Art4Math: Handwritten Mathematical Expression Recognition via Multimodal Sketch Grounding

Yang Zhou
22260043@zju.edu.cn
Zhejiang University
Hangzhou, China

Jin Wang*
dwjcom@zju.edu.cn
Zhejiang University
Hangzhou, China

Yuxiao Zhang
12225049@zju.edu.cn
Zhejiang University
Hangzhou, China

Kaixiang Huang
kaixianghuang@zju.edu.cn
Zhejiang University
Hangzhou, China

Guodong Lu
lugd@zju.edu.cn
Zhejiang University
Hangzhou, China

Jingru Yang
jingruy@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, USA

Shengfeng He
shengfenghe@smu.edu.sg
Singapore Management University
Singapore

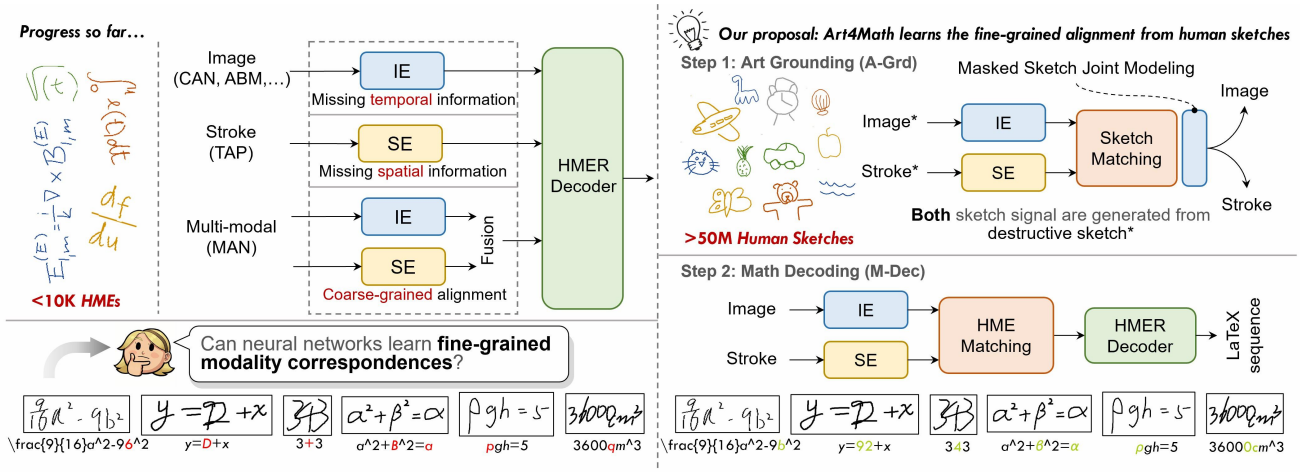

Figure 1: We propose Art4Math, a novel approach that leverages human sketches to enhance HMER. In contrast to prior unimodal methods, Art4Math supports multimodal inputs and enables fine-grained modality alignment. By pre-training on large-scale, easily accessible sketches, the model learns rich cross-modal correspondences, which in turn improve its ability to resolve ambiguities and structural issues in HMER. IE and SE denote the image and stroke encoder, respectively.

Abstract

Handwritten Mathematical Expression Recognition (HMER) remains a challenging task due to the structural complexity of mathematical notation and the ambiguity of handwritten symbols—e.g., "ρ" vs. "p" or "B" vs. "β". While stroke-based models offer disambiguation via temporal cues, most existing methods are constrained by coarse modality fusion and a lack of fine-grained cross-modal alignment, further hindered by limited annotated data. We introduce **Art for Math (Art4Math)**, a novel framework that leverages the structural richness of human sketches to enhance HMER through fine-grained, modality-aware learning. Art4Math follows a two-stage training paradigm: *Art Grounding (A-Grd)* and *Math Decoding (M-Dec)*. In A-Grd, the model is trained to reconstruct masked regions of sketches via joint modeling of visual and stroke-level features, encouraging sensitivity to local structural cues and inter-modality alignment. This Art Grounding cultivates a strong inductive bias for parsing abstract, sparse visual forms. M-Dec then adapts this representation to the HMER domain, enabling more precise symbol disambiguation and structural decoding with limited

*Corresponding author.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755247

supervision. Extensive experiments across sketch and handwriting-related tasks, including sketch recognition, retrieval, and HMER, demonstrate that Art4Math significantly outperforms existing self-supervised methods, revealing the overlooked synergy between artistic abstraction and mathematical expression.

CCS Concepts

• **Computing methodologies** → **Computer vision**.

Keywords

Multi-modal Learning, HMER, Sketch Representation Learning

ACM Reference Format:

Yang Zhou, Jin Wang, Yuxiao Zhang, Kaixiang Huang, Guodong Lu, Jingru Yang, and Shengfeng He. 2025. Art4Math: Handwritten Mathematical Expression Recognition via Multimodal Sketch Grounding. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755247

1 Introduction

Handwritten Mathematical Expression Recognition (HMER) plays a critical role in smart education, digital note-taking, and scientific writing. Despite the rapid advancements in Optical Character Recognition (OCR) technologies [47, 66], HMER remains an open challenge due to the intricate visual syntax of mathematical notation and the high variability in human handwriting.

