeX JavaScript library to perform semantic-level tokenization. During tokenization, we also remove any syntactically invalid expressions that fail to parse. Finally, the expression `\frac{a^2}{2}` is tokenized as `\frac a ^ 2 2`, where each $\LaTeX$ token is separated by a space.

Data Normalization After tokenization, we normalize the ground-truth $\LaTeX$ sequences across all training datasets to ensure a consistent format. Different datasets might represent the same mathematical concept in slightly different $\LaTeX$ forms. Such variations include notation differences (e.g., using `\leq` and `\le`), optional braces (`a^2` and `a^{2}`), or stylistic markup (e.g., using `\textbf{a}` vs. just `a`), which is typically not visible in handwritten notes. If left unnormalized, these variations could confuse the model or reduce its ability to generalize between datasets. We therefore systematically unify these to a single canonical form. For example, the expression `\frac{a^2}{2}` is normalized as `\frac { a ^ { 2 } } { 2 }`. Detailed normalization transformations are summarized in Tab. 8, illustrating the specific $\LaTeX$ syntax standardization rules we applied, extending and optimizing the practices initially established by MathNet [44].

Table 8: **Examples of $\LaTeX$ math normalization.** Comparison of original math expressions, normalized versions, and their rendered figures. Following the configurations of HiE-VL [21] and MathNet [44], we apply these transformations to make training more consistent.

Transform	Original	Normalized	Figure
Brace	<code>H^I</code>	<code>H^{I}</code>	H^I
SubSup	<code>\int^{b}_{a}</code>	<code>\int_{a}^{b}</code>	$\int_a^b$
Root	<code>\sqrt[2]{a}</code>	<code>\sqrt{2}{a}</code>	$\sqrt[2]{a}$
Stylized	<code>\textbf{a}</code>	<code>a</code>	a
...	...	...	...

Table 9: **Dataset filtering summary and multi-task sample counts.**

Category	CROHME	CROHME2023	HME100K	Mathwriting	Im2LaTeXV2
raw	8834	12204	74502	229864	74244
rendering errors	0	587	1166	624	1153
normalization step	0	1	590	4749	3954
brackets invalid	0	0	0	65	82
rendering errors	0	0	0	293	1
Total (after deletion)	8834	11616	72746	224133	69054
Tree-CoT	8834	11616	72746	224133	69054
Error Detection	9010	6287	49959	125316	8828
Error Correction	11434	8860	58884	166387	17047
Symbol Counting	8834	11616	72746	224133	69054
Total (with four tasks)	38112	38379	254335	739969	163983

Dataset Filtering Each dataset contributed a different number of valid expressions after preprocessing. Tab. 9 summarizes the filtering outcomes for the major datasets we used: CROHME, CROHME2023, HME100K, MathWriting, and Im2LaTeXv2. For each dataset, we list the original number of expressions and how many were removed for various reasons, such as rendering errors or invalid syntax detected during normalization. The final row gives the count of expressions remaining in each dataset after all filtering steps.

Task Data Construction With the cleaned datasets in hand, we construct the multi-task training samples. Each image can yield multiple training samples, one for each task. The lower part of Tab. 9 lists the number of training samples per task for each dataset. For the Tree-CoT, Symbol Counting, and Vanilla HMER tasks, we typically generate one sample per image, so their counts per dataset are equal to the number of remaining images. For the Error Detection and Error Correction subtasks, we collect all error corpus generated by each dataset. Further inspired by recent work on incremental refinement LATTE [26], our error-correction task is specifically constrained to correcting one error per round, which is more effective than inputting the whole incorrect expression. In total, across all datasets, our training set comprises roughly 1.6 million samples derived from about 392k unique images, spanning the three tasks.

B.2 Evaluation Datasets and Metrics

Datasets The **MNE (Multi-level Nested Expression)** dataset is specifically designed to evaluate models on recognizing complex handwritten mathematical expressions [18]. The dataset is organized into subsets designated as N1, N2, and N3, based on the structural complexity of the expressions, determined by the maximum nested levels within their substructures. Fig. 7 illustrates various mathematical expressions along with their corresponding tree-based structural complexities. For instance, an expression such as $\frac{a^2+1}{b}$ possesses a structural complexity level of 2, indicating two hierarchical nesting layers.

Evaluation Metric The **Character Error Rate (CER)** CER is computed via Levenshtein distance at the Unicode character level, normalized by ground-truth length. It is widely used in online handwriting datasets (e.g., *MathWriting*) to detect localization errors, brace imbalances, and missing superscripts.

EditScore Following *MathNet*, we adopt EditScore, defined as $1 - \frac{\text{Edit distance}}{\max(|\hat{y}|, |y|)}$. EditScore, the sequence-level counterpart to CER, strongly correlates with human judgments in printed Mathematical Expression Recognition (MER) tasks and tolerates minor reordering and length variations while preserving semantics. **BLEU-4** For consistency and comparability with existing corpus in the im2latexv2 domain, we incorporate the BLEU-4 metric, calculated based on 4-gram overlap after $\LaTeX$ normalization, which effectively quantifies n-gram overlaps and remains a valuable metric.

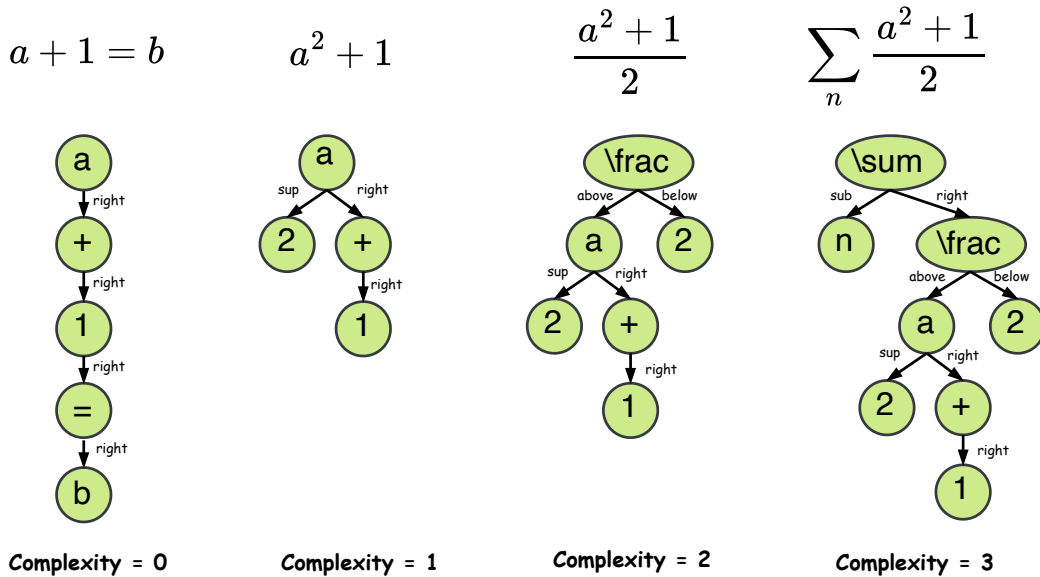


Figure 7: Examples of structural complexity for different expressions

B.3 Evaluation Prompting Protocol

When evaluating the fine-tuned Uni-MuMER model and comparing it to various baselines, we standardize the prompting method to ensure a fair and consistent evaluation across models: Below, we describe the exact prompts and settings used in different evaluation scenarios.

Uni-MuMER (Ours) We used the Vanilla HMER prompt to elicit the final $\LaTeX$ output when evaluating our fine-tuned model on the test sets. Instead of explicitly prompting the model to show its chain-of-thought or to count symbols during this evaluation phase, we directly request the final $\LaTeX$ output. The rationale is that our multi-task training has already imbued the model with better internal reasoning and understanding; at test time, we just want the final answer, not the intermediate steps. Thus, for each test image, we simply provided the vanilla prompt like: *"I have an image of a handwritten mathematical expression. Please write out the expression of the formula in the image using $\LaTeX$ format."* to the model, along with the handwritten expression image.

Baseline VLMs (Open-Source and Closed-Source) We evaluate several open-source VLMs and closed-source VLMs in a zero-shot setting on the CROHME and HME100K datasets. However, one challenge was that some general models might produce additional explanatory text or not output the expression alone. To mitigate this, we minimally revise the prompt to ask the model to output the $\LaTeX$ expression in the `\boxed{ }`. The prompt is *"I have an image of a handwritten mathematical expression. Please write out the expression of the formula in the image using $\LaTeX$ format and place it inside `\boxed{ }`"*. We then post-processed the output by removing the `\boxed{ }` wrapper to retrieve the raw $\LaTeX$.

Output Normalization After obtaining outputs from all VLMs in a zero-shot setting, we apply the same cleaning and tokenization steps as used in training to each output before evaluation. This is important so that metrics like ExpRate or CER treat all outputs on equal footing. In particular, for the HME100K test set, if a model produced any non- $\LaTeX$ characters or CJK text, we dropped those outputs entirely.

Reproducibility of Baselines To ensure transparency and reproducibility, we explicitly detail the exact versions of the models evaluated in this study. Closed-source models were accessed through their respective APIs or publicly available releases at specific reference dates: Gemini 2.5-flash (version: *gemini-2.5-flash-preview-04-17*), Doubao-1.5-pro (version: *doubao-1.5-vision-pro-250328*), and GPT-4o (version: *gpt-4o-2024-11-20*). For open-source models such as Qwen2.5-VL (3B, 7B, and 72B), we adopted the latest publicly accessible checkpoints available at the time of experimentation. All models underwent evaluation on identical test datasets, specifically the CROHME 2014, 2016, and 2019 datasets, as well as the HME100K test set. We meticulously ensured consistent preprocessing procedures and uniformly applied evaluation metrics across all model outputs.

By enforcing this consistent prompting and processing protocol, we aimed to make the comparison as fair as possible. All outputs were judged by exactly the same criteria. This way, the results reflect the true capability of each model on HMER, without artifacts from prompting differences.

B.4 Training Detail

Implementation Details We fine-tune Qwen2.5-VL-3B using a full-parameter update. Besides, we employ a cosine decay schedule to gradually reduce the learning rate to 10% of its initial value (1×10^{-5}) by the end of training. Training is conducted for a single epoch using all tasks and mixed dataset batches. Due to memory constraints, we adopt mixed-precision training, which enables us to fit a batch size of 512 within approximately 80 GB of VRAM per GPU (8xA800). This large batch size was specifically chosen to accommodate the diversity of tasks and stabilize the training process.

Training time We trained for a single epoch over the combined multi-task dataset. One epoch in this context amounted to roughly 1.6 million training samples, since each image contributes multiple task-specific samples. Iterating over 1.6M samples with batch 512 yields about 3152 optimization steps. In practice, this single epoch took 20 hours to complete on the $8 \times$ A800 setup. Although this runtime slightly exceeds that of training lightweight specialized models from scratch, our approach achieves convergence to a high-performance level within just a single epoch.

C Extended Analysis

C.1 Effects of Data Diversity Compared with Specialized HMER Model

In this subsection, we examine how incrementally scaling training data impacts ExpRate@CDM . As noted in Sec. 1, previous specialized models typically train on individual datasets, limited by the available training data and rarely assessed using visual metrics. Building upon the analysis in Sec. 5.2, we start with the CROHME dataset alone and then progressively add more training data from four additional sources. At each stage, we train three models from scratch on the combined data up to that point: (i) a lightweight specialized model, CoMER, which is a strong transformer-based architecture representative of specialized models; (ii) our vanilla baseline, which is only fine-tuned on 392k unique images; (iii) Uni-MuMER which is multi-task fine-tuned on the 1.6M samples).