Most existing HMER approaches treat handwritten mathematical expressions (HMEs) as 2D images [3, 12, 29, 59, 65]. While effective to an extent, this image-centric paradigm struggles with core ambiguities inherent in HME, such as visually similar symbols (e.g., "ρ" vs. "p") and overlapping strokes, that often turn recognition into a heuristic guessing game, as illustrated in Figure 1.

With the increasing ubiquity of touch-screen devices and stylus input, handwritten data is now naturally recorded as stroke sequences [14], which preserve the temporal dynamics of writing in addition to its visual form. Stroke-based representations offer a unique advantage: they encode the order in which symbols are written, providing critical disambiguation signals. However, these representations lack explicit spatial structure, which is essential for decoding the hierarchical layout of mathematical expressions. As a result, most stroke-based or multimodal HMER methods fuse strokes with images in a coarse fashion, illustrated as TAP [48] and MAN [60] in Figure 1, often treating strokes as auxiliary features. Consequently, performance improvements largely stem from image-based backbones, and the potential of multimodal learning remains underexploited.

To move beyond this bottleneck, we ask: *How can we learn fine-grained, modality-aware correspondences between strokes and pixels to improve HMER?* Achieving this goal is nontrivial, especially in light of the data scarcity in HMER. Existing datasets are insufficient to support the learning of complex, local cross-modal relationships [13, 35]. Interestingly, handwritten mathematical expressions share key properties with human sketches, both are sparse, symbolic abstractions rendered through strokes. Unlike HMEs, however, sketches are abundant and broadly accessible: anyone can produce a sketch in seconds [17, 56]. This insight motivates our central question: *What can human sketches do for HMER?*

We present **Art4Math** (Art for Math), the first framework to leverage large-scale human sketches to pre-train a fine-grained multimodal model for HMER. Art4Math adopts a two-stage training paradigm. In the first stage, *Art Grounding (A-Grd)*, the model learns to reconstruct corrupted sketches via joint modeling of visual and stroke modalities. To enable this, we introduce a global sketch-matching module to establish loose cross-modal alignment, inspired by vision-and-language pre-training methods [23, 40]. However, unlike natural image-text pairs where cross-modal alignment is ambiguous, sketch images and stroke sequences are two views of the same semantic signal. This unique property allows us to pursue precise, fine-grained alignment between pixels and coordinate points. To exploit this, we propose a *collaborative masking* (co-masking) mechanism that simultaneously masks both modalities while preserving the sketch's global semantics. We then design a self-supervised learning task, *masked sketch joint modeling*, that encourages the two modalities to interact and reconstruct the full sketch collaboratively. In the second stage, *Math Decoding (M-Dec)*, we adapt the Art Grounding model to HMER by attaching a standard decoder. The model, now equipped with fine-grained modality-aligned features, enables more robust recognition even under data constraints.

In summary, our contributions are threefold:

(1) We propose **Art4Math**, the first framework to explore the use of human sketches as a pretext modality to enhance handwritten mathematical expression recognition.
(2) We introduce a self-supervised learning paradigm that leverages masked sketch joint modeling to learn fine-grained modalities alignment during Art Grounding, which is then adapted to HMER in Math Decoding.
(3) We validate Art4Math through comprehensive experiments, demonstrating state-of-the-art performance on HMER and sketch-related tasks, and showcasing the untapped synergy between artistic abstraction and mathematical structure.

2 Related Works

Handwritten Mathematical Expression Recognition. HMER frameworks typically consist of an encoder that extracts image or stroke features and a decoder that parses the corresponding LaTeX sequence. Deng *et al.*[7] introduce a fine-grained attention mechanism to guide LaTeX generation. WAP[62], inspired by image captioning [53], employs coverage-based attention to dynamically focus on image regions during decoding. ABM [3] extends this idea by combining hidden states with image features and adopting a bidirectional decoder to generate LaTeX sequences in both directions. To enhance symbol awareness, CAN [26] introduces explicit symbol counting, while GCN [63] formulates a general category recognition task to classify symbol types. PosFormer [16] further improves positional understanding by integrating a position-forest structure into a Transformer backbone. Beyond raster-based inputs, stroke trajectories offer rich local and temporal cues. TAP [60] is a stroke-only approach, yet its performance is limited by the lack of spatial structure inherent in HME. To address this, MAN [48] fuses image and stroke modalities at the decoder level.