The results, summarized in the Tab. 10, indicate that CoMER initially benefits from increased data diversity but subsequently exhibits significant performance degradation, settling at modest performance. Conversely, the vanilla baseline and Uni-MuMER show a much more robust improvement curve. Finally, Uni-MuMER consistently outperforms the Vanilla model at every stage and shows continuous gains. This demonstrates that the multi-task fine-tuning is effective at harnessing diverse data. Furthermore, specialized models like CoMER typically require extensive hyperparameter tuning, including adjustments to learning rates and training epochs. In contrast, Uni-MuMER attains superior performance after only a single training epoch without additional optimization efforts. This underscores Uni-MuMER’s robustness and efficiency.

Table 10: **Effect of progressively adding training datasets on CROHME-Avg ExpRate@CDM .** At each step, models are trained from scratch on the cumulative data. Uni-MuMER improves monotonically and consistently outperforms the Vanilla (single-task) and CoMER baselines.

Model	CROHME Train	+ CROHME-23	+ HME100K	+ MathWriting	+ Im2Latexv2
CoMER	58.14	60.08	63.10	52.17	52.29
Vanilla (baseline)	71.14	75.07	75.13	80.87	80.35
Uni-MuMER	75.36	76.27	78.01	83.20	82.89
Uni-MuMER-7B	77.45	81.17	80.35	83.35	83.34

C.2 Inference Efficiency

We evaluated the inference speed and model size of our proposed Uni-MuMER model against the baseline methods CoMER and TAMER, utilizing a single NVIDIA A800 GPU. To ensure a consistent and fair comparison, we used the HME100K test dataset, comprising 24,607 real-scene handwritten mathematical expression images. Additionally, we leveraged vLLM, an optimized transformer-based inference engine known for accelerating decoding processes through efficient batch processing and caching mechanisms. The comparative efficiency metrics are summarized in Table 11. As demonstrated, Uni-MuMER significantly surpasses previous HMER models in terms of inference speed, achieving an FPS approximately eightfold higher than existing baselines under identical hardware conditions. As a result, Uni-MuMER is not only accurate but also efficient enough for practical use.

Table 11: **Comparison of methods on efficiency metrics.** Throughput is reported in frames per second (FPS, higher is better) and model capacity in trainable parameters.

Method	FPS $\uparrow$	Parameters
CoMER	5.23	6.39 M
TAMER	4.50	8.23 M
Uni-MuMER[†]	43.40	3.75 B

C.3 Alternative explorations format

In developing the Tree-CoT and Symbol Counting tasks, we explored various output representation formats and their impact on model performance. Our goal was to find a format that is expressive enough to capture structure/counts but concise enough for the model to learn effectively. Here we report experiments with alternative formats for the expression tree serialization in Tree-CoT, as well as for the symbol counting task. We compare each alternative to the format ultimately used in Uni-MuMER (referred to as “our format”). All experiments in this section were conducted using

only CROHME data for training (no external data), so that differences reflect format efficacy rather than additional data. We evaluate both in a Vanilla+X setting and in the full Uni-MuMER setting (all three tasks, but substituting the alternative format for the relevant task) to assess each format’s influence. Below, we outline each representation with an example.

Example C.1: S-expression format for Tree CoT

$\backslash\mu_{\{pj\}}$

Example C.2: Json AST format for Tree CoT

`['type': 'supsub', 'value': 'base': 'type': 'mathord', 'value': '\mu', 'sub': 'type': 'ordgroup', 'value': ['type': 'mathord', 'value': 'p', 'type': 'mathord', 'value': 'j']]`

Example C.3: Simplified Json AST format for Tree CoT

`[supsub: base: '\mu', sub: ordgroup: ['p', 'j']]`

The S-expression is merely the formula itself and adds no metadata. Its nested structure also hurts readability and organization. The JSON AST is excessively complex—it reflects JavaScript’s internal LaTeX format—and includes substantial redundancy. The “Simplified JSON AST” was manually trimmed and lacks a consistent standard. In contrast, our AST format (used in Uni-MuMER) strikes a balance between structural clarity and brevity with special delimiters to indicate hierarchical relationships. Our format avoids the heavy overhead of full JSON while preserving the necessary structure of the expression tree.

Tab. 12 reports the model performance using each of these tree formats, which shows that the choice of format significantly impacts accuracy. We attribute this gain to our format’s compactness and structured guidance: it provides clear structural cues (like JSON) but with far fewer extraneous tokens, making it easier for the decoder to learn the task.

Table 12: **Performance impact of different expression tree formats for Tree-CoT.** Bold numbers are best in column. Setup: Models are trained on the CROHME dataset and evaluated on CROHME 2014, 2016, and 2019 test sets and their average. Results are reported in terms of ExpRate and ExpRate@CDM. We compare the baseline Vanilla model against variants that incorporate the Tree-CoT task using various AST output formats, and likewise compare full Uni-MuMER models using those formats.

Model	CROHME 2014		CROHME 2016		CROHME 2019		CROHME Average	
	Exp	@CDM	Exp	@CDM	Exp	@CDM	Exp	@CDM
Vanilla	72.62	75.66	64.26	67.83	69.14	71.98	68.67	71.82
Vanilla + S-expression	72.21	75.56	66.87	70.97	68.06	71.73	69.05	72.75
Vanilla + Json AST	71.70	75.66	66.70	70.51	70.56	74.04	69.65	73.40
Vanilla + Simplified Json AST	72.62	76.47	66.96	70.18	69.89	73.06	69.82	73.24
Vanilla + our AST	73.33	75.76	68.00	70.97	71.23	74.40	70.85	73.71
Uni-MuMER (w/ S-expression)	74.85	78.07	68.07	73.14	72.56	76.14	71.83	75.78
Uni-MuMER (w/ Json AST)	75.46	79.51	70.10	74.28	72.98	76.06	72.85	76.62
Uni-MuMER (w/ Simplified Json AST)	74.44	78.27	70.36	74.63	73.90	77.48	72.90	76.79
Uni-MuMER (w/ our AST)	75.36	79.11	70.79	75.24	73.73	77.23	73.29	77.19

In addition, we investigated alternative formats for the Symbol Counting (SC) auxiliary task. The default format in Uni-MuMER expresses the symbol counts as a simple list of symbol-count pairs, separated by commas. We compared this with a more explicit map/dictionary format. For example, consider the expression $x^{\{2\}} + y^{\{2\}} < 1$.

Example C.4: Map Counting format for Symbol Counting

`{'x': 1, '^': 2, '2': 2, '+': 1, 'y': 1, '<': 1, '1': 1}`

Example C.5: Our Counting format for Symbol Counting

`x: 1, ^: 2, 2: 2, +: 1, y: 1, <: 1, 1: 1`

The Map Counting format is semantically equivalent to ours but clearly more redundant (it introduces additional punctuation and quotation characters). We found that this extra overhead can make the sequence longer and potentially confuse the model (especially because symbols like braces and quotes have no direct meaning in the counting task itself). As a result, as shown in Tab. 13, the map format yielded slightly worse performance. The differences are modest, but consistent.

Table 13: **Impact of Symbol Counting formats.** Bold numbers are best in column. Setup: Models are trained on the CROHME dataset and evaluated on CROHME 2014, 2016, and 2019 test sets and their average. Results are reported in terms of ExpRate and ExpRate@CDM. We compare using a JSON-like Map Counting format versus our simple listing format for the SC task.

Model	CROHME 2014		CROHME 2016		CROHME 2019		CROHME Average	
	Exp	@CDM	Exp	@CDM	Exp	@CDM	Exp	@CDM
Vanilla	72.62	75.66	64.26	67.83	69.14	71.98	68.67	71.82
Vanilla + Map Counting	72.92	76.67	65.39	70.01	69.89	73.65	69.40	73.44
Vanilla + our Counting	71.91	76.85	67.04	72.28	70.64	75.15	69.86	74.76
Uni-MuMER (w/ Map Counting)	75.05	78.80	70.62	74.43	72.48	76.21	72.72	76.48
Uni-MuMER (w/ our Counting)	75.36	79.11	70.79	75.24	73.73	77.23	73.29	77.19

C.4 Hallucination for HMER

Large generative models are known to occasionally hallucinate output tokens that do not belong or omit tokens that should be present. In the context of HMER, both big VLMs and prior specialized models can produce spurious symbols. To quantitatively analyze this, we preliminarily adopt two metrics: Hallucination Token Rate (HTR) (percentage of predicted tokens not in the ground truth) and Missing Token Rate (MTR) (percentage of ground truth tokens not in the pred). Additionally, we categorized hallucinations by token type (numerical(0-9), variable(a-zA-Z), structural delimiters (e.g., [], {}, ()). This gives finer-grained error counts: HTR_num, HTR_var, HTR_delim for hallucinated numbers, letters, and delimiters, respectively. We evaluated these metrics across the CROHME 2014, 2016, and 2019 test sets, ensuring that no external data was used during training. The results are summarized below:

Table 14: **Error analysis of hallucinations and omissions on CROHME.** Results include ExpRate (Exp), ExpRate@CDM, overall HTR and MTR errors, as well as specific error types (HTR_num, HTR_var, HTR_delim). We compare two previous SOTA models (CoMER and TAMER), the Vanilla baseline, ablations adding each of our tasks individually (Tree-CoT, EDL, SC), and the full Uni-MuMER.

Model	Exp	@CDM	HTR	MTR	HTR_num	HTR_var	HTR_delim
CoMER (SOTA'2022)	59.18	62.47	2.55	3.91	0.37	0.33	1.18
TAMER (SOTA'2024)	60.29	63.18	2.17	4.02	0.29	0.29	1.05
Vanilla (baseline)	68.64	71.74	4.08	2.42	1.08	0.29	1.68
Vanilla + Tree-CoT	70.85	73.71	1.67	1.56	0.14	0.21	0.85
Vanilla + EDL	70.30	73.76	2.58	1.94	1.09	0.22	0.83
Vanilla + CAN	69.86	74.76	2.15	2.06	0.23	0.27	1.08
Uni-MuMER (w/ 3 tasks)	73.29	77.19	1.18	1.39	0.12	0.15	0.60

Tab. 14 shows that our full model yields the highest accuracy while also having the lowest HTR and MTR of all methods. Besides, incorporating each auxiliary task individually already helps: adding the Tree-CoT task slashes hallucinations to 1.67% HTR, adding the Symbol Counting task alone significantly lowers omissions (MTR), and adding the Error-Driven Learning task reduces hallucinations of delim and var. These improvements suggest that Tree-CoT primarily mitigates structural hallucinations, EDL corrects symbol-level mistakes, and SC enforces global consistency.

C.5 CJK influence for HMER

In Sec. 4.2, we excluded samples containing CJK characters of HME100K. This is primarily for two reasons:

1. Our prompt specifically targets recognizing mathematical expressions, not general textual content such as CJK characters. This filtering ensures a clear evaluation of expression recognition capability.
2. Some CJK-containing samples introduced problematic annotations unrelated to pure expressions (e.g., textual descriptions or numeric lists)

Table 15: Examples of CJK-containing image-caption pairs in HME100K that lie outside pure expression recognition, which influence the robustness of expression recognition and are outside the scope of math expression recognition.

image id	caption
train_63827	$= 88 \div 5$
train_11042	$1 \div 3 \div 14 \times (9.42 \div 3.14)$
train_70305	39的因数: 1, 3,

Having justified the filtering strategy, we additionally examine the effect of **CJK** characters. We first list typical CJK-containing captions to illustrate the distributional shift (Tab. 15), then compare performance under two evaluation settings: (i) test set filtered to remove CJK, and (ii) test set retaining CJK. We also test training variants that include or exclude CJK samples. For reference, we also include results for CoMER and TAMER (previous specialized models) evaluated on the same splits.