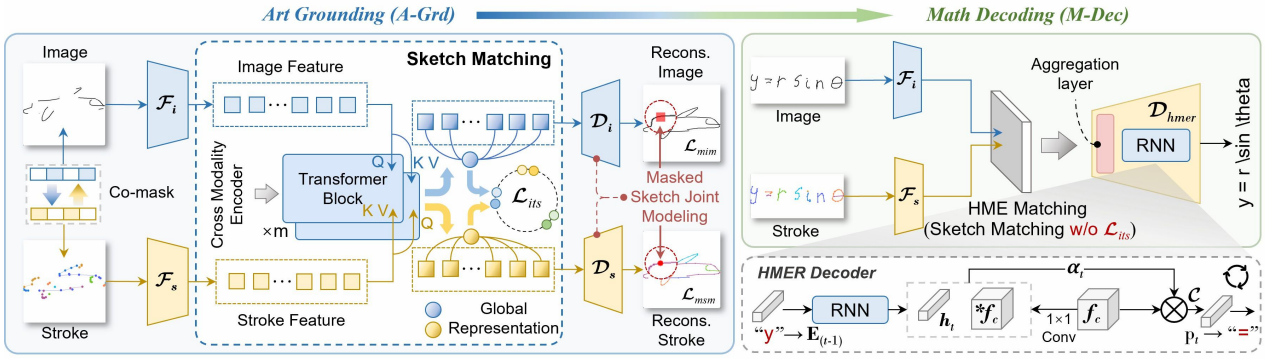

Figure 2: Illustration of Art4Math. It consists of an image encoder and a stroke encoder. *(i)* In A-Grd, we co-mask the sketch while retaining all the signals. Then the sketch matching module aligns modalities at the global level for better cross-modal representation. The sketch is reconstructed from the destructive signals by masked sketch joint modeling. *(ii)* In M-Dec, all parameters are transferred to the HMER, except for the sketch matching and masked sketch joint modeling components.

Sketch Representation Learning. Sketches can be represented either as sequences of stroke coordinates or as rasterized images. Effective sketch representation is crucial for downstream tasks such as recognition [38], retrieval [30, 33, 43–45], and generation [24, 39]. Typically, sketch images are processed using CNNs [57] or Vision Transformers (ViTs) [30], while stroke data is modeled with RNNs [28] or Transformer-based encoders [31, 38]. Unlike generic multimodal learning, where modalities may convey loosely aligned semantics, sketch data exhibits strong co-referentiality: both stroke and image modalities represent the same underlying concept. While sketch images capture spatial layout more effectively, stroke sequences preserve fine-grained temporal details and stylistic nuances.

Self-supervised Learning. Self-supervised learning approaches are generally divided into generative and contrastive paradigms. Generative models learn to reconstruct masked or corrupted data distributions, facilitating the discovery of latent structure. Inspired by masked language modeling (MLM) in NLP [8, 41], the computer vision community has developed masked image modeling (MIM), which has shown strong performance across various tasks [1, 4, 18, 52]. In contrast, contrastive methods learn global representations by pulling positive samples closer in the latent space [5, 15, 19]. Within the sketch domain, generative self-supervised approaches are more common. SketchBERT [31] masks stroke coordinates and predicts them using a Transformer encoder. Bhunia *et al.* [2] introduce dual tasks—rasterization and vectorization—that encourage cross-modal prediction between strokes and images.

3 Art4Math

Overview. In this section, we introduce Art4Math and its detailed implementation. There are two steps in our framework: *art grounding* (A-Grd) and *math decoding* (M-Dec). During A-Grd, the model is trained on human sketch data through a masked sketch joint modeling task. For M-Dec, the Art4Math model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using labeled data from the HMER task. The overall Art4Math framework is illustrated in Figure 2.

Input Representations. Touchscreen devices can easily capture handwritten data trajectories, which can be rasterized into images. We use a 5-element vector $v_t = (x_t, y_t, s_t^1, s_t^2, s_t^3)$ to characterize the trajectories. In particular, (x_t, y_t) denotes the absolute coordinates of the trajectory in a normalized $H \times W$ canvas, while the last three elements represent the binary one-hot vector of three pen states: pen touching the screen, pen being lifted and end of drawing. Thus, the size of the trajectory data is $S \in \mathbb{R}^{N\times 5}$, where N is the sequence length. The raster image is represented by a three-channel RGB image $I \in \mathbb{R}^{H\times W\times 3}$.

Model Architecture. Art4Math implements a parallel dual encoder architecture that processes images and strokes, respectively. The image encoder can be any convolutional neural network (CNN) or vision transformer (ViT). For ViT-based encoders, we use a 12-layer ViT-B/16 [10] as the image encoder $\mathcal{F}_i$, an input sketch image I is embedded into $L+1$ patches sequence: $\{p_{cls}, p_1, ..., p_L\}$, where p_{cls} is the global representation, which is the [CLS] token in transformer. For CNN-based encoders, we adopt DenseNet [21] following most previous works [26, 32, 63]. The feature sequence can be obtained by flattening the feature map before the average pooling operation. The global representation can be obtained by global average pooling. The stroke encoder $\mathcal{F}_s$ transforms an input stroke S into a sequence of embeddings $\{w_{cls}, w_1, ..., w_N\}$ by 5 trainable linear layers with hidden sizes $5-128-256-512-768$. The image features are fused with the sketch matching module.

3.1 Art Grounding

Art4Math first learns the alignment of images and strokes in large-scale human sketch data, described in this section. This step is presented in the left part of Figure 2.

Collaborative Mask. Previous cross-modal alignment models, such as the Vision-and-Language Pre-training (VLP) [22, 23, 27, 40], typically bring cross-modal representations closer to or push them farther away in latent space by optimizing contrastive loss or triplet loss. However, these methods only model global cross-modal correlations. To model the pixel and coordinate level correspondence of the sketch, inspired by MLM [9, 41], we randomly mask the stroke data, retaining 80% of the points. The masking strategy on the image side is equally essential. MAE [18] utilizes random mask sampling. However, this method may destroy the original sketch

signal, as the masks of the strokes and the image may overlap. To address this, our co-mask also uses a 20% mask rate, but samples from the remaining unmasked points and rasterizes them into the sketch image. Just this change, we can destroy the sketch signal while retaining all of it.

Sketch Matching Module. The two modalities of the sketch are first passed through a unimodal encoder to obtain the image feature $f_i \in \mathbb{R}^{L\times c}$ and the stroke feature $f_s \in \mathbb{R}^{N\times d}$, and their corresponding global representations p_{cls} and w_{cls}, respectively. Then the modality representations are projected to a shared feature space of dimension 512 by two separate MLPs. To fuse the sketch modality features, as shown in Figure 2, we devise a sketch matching module for cross-modality alignment and interaction. We adopt m-layer image-to-stroke and stroke-to-image cross-attention layers to explore cross-modal interaction ($m = 3$). This can be achieved by swapping the image query Q_i and stroke query Q_s, obtaining a new Query, Key and Value tuples, *i.e.*, (Q_s, K_i, V_i) and (Q_i, K_s, V_i). The Query, Key, and Value can be obtained by three different projection layers $[W_q, W_k, W_v]$. The cross-modal attention is obtained by

$$\mathrm{CA}(Q_i, K_s, V_s) = \texttt{softmax}(\frac{Q_i K_s^T}{\sqrt{d_{K_i}}})V_s. \tag{1}$$

In this way, stroke embeddings are updated by the information from image tokens, and the image embeddings do the same. Thus the destructive information can be complemented from different modalities.

We next optimize the image-to-stroke contrastive loss by computing softmax-normalized image-to-stroke and stroke-to-image similarities with batch size $\mathcal{B}$:

$$p_{i2s} = \frac{\exp(s(p_{cls}, w_{cls})/\tau)}{\sum_{b=1}^{\mathcal{B}} exp(s(p_{cls}, w_{cls})/\tau)}, p_{s2i} = \frac{\exp(s(w_{cls}, p_{cls})/\tau)}{\sum_{b=1}^{\mathcal{B}} \exp(s(w_{cls}, p_{cls})/\tau)} \tag{2}$$

where s is cosine similarity function, τ is a learnable temperature parameter, and b is the sample index within a mini-batch. Let y_{i2s} and y_{s2i} denote the ground-truth one-hot similarity, where the positive pair has a probability of 1 and the negative pair has a probability of 0. The image-stroke contrastive loss $\mathcal{L}_{its}$ is defined as the cross-entropy H between p and y:

$$\mathcal{L}_{its} = \frac{1}{2}\mathbb{E}_{(p_{cls}, w_{cls}\sim D)}[H(y_{i2s}, p_{i2s})] + [H(y_{s2i}, p_{s2i})] \tag{3}$$

Masked Sketch Joint Modeling. While the sketch matching module achieves global-level alignment, we further explore fine-grained modal correspondences. Therefore, we devise a masked sketch joint modeling task as our training objectives to reconstruct the original sketch images and strokes via an image decoder and a stroke decoder. These two objectives are analogous to MLM [8] and MIM [18]. However, the main difference is that Art4Math induces modality feature-level interactions through masked modeling. Co-mask, in turn, better achieves this through complementary sketch destruction. This implies that Art4Math leverages cross-modal relationships to guide the generated results rather than treating them as separate tasks. Thus, the results generated have higher fidelity, which matches downstream tasks like HMER.

Masked Sketch Image Modeling. We use a standard convolutional decoder $\mathcal{D}_i(\cdot)$ to reconstruct the original image from the image feature f_i. $\mathcal{D}_i$ consists of a series of deconvolutional layers that upsample the spatial size of the image feature to $H \times W$. Previous methods calculate the mean square error (MSE) loss between the generated image and the original image [2]. However, since sketch pixels are either 0 or 1, we treat this as a pixel prediction task to generate more precise sketch images and calculate the prediction probability for the i-th generated pixel f_i with $\hat{f}_i$ being the ground truth as:

$$\mathcal{L}_{mim} = -\frac{1}{L}\sum_{n=1}^{L}\sum_{i=1}^{2}\hat{f}_i \log(\frac{\exp(f_i)}{\sum_{j=1}^{2}\exp(f_j)}) \tag{4}$$

Masked sketch image modeling not only utilizes information from the image side but also incorporates the stroke side to obtain a complete reconstruction result. In this way, we enable the model to understand the correlation between the stroke coordinates and the image pixels to guide the reconstruction process.