Table 16: Comparison of recognition performance on HME100K test sets with and without CJK characters. Results are reported in terms of ExpRate and ExpRate@CDM. We evaluate three models on each subset: CoMER (specialized, SOTA 2022), TAMER (specialized, SOTA 2024), and Uni-MuMER. For Uni-MuMER, we report two training conditions: w/o CJK and w/ CJK. For CoMER and TAMER, we evaluated open-sourced models under zero-shot inference settings.

Model	HME100K Test w/ CJK			HME100K Test w/o CJK		
	ExpRate	ExpRate@CDM	CDM	ExpRate	ExpRate@CDM	CDM
CoMER	68.47	70.19	96.79	68.70	70.29	96.81
TAMER	69.50	71.12	97.02	69.74	71.25	97.03
Uni-MuMER [†] (w/o CJK)	69.46	72.49	97.08	72.66	74.30	97.38
Uni-MuMER [†] (w/ CJK)	71.63	73.60	97.32	70.98	74.32	97.38

As is shown in Tab. 16, the performance of CoMER and TAMER remained nearly identical before and after filtering. Our model still outperforms previous methods, with the CJK characters the same in the training set, which suggests minimal effect on comparability, and including vs. excluding the CJK samples in training had minimal effect on overall HMER accuracy.

We further examine whether including CJK characters in training has any noticeable effect on the cross-domain performance of Uni-MuMER. Tab. 16 and Tab. 17 summarizes ExpRate@CDM results for two model variants across benchmarks. These results indicate that adding the small set of CJK-containing samples introduces a tiny trade-off: it may slightly compromise performance on the main handwriting benchmarks, while providing a negligible benefit to certain out-of-domain cases. We conclude that excluding the CJK data is a reasonable choice to keep the training focused, with no meaningful loss in generality.

Table 17: Effect of including CJK during training evaluated on ExpRate@CDM across benchmarks. Results are reported on CROHME-Average, CROHME 2023 test, Mathwriting test, and Im2Latexv2 test. The best results for each column are highlighted in bold.

Model	CROHME-Avg	CROHME 2023	Mathwriting	Im2Latexv2
Uni-MuMER [†] (w/o CJK)	82.89	74.42	69.11	88.99
Uni-MuMER [†] (w/ CJK)	82.21	72.14	68.91	90.28

D More Experimental Results

D.1 Results on Handwritten Mathematical Expression

Tab. 1 and Tab. 2 in the main paper, we focused on the strict ExpRate and visual CDM metrics. Following past lightweight specialized models in HMER, we report ≤ 1 and ≤ 2 tolerating up to one or two token prediction errors within the generated $\text{\LaTeX}$ sequence. Here, we provide additional metrics and comparisons on multiple datasets to fully characterize Uni-MuMER’s strengths.

Tab. 18 is an extended version of Tab. 1. Specifically, Tab. 18 reports not only ExpRate but also the lenient metrics allowing up to ≤ 1 and ≤ 2 token errors. We compare Uni-MuMER against a broad range of baselines: open-source VLMs, closed-source VLMs, previous SOTA methods on HMER, and their data augmentation versions.

Tab. 19 presents an extended evaluation of our method on the CROHME2023 test set, providing metrics including ExpRate, ≤ 1 and ≤ 2 token errors, as well as CDM-based evaluations. We benchmark Uni-MuMER and Uni-MuMER[†] against various baselines. Our baseline model already demonstrates competitive performance against previous SOTA methods. Notably, after incorporating external data, Uni-MuMER[†] exhibits a drop in ExpRate due to changes in LaTeX-style outputs but achieves outstanding performance in ExpRate@CDM.

Tab. 20 and Tab. 21 shows the results on the MNE dataset. This shows that the vanilla model is already competitive with the prior SOTA on these challenging sets. By introducing three data-driven tasks and external data, the Uni-MuMER and Uni-MuMER[†] outperform the prior best methods.

Tab. 22 shows a comparison of MathWriting using CER. Mathwriting is a dataset of handwritten expressions paired with online handwriting traces and offline rendered images. With its introduction, Google released two foundation models, PaLI [14] and PaLM-E [13], which are capable of reading both images and sequences of pen strokes as input [14]. While online handwriting traces typically aid models in mitigating ambiguity among visually similar characters, resulting in enhanced performance, our proposed Uni-MuMER model demonstrates superior results despite relying exclusively on offline image data for both training and evaluation.

Tab. 23 is an extended version of Tab. 2, presenting the performance of various models on the HME100K test set. We report ExpRate, ≤ 1 , ≤ 2 , and the CDM-based metrics for this dataset.

Table 18: **Performance comparison on CROHME datasets (%)**. We evaluate our model and four baselines using expression recognition rate (Exp), ≤ 1 , and ≤ 2 on the CROHME 2014, 2016, and 2019 test sets. We use the following notation for reported results: All results are zero-shot inference; ^a: data augmentation, ^p: reproduced results, [†]: external data used.