Masked Sketch Stroke Modeling. Instead of adopting the strategy of predicting mask information like BERT [8, 31], we choose to reconstruct the whole stroke information to ensure that Art4Math can focus on the relationship between different stroke structures. We use a separate stroke decoder to reconstruct the stroke signal. The stroke decoder $\mathcal{D}_s(\cdot)$ is only used during A-Grd. We intend to focus Art4Math's cross-modal learning capabilities on the encoder side. Therefore, we adopt a lightweight independent decoder, which is an MLP with hidden sizes $512-256-128-64-5$. The output of $\mathcal{D}_s$ predicts 5-element stroke data. We use mean-square error and categorical cross-entropy loss to optimize the absolute coordinate and pen state, respectively. Thus, $(\hat{x_t}, \hat{y_t}, \hat{s_t^1}, \hat{s_t^2}, \hat{s_t^3})$ being the ground truth coordinate at t-th step, the training loss is:

$$\mathcal{L}_{msm} = \mathcal{L}_{coord} + \mathcal{L}_{state}, \tag{5}$$

where $\mathcal{L}_{coord}$ indicates the loss between the predicted coordinate value and groundtruth coordinate value, and $\mathcal{L}_{state}$ indicates the loss between the predicted drawing state and groundtruth state. The calculate method is:

$$\begin{aligned}\mathcal{L}_{state} &= -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{3}\hat{s_t^i}\log(\frac{\exp(s_t^i)}{\sum_{j=1}^{2}\exp(s_t^j)}),\\ \mathcal{L}_{coord} &= \frac{1}{N}\sum_{t=1}^{N}||\hat{x_t} - x_t||_2 + ||\hat{y_t} - y_t||_2.\end{aligned} \tag{6}$$

The full pre-training objective of Art4Math is:

$$\mathcal{L} = \mathcal{L}_{its} + \mathcal{L}_{mim} + \mathcal{L}_{msm} \tag{7}$$

3.2 Math Decoding

At the math decoding stage, we fine-tune the arbitrary HMER dataset described in the experiment and replace the sketch matching module with the HME matching module, which is straightforward to realize by just eliminating the $\mathcal{L}_{its}$. We then plug an HMER decoder $\mathcal{D}_{hmer}(\cdot)$ after the HME matching module, and its design is flexible. Instead of devising a decoder from the ground up, we use an RNN as the decoder $\mathcal{D}_{hmer}$ like previous HMER models [26] as shown in Figure 2. We calculate one stroke-to-image cross-attention to aggregate cross-modal representation, and then reshape the output to the required multimodal feature $f_c \in \mathbb{R}^{h\times w\times c_m}$, where $\frac{H}{h} = \frac{W}{w} = 16$ in our implementation. A convolutional operation with the 1×1 kernel is applied to obtain the transformed multimodal feature ${}^*f_c$. At time step t, the probability p_t of predicting the

symbol depends on context C, word embedding $E(y_t-1)$ from time step $t-1$ and current hidden state h_t:

$$p_t = \texttt{softmax}(\omega(W_c C + W_t h_t + W_e E(y_t - 1))) + b_o \quad (8)$$

where $\omega, W_c, W_t, W_e, b_o$ are trainable weights. Context C is computed as a weighted sum of multimodal feature ${}^{*}f_c^{i,j}$:

$$C = \sum_{i,j} \alpha_t^{i,j} {}^{*}f_c^{i,j}, (i = 1, 2, ..., h; j = 1, 2, ..., w) \quad (9)$$

where $\alpha_t^{i,j}$ is the attention weight of multimodal feature ${}^{*}f_c^{i,j}$ at time step t following the implementation of Bian *et al.* [3]. The decoder predicts the sequence iteratively until the special mark "end of sequence" is predicted. The HMER loss function $\mathcal{L}_{hmer}$ is a commonly used cross-entropy loss of the predicted probability to its ground truth.

4 Experiments

4.1 Datasets and Settings

Dataset. We use two popular datasets for evaluation on sketch-related tasks and two datasets for evaluation on the HMER task. (i) **QuickDraw** [17] contains 50 million sketches from 345 classes. These sketches are drawn by human players in the Quick, Draw! game and cover a wide range of common objects, animals, vehicles, everyday items, and abstract concepts. Each image sample is saved in the form of serialized handwritten stroke data. We use a split where each class has 70K training samples, 2.5K validation, and 2.5K test samples to pre-train Art4Math and evaluate sketch-related downstream tasks. (ii) **TU-Berlin** [11] is also used for the testing of sketch-related downstream tasks. It includes 250 categories, each containing 80 sketches. TU-Berlin sketches are drawn more finely (completed within 30 minutes) and stored in vector format. (iii) **CROHME** [35] is the most widely-used public dataset in the field of HMER, which is from the competition on recognition of online handwritten mathematical expressions. We use it to evaluate HMER. It contains 8836 HMEs, and three testing sets are provided: CROHME 2014, 2016, and 2019 with 986, 1147, and 1199 handwritten mathematical expressions, respectively. The number of symbol classes C is 111. (iv) **HME100K** [58] is a dataset of handwritten mathematical expressions written on paper, consisting of 74,502 images for training and 24,607 images for testing. Due to the lack of stroke sequences, we use it for testing Art4Math's image encoder.