Method	CROHME 2014			CROHME 2016			CROHME 2019		
	Exp $\uparrow$	$\leq 1 \uparrow$	$\leq 2 \uparrow$	Exp $\uparrow$	$\leq 1 \uparrow$	$\leq 2 \uparrow$	Exp $\uparrow$	$\leq 1 \uparrow$	$\leq 2 \uparrow$
GPT-4o	50.61	65.21	77.69	46.03	61.99	73.85	49.79	64.79	75.06
Doubao-1.5-pro	50.41	68.15	77.69	49.00	61.99	73.85	46.87	68.81	78.65
Gemini2.5-flash	58.01	71.60	81.14	52.74	667.28	76.24	55.21	70.14	79.48
Qwen2.5-VL-3B	38.64	50.81	60.55	37.75	51.00	61.38	37.45	53.88	63.39
Qwen2.5-VL-7B	55.98	68.76	77.28	50.92	67.13	76.37	49.62	67.22	76.73
Qwen2.5-VL-72B	59.74	70.39	78.40	54.32	69.53	77.51	55.13	69.81	78.65
DenseWAP	50.1	–	–	47.5	–	–	–	–	–
BTTR	53.96	66.02	70.28	52.31	63.90	68.61	52.96	65.97	69.14
ABM	56.85	73.73	80.61	52.92	69.66	78.73	53.96	71.06	78.65
CAN-ABM	57.26	74.52	82.03	56.15	72.71	80.30	55.96	72.73	80.57
CoMER	58.38	74.48	81.14	56.98	74.44	81.87	59.12	77.45	83.87
PosFormer	60.45	77.28	83.68	60.94	76.72	83.87	62.22	79.40	86.57
TAMER	61.36	76.77	83.25	60.26	76.91	84.05	61.97	78.97	85.80
SSAN	60.85	75.56	82.25	60.02	76.22	83.28	61.83	79.08	86.08
CoMER ^a	59.33	71.70	75.66	59.81	74.37	80.30	62.97	77.40	81.40
CoMER ^{a,p,†}	46.96	60.65	69.27	43.33	56.84	64.60	48.29	64.22	71.64
PosFormer ^a	62.68	79.01	84.69	61.03	77.86	84.74	64.97	82.49	87.24
SSAN ^{a,†}	62.58	–	–	62.51	–	–	65.30	–	–
VLP [†]	60.41	74.14	78.90	60.51	75.50	79.86	62.34	76.81	81.15
HiE-VL ^{a,†}	73.30	84.00	87.00	70.70	82.10	86.40	69.30	81.90	86.20
Vanilla (baseline)	72.62	81.34	86.82	64.26	77.77	84.22	69.06	82.07	88.07
Uni-MuMER	75.36 ^{+2.74}	85.09 ^{+3.75}	89.05 ^{+2.23}	70.79 ^{+6.53}	83.44 ^{+5.67}	89.54 ^{+5.32}	73.73 ^{+4.67}	86.74 ^{+4.67}	91.83 ^{+3.76}
Uni-MuMER [†]	82.05 ^{+9.43}	89.45 ^{+8.11}	93.00 ^{+6.18}	77.94 ^{+13.68}	87.45 ^{+9.68}	92.07 ^{+7.85}	79.23 ^{+10.17}	90.91 ^{+8.84}	93.83 ^{+5.76}

Table 19: **Performance on CROHME2023 Test set (%)**. Evaluated by ExpRate, ExpRate@CDM, and CDM. The $\dagger$ denotes the use of external data. The p denotes the results reproduced by [55].

Method	CROHME 2023				
	ExpRate $\uparrow$	$\leq 1 \uparrow$	$\leq 2 \uparrow$	ExpRate@CDM $\uparrow$	CDM $\uparrow$
VLP	56.47	68.54	73.63	–	–
CoMER p	59.82	69.23	74.12	–	–
TST	60.72	70.51	75.00	–	–
TST †	62.59	73.59	77.51	–	–
Vanilla (baseline)	68.39	79.13	87.57	69.24	94.43
Uni-MuMER	70.26 ^{+1.87}	81.13 ^{+2.00}	88.35 ^{+0.78}	71.47 ^{+2.23}	95.12 ^{+0.69}
Uni-MuMER †	70.00 ^{+1.61}	79.78 ^{+0.65}	87.00 ^{-0.57}	74.42 ^{+5.18}	95.33 ^{+0.90}

Table 20: **Performance comparison on the MNE dataset (%)**. We compare the expression recognition rate (ExpRate) and CDM-based metrics between our model and previous SOTA models on the MNE N1/N2/N3 test sets. The $\dagger$ denotes the use of external data.

Method	N1			N2			N3		
	Exp $\uparrow$	@CDM $\uparrow$	CDM $\uparrow$	Exp $\uparrow$	@CDM $\uparrow$	CDM $\uparrow$	Exp $\uparrow$	@CDM $\uparrow$	CDM $\uparrow$
CoMER	59.73	–	–	37.17	–	–	24.04	–	–
PosFormer	60.59	–	–	38.82	–	–	36.82	–	–
MMHMER	62.03	–	–	39.47	–	–	34.29	–	–
Vanilla (baseline)	64.37	68.21	93.94	51.65	58.22	92.89	35.66	46.59	91.75
Uni-MuMER	70.08 ^{+5.71}	74.19 ^{+5.98}	95.33 ^{+1.39}	55.59 ^{+3.94}	62.83 ^{+4.61}	94.64 ^{+1.75}	38.53 ^{+2.87}	50.96 ^{+4.37}	93.63 ^{+1.88}
Uni-MuMER †	76.00 ^{+11.63}	81.44 ^{+13.23}	96.37 ^{+2.43}	58.88 ^{+7.23}	71.38 ^{+13.16}	96.08 ^{+3.19}	46.45 ^{+10.79}	66.67 ^{+20.08}	95.13 ^{+3.38}

Table 21: **Performance comparison on the MNE dataset (%)**. We compare the expression recognition rate (ExpRate), ≤ 1 , and ≤ 2 between our model and previous SOTA models on the MNE N1/N2/N3 test sets. The $\dagger$ denotes the use of external data.