Metrics. We report ExpRate (%) and ≤ 1 error to evaluate expression recognition accuracy [3, 26, 32, 64]. For sketch downstream tasks, Top-1 and Top-5 accuracy is used for recognition. We employ Acc@top1 and mAP@top10 as the evaluation metric for sketch retrieval.

Implementation Details. We pre-train Art4Math for 30 epochs with a batch size of 20 on 4 NVIDIA RTX4090 GPUs. For A-Grd, an AdamW optimizer with a weight decay of 0.005 is used. The image encoder is a ViT-B/16 [10] or DenseNet [21]. The stroke encoder is an 8-layer transformer with a hidden size of 768. The number of self-attention heads is 8. The learning rate is $1e-4$ in the first epoch and decays to $1e-7$ following a cosine schedule. The spatial size of sketch images is fixed at 224×224. In the M-Dec stage, for CROHME, an Adadelta optimizer with a weight decay of 0.0005 is used to fine-tune Art4Math for 240 epochs with batch size 12. The learning rate starts from 0 and monotonously increases to 0.01 at the end of the first epoch and decays to 0 following the cosine schedules. The HME image fixes the maximum spatial size of images to 768×192. For a transformer-like image encoder, we interpolate the positional embeddings for fine-tuning. The channel of transformed multimodal feature c_m is 512. For HME100k, the training epoch is set to 30.

Table 1: Comparison of Art4Math and state-of-the-art on CROHME. "†" indicates Art4Math with A-Grd, "*" indicates our reproduction.

	Methods	CROHME 2014		CROHME 2016		CROHME 2019	
		ExpRate	≤ 1	ExpRate	≤ 1	ExpRate	≤ 1
Image	UPV [35]	37.22	44.22	-	-	-	-
	TOKYO [36]	-	-	43.94	50.91	-	-
	PAL [50]	39.66	56.80	-	-	-	-
	WAP [62]	46.55	61.16	44.55	57.10	-	-
	PAL-v2 [51]	48.88	64.50	49.61	64.08	-	-
	DLA [25]	49.85	-	47.34	-	-	-
	DWAP [59]	50.10	-	47.50	-	-	-
	DWAP-TD [61]	49.10	64.20	48.50	62.30	51.40	66.10
	DWAP-MSA [59]	52.80	68.10	50.10	63.80	47.70	59.50
	WS-WAP [46]	53.65	-	51.96	64.34	-	-
	BTTR [65]	53.96	66.02	52.31	63.90	52.96	65.97
	ABM [3]	56.85	73.73	52.92	69.66	53.96	71.06
	CAN [26]	57.26	74.52	56.15	72.71	55.96	72.73
	TD-V2 [49]	53.62	-	55.18	-	58.72	-
	CoMER [64]	58.57	-	57.89	-	59.71	-
	SAM-CAN [34]	58.01	-	56.67	-	57.96	-
	TAMER* [67]	59.72	74.77	58.64	73.22	60.72	74.80
	GCN [63]	60.00	-	58.94	-	61.63	-
	NAMER [32]	60.51	75.03	60.24	73.50	61.72	75.31
	PosFormer [16]	60.45	77.28	60.94	76.72	62.22	79.40
w/ Stroke	TAP [60]	48.47	63.28	44.81	59.72	-	-
	MAN [48]	54.05	68.76	50.56	64.78	-	-
	Art4Math (ViT-B/16)	51.47	68.01	51.95	68.66	52.52	69.27
	Art4Math† (ViT-B/16)	58.11	75.40	57.92	74.79	59.15	78.24
	Art4Math (DenseNet)	56.62	74.31	55.38	73.86	55.97	74.54
	Art4Math†(DenseNet)	**63.18**	**79.34**	**62.21**	**78.52**	**63.87**	**80.40**

4.2 Comparison with SOTAs On HMER

HMER. First, we compare Art4Math against state-of-the-arts for HMER. We report their results directly from original papers, as shown in Table 1. Some of the works also report results with data augmentation, and due to the lack of specific implementation details, for a fair comparison, we focus only on results without data augmentation. We first test the performance of Art4Math without A-Grd, which still outperforms existing models using stroke sequences, thus providing a basis for better cross-modal alignment. While image-based models offer higher performance thanks to their respective strategies of symbol counting (*CAN*), non-autoregressive modeling (*NAMER*) or position forest (*PosFormer*), coarse-grained recognition (*GCN*), and other setups, these more complex architectural designs are not the focus of this paper given the generalization, Art4Math uses the most basic Encoder-Decoder architecture without adding "fancy" modules to better adapt to other designs in the future. However, Art4Math with A-Grd exploits the generalization potential of human sketches and surpasses SOTAs in all three testing sets. Specifically, it outperforms SOTAs by 2.67%, 1.27%, and 1.65% on the CROHME 2014, 2016, and 2019 datasets. We also evaluate a ViT-based image encoder that is not considered in previous works, and although it yields no top performance, the model after A-Grd obtains 6.64%, 5.97%, and 6.63% improvements

Table 2: Comparison of Art4Math and existing self-supervised learning methods on HMER datasets. l, h, FF, and N_h indicate the number of layers, hidden dimension, feed-forward size, and number of attention heads in a transformer. Art4Math-I and Art4Math-S are pre-trained Art4Math image encoder and stroke encoder, respectively.

Method	Modality		Arch.	Datasets			
	Stroke Space	Image Space		CROHME 2014	CROHME 2016	CROHME 2019	HME-100K
Vector [2]		✓	ViT-B/16	51.03%	49.71%	50.22%	62.21%
MAE [18]		✓	ViT-B/16	49.73%	48.88%	50.10%	58.66%
DINOV2 [37]		✓	ViT-B/14	51.41%	51.10%	51.32%	61.73%
MoCoV3 [6]		✓	ViT-B/16	47.83%	47.18%	48.02%	59.65%
Art4Math-I		✓	ViT-B/16	**53.56%**	**53.37%**	**54.00%**	**66.72%**
SketchBERT [31]	✓		$l = 8, h = 768, FF = 3072, N_h = 12$	35.22%	35.14%	36.31%	-
Raster [2]	✓		$l = 8, h = 768, FF = 2048, N_h = 12$	31.14%	29.82%	31.74%	-
Art4Math-S	✓		$l = 8, h = 768, FF = 1024, N_h = 8$	**42.83%**	**42.01%**	**43.44%**	-
Art4Math	✓	✓	Art4Math-full	**58.11%**	**57.92%**	**59.15%**	-

Table 3: HMER fine-tuning using 50% labelled training data on CROHME. All image encoders are DenseNet.

Methods	CROHME 2014 50% Training	CROHME 2016 50% Training	CROHME 2019 50% Training
SketchBERT [31]	13.42%	10.13%	12.91%
Vector [2]	39.12%	38.94%	39.64%
Raster [2]	8.37%	7.72%	9.11%
Art4Math	**47.38%**	**47.26%**	**46.11%**

on the three datasets, respectively. We attribute ViT's lower performance compared to DenseNet to the fact that CNNs are more suited for structure-sensitive HME images, whereas ViT relies on positional embedding to differentiate spatial information, making it more sensitive to resolution changes during fine-tuning. The results demonstrate that pre-training on human sketches can benefit HMER and thus motivate us to focus on other pre-training or self-supervised learning frameworks.

Self-supervised Learning for HMER. We compare our pretext task with existing self-supervised learning methods including *MoCoV3* [6], *MAE* [18], *DINOV2* [37], and *Vector* [2] for image input, while *SketchBERT* [31] and *Raster* [2] for stroke input. We fine-tune all the parameters of the networks. Most of these self-supervised learning methods are oriented toward RGB photos rather than sketches or handwritten data. *SketchBERT* and *Raster* are specialized methods for handling sketch sequences, while *Vector* is for handling sketch images. We use the same encoder for image input, except that *DINOV2* has no framework to provide ViT-B/16. For the stroke encoder, we follow the implementations in the original papers. In particular, we also provide two baselines for Art4Math. *Art4Math-I* and *Art4Math-S* directly utilize the unimodal feature outputs of the pre-trained image encoder and stroke encoder as inputs to the HMER decoder without using multimodal features.

The results in Table 2 demonstrate that Art4Math, even with a uni-modal encoder, surpasses existing self-supervised learning methods (53.56% vs. 51.41% on CROHME 2014). While *Vector* and *SketchBERT* also provide reasonable performance, there remains scope for improvement in cross-modal understanding, as they still lack effective temporal and spatial structural cues. By incorporating information from different modalities during the self-supervised learning stage, each single-modal encoder in Art4Math acquires a feature distribution that closely aligns with the other modality. Therefore, *Art4Math-S* easily surpasses them by a maximum of 6.52% ExpRate on CROHOME 2014. Finally, the full version of Art4Math, which supports multi-modal inputs, achieves the highest accuracy, further validating the benefit of multi-modal information for handwritten data.

Fine-tuning HMER with Reduced Data. Compared to human sketches, HME data is significantly scarcer. We next evaluate the methods trained on reduced data, including *Vector*, *Raster*, and *SketchBERT*, where we randomly sample 50% of the training set to create subsets for fine-tuning the whole network, ensuring that all symbols in the test set are covered. The results are shown in Table 3. Benefiting from multimodal data and a well-initialized representation, our approach effectively expands each sample into two modalities (image and stroke representation), achieving a significant performance gain over other self-supervised methods (8.26%, 8.32%, and 6.47% on CROHME 2014, 2016, and 2019, respectively). This shows the potential of using human sketches to boost handwriting recognition with reasonable results on small datasets.

4.3 Results on Sketch-Based Downstream Tasks

Art4Math is pre-trained on human sketches, and in addition to being generalizable in the HMER task, we are naturally tempted to evaluate its performance in downstream sketch-related tasks, including sketch recognition and sketch retrieval. Since the sketch-based downstream tasks do not involve variations in image size like HMER, for a fair comparison, we follow the traditional protocol of evaluation for self-supervised learning [2, 18, 37] and evaluate the linear probing performance of our method. The Art4Math image encoder uses ViT-B/16.