Method	N1			N2			N3		
	Exp $\uparrow$	$\leq 1 \uparrow$	$\leq 2 \uparrow$	Exp $\uparrow$	$\leq 1 \uparrow$	$\leq 2 \uparrow$	Exp $\uparrow$	$\leq 1 \uparrow$	$\leq 2 \uparrow$
CoMER	59.73	77.55	84.11	37.17	53.95	65.13	24.04	32.31	36.34
PosFormer	60.59	77.97	84.32	38.82	56.91	66.12	36.82	40.30	43.10
MMHMER	62.03	78.13	84.32	39.47	56.57	65.46	34.29	36.68	40.51
Vanilla (baseline)	64.37	78.61	85.65	51.65	59.54	69.08	35.66	44.81	52.12
Uni-MuMER	70.08 ^{+5.71}	83.73 ^{+5.12}	90.29 ^{+4.64}	55.59 ^{+3.94}	64.80 ^{+5.26}	73.36 ^{+4.28}	38.53 ^{+2.87}	49.18 ^{+4.37}	57.24 ^{+5.12}
Uni-MuMER †	76.00 ^{+11.63}	87.31 ^{+8.70}	92.16 ^{+6.51}	58.88 ^{+7.23}	74.34 ^{+14.80}	82.90 ^{+13.82}	46.45 ^{+10.79}	54.37 ^{+9.56}	63.66 ^{+11.54}

Table 22: **Comparison with multi-modal models PaLI and PaLM-E on MathWriting**. CER is the character error rate (lower is better). The $\dagger$ denotes the use of external data.

Model	Input	CER $\downarrow$	Exp $\uparrow$	@CDM $\uparrow$	CDM $\uparrow$
PaLI	Image	8.07	–	–	–
PaLI	Ink	4.64	–	–	–
PaLI	Ink + Image	4.55	–	–	–
PaLM-E	Image	4.87	–	–	–
PaLM-E	Ink	6.46	–	–	–
PaLM-E	Ink + Image	4.22	–	–	–
Vanilla	Image	4.15	50.51	66.46	94.20
Uni-MuMER	Image	4.01	51.19	68.73	94.80
Uni-MuMER †	Image	4.00	51.45	69.11	94.84

D.2 Results on Printed Mathematical Expressions

An important aspect of our approach is whether a model fine-tuned primarily for handwritten expressions also performs well on printed mathematical expressions. To test this, we evaluate on the Im2LaTeXv2 dataset, a dataset of printed formulas rendered from $\text{\LaTeX}$. Minor normalization adjustments were applied to align the dataset with our preprocessing standards. Specifically, we performed synonym replacements such as changing expressions from $\text{\code{\le}}$ to $\text{\code{\leq}}$, inserted spaces to separate commands clearly (e.g., converting $\text{\code{\begin{align}}}$ to $\text{\code{\begin {align}}}$), and standardized

Table 23: **Performance on HME100K dataset (%)**. Evaluated by Exp, @CDM, and CDM. The [†] denotes the use of external data.

Method	HME100K				
	Exp $\uparrow$	$\leq 1 \uparrow$	$\leq 2 \uparrow$	@CDM $\uparrow$	CDM $\uparrow$
GPT-4o	22.96	37.32	47.40	27.02	83.13
Gemini2.5-flash	28.14	44.23	54.75	34.32	85.43
Doubao1.5-pro	43.90	64.52	74.09	55.08	91.28
Qwen2.5-VL-3B	44.42	61.47	70.38	47.41	89.96
Qwen2.5-VL-7B	54.57	71.05	78.44	58.84	92.90
Qwen2.5-VL-72B	49.59	66.21	73.79	55.09	93.08
DenseWAP	61.85	70.63	77.14	–	–
CoMER	68.12	84.20	89.71	70.36	96.83
TAMER	69.50	85.48	90.80	71.30	97.07
HiE-VL [†]	64.20	–	–	–	–
Vanilla (baseline)	70.15	85.81	91.25	71.80	96.97
Uni-MuMER	71.62 ^{+1.47}	86.92 ^{+1.11}	92.16 ^{+0.91}	73.40 ^{+1.60}	97.30 ^{+0.33}
Uni-MuMER [†]	72.66 ^{+2.51}	87.67 ^{+1.86}	92.60 ^{+1.35}	74.30 ^{+2.50}	97.38 ^{+0.41}

notation by enclosing single-element subscripts like a_i within braces as $a_{\{i\}}$. Other than these normalization steps, no further modifications were made, and the mathematical semantics remained entirely unaffected. These normalization procedures are consistent with the preprocessing approach detailed in Appendix B. For transparency, we also include results from evaluations performed on the original, unnormalized test set. Such normalization adjustments were not applied to other test sets in our study, thus preserving fairness and comparability across datasets.

Tab. 24 summarizes these results using common metrics from the Im2LaTeXv2: Edit Score, BLEU-4, and ExpRate. We also include ExpRate@CDM for completeness, especially since our models output $\LaTeX$ that could be visually equivalent even if not exactly matching. Our Vanilla baseline performed slightly better than the previous SOTA. After completing three data-driven tasks, Uni-MuMER further reached an ExpRate of 88.52. We also check our Uni-MuMER[†], which is primarily optimized for handwriting. Interestingly, its performance on purely printed data is a bit lower than Uni-MuMER, perhaps because training on such a broad mix causes a slight domain shift. Nonetheless, all our models outperform MathNet [44].

Table 24: **Comparison of model performance on the *im2latexv2* test set**. We compare our models to prior Im2LaTeX systems. Metrics shown are Edit Score, BLEU-4, ExpRate, and ExpRate@CDM. All metrics are percentages (%), and higher values $\uparrow$ indicate better performance. The lower half reports results on the original, unnormalized test set. Missing values indicated by ‘–’.

Model	Train Dataset	Edit	Bleu-4	ExpRate	ExpRate@CDM
WYGIWYS	im2latex-100k	37.2	23.9	0.0	–
I2l-strips	im2latex-140k	75.9	65.9	10.3	–
I2l-nopool	im2latex-140k	76.0	66.4	10.4	–
MathNet	im2latexv2	97.2	96.8	83.9	–
Vanilla (baseline)	im2latexv2	99.12	98.44	88.08	93.19
Uni-MuMER	im2latexv2	99.16	98.70	88.52	93.46
Uni-MuMER [†]	Uni-MuMER-Data	99.33	98.71	88.19	93.29
<i>Evaluation on Unnormalized Test Set</i>					
Vanilla (baseline)	im2latexv2	98.71	97.59	79.85	93.19
Uni-MuMER	im2latexv2	98.75	97.85	80.27	93.46
Uni-MuMER [†]	Uni-MuMER-Data	98.54	96.72	76.36	93.29

E Case Study

We present a qualitative comparison to illustrate how Uni-MuMER handles a challenging expression versus other models. As shown in Fig. 8, we compare the outputs of closed-source VLMs, a prior lightweight HMER model (CoMER), our fine-tuned baseline (vanilla), and our Uni-MuMER. The figure displays each model’s predicted $\LaTeX$, highlighting any incorrect tokens in red. These case studies reinforce our quantitative findings: Uni-MuMER is not only quantitatively better but qualitatively produces significantly cleaner and more complete outputs, especially on expressions that involve several levels of nesting and a variety of symbols.

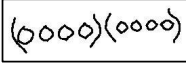
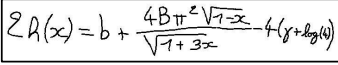
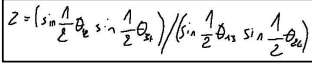
Images			
GPT4o	$(\infty \infty \infty \infty \infty \infty)$ $(\infty \infty \infty \infty \infty \infty)$	$2A(x) = b + \frac{4B\pi^2\sqrt{1-x}}{\sqrt{1+3x}} - 4(\gamma + \log(4))$	$z = \frac{\sin(\frac{1}{2}\theta_{k_1})\sin(\frac{1}{2}\theta_{k_2})}{\sin(\frac{1}{2}\theta_{k_1+k_2})\sin(\frac{1}{2}\theta_{k_1-k_2})}$
Gemini2.5-flash	$(ooo)(oooo)$	$2R(x) = b + \frac{4B\pi^2\sqrt{1-x}}{\sqrt{1+3x}} - 4(\gamma + \log(4))$	$z = \frac{\sin(\frac{1}{2}\theta_{k_1})\sin(\frac{1}{2}\theta_{k_2})}{\sin(\frac{1}{2}\theta_{k_1+k_2})\sin(\frac{1}{2}\theta_{k_1-k_2})}$
Doubao1.5-pro	$(a_{11}a_{22}a_{33}a_{44})$ $(b_{11}b_{22}b_{33}b_{44})$	$2R(x) = b + \frac{4B\pi^2\sqrt{1-x}}{\sqrt{1+3x}} - 4(\gamma + \log(4))$	$z = \frac{\sin(\frac{1}{2}\theta_{k_1})\sin(\frac{1}{2}\theta_{k_2})}{\sin(\frac{1}{2}\theta_{k_1+k_2})\sin(\frac{1}{2}\theta_{k_1-k_2})}$
CoMER	$(00000)(00000)$	$2h(x) = b + \frac{4B\pi^2\sqrt{1-x}}{\sqrt{1+3x}} - 4(\gamma + \log(4))$	$z = \frac{\sin(\frac{1}{2}\theta_{k_1})\sin(\frac{1}{2}\theta_{k_2})}{\sin(\frac{1}{2}\theta_{k_1+k_2})\sin(\frac{1}{2}\theta_{k_1-k_2})}$
Vanilla (baseline)	$(0000)(0000)$	$2h(x) = b + \frac{4B\pi^2\sqrt{1-x}}{\sqrt{1+3x}} - 4(\gamma + \log(4))$	$z = \frac{\sin(\frac{1}{2}\theta_{k_1})\sin(\frac{1}{2}\theta_{k_2})}{\sin(\frac{1}{2}\theta_{k_1+k_2})\sin(\frac{1}{2}\theta_{k_1-k_2})}$
Uni-MuMER [†]	$(0000)(0000)$	$2h(x) = b + \frac{4B\pi^2\sqrt{1-x}}{\sqrt{1+3x}} - 4(\gamma + \log(4))$	$z = \frac{\sin(\frac{1}{2}\theta_{k_1})\sin(\frac{1}{2}\theta_{k_2})}{\sin(\frac{1}{2}\theta_{k_1+k_2})\sin(\frac{1}{2}\theta_{k_1-k_2})}$

Figure 8: **Qualitative comparison on a challenging handwritten expression.** Compared with closed-source VLMs, CoMER, Vanilla(baseline), and Uni-MuMER[†]. The red symbols represent incorrect predictions.

CDM is not the ultimate evaluation metric yet

Although CDM metrics [50] quantify the visual equivalence between predicted and ground-truth $\LaTeX$ sequences beyond ExpRate, they are not definitive for evaluating $\LaTeX$ recognition accuracy. CDM tends to prioritize local character matching, potentially neglecting significant global structural discrepancies. This oversight can lead to inaccurate recognition assessments, as is shown in Fig. 9, CDM fails to detect the incorrect subscript structure of the character $u_x u_x$, mistakenly equating it with $u u_{xx}$, due to relying solely on local character matching without considering global hierarchical relationships.

$$u_{xt} = u_{xxx} - 6u u_{xx}$$

$$u_{xt} = u_{xxx} - 6u_x u_x$$

Figure 9: An example of CDM leading to incorrect recognition assessment.