Sketch Recognition. For sketch recognition, we first apply average pooling to the multimodal feature to obtain a multimodal sketch representation and fine-tune the linear classifier. The linear probing result of Art4Math is significantly higher than existing self-supervised learning frameworks (76.92% vs.71.90% on QuickDraw) and remains competitive against the SOTA supervised method as

Table 4: Comparison of Art4Math and existing sketch recognition networks. "*" indicates our reproduction.

	Methods	Modality		Recognition Acc.	
		Image	Stroke	QuickDraw	TU-Berlin
Supervised	ResNet-50* [20]	✓		78.53%	73.40%
	Sketch-a-Net [57]	✓		68.71%	74.9%
	SketchMate [54]	✓	✓	80.51%	-
	SketchFormer [42]		✓	78.34%	-
	SketchAA [55]	✓		81.51%	-
	Sketch-R2CNN* [28]		✓	83.41%	74.71%
	SketchXAI [38]		✓	87.21%	-
Linear-probing	MoCoV3* [6]	✓		65.70%	68.12%
	MAE* [18]	✓		66.34%	68.52%
	SketchBERT* [31]		✓	66.1%	53.3%
	Raster [2]		✓	67.2%	55.6%
	Vector [2]	✓		71.9%	70.6%
	Art4Math	✓	✓	**76.92%**	**73.21%**

Table 5: Sketch retrieval results on QuickDraw

Sketch Retrieval			Image to Stroke Retrieval		
Method	AP@1	mAP@10	Method	AP@1	mAP@10
Vector [2]	62.49%	93.50%	Vecot-Raster [2]	10.85%	32.00%
Raster [2]	58.72%	93.00%			
SketchBERT [31]	55.90%	92.43%	**Art4Math**	**37.84%**	**82.13%**
Art4Math	**67.00%**	**94.07%**			

shown in Table 4, proving the superiority of Art4Math in the sketch-based task. We also provide end-to-end fine-tuning results in the Appendix.

Sketch Retrieval. We follow the implementation of Lin *et al.* [31], using 100 categories with 5K train samples, 2.5K validation samples, and 2.5K test samples to create a subset. We use the average pooled multimodal feature as the latent feature and fine-tune all the parameters. However, since Art4Math is designed for fine-grained cross-modal alignment, we design an image-to-stroke retrieval task to assess its capabilities. Specifically, we replace the global [CLS] token with a [RET] token, where an image embedding query is used to retrieve stroke embeddings. Due to the inherent modalities heterogeneity, this task is significantly more challenging than conventional unimodal sketch retrieval. The performance in Table 5 shows that Art4Math outperforms existing sketch self-supervised learning frameworks with 4.51% improvement on AP@1 at the general sketch retrieval task. Notably, Art4Math significantly surpasses the multimodal method *Vector-Raster* in cross-modal retrieval (37.84%*vs*.10.85%). We argue that *Vector* and *Raster* use two separate tasks to construct sketch relationships, which makes both positive and negative stroke embeddings far away from the anchor image embedding in latent space, resulting in hard optimization on the triplet. Art4Math learns sketch modality alignment jointly via sketch matching and masked sketch joint modeling, which is more advantageous in the cross-modal retrieval task. The retrieval results and visualization are provided in the Appendix.

4.4 Quantitative Analysis

Data Volume. We fine-tune the whole network to evaluate our method on HMER and sketch recognition with different training data volumes, shown in Figure 3. We reproduce two SOTA methods for fair comparison, *Vector* [2] for image modality and *SketchBERT*

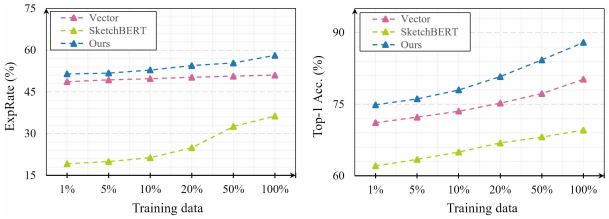

Figure 3: Fine-tuning at different training data sizes for HMER (left) and sketch recognition (right). ViT-B/16 is used for image modality, and the Transformer is for stroke modality inputs.

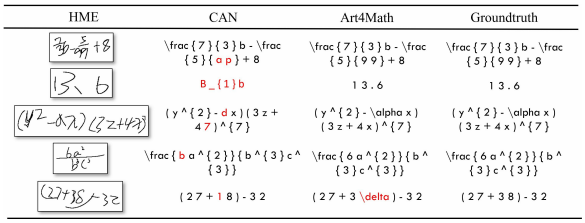

Figure 4: Some cases of CAN and our Art4Math.

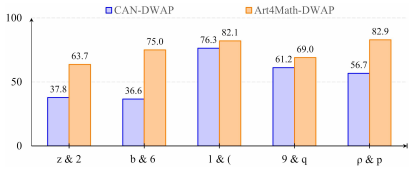

Figure 5: Accuracy statistics for confusing symbol examples.

[31] for stroke modality. It can be seen that the image-based models are generally better than the stroke-based models, but Vector's gain on HMER is limited. As the training data increases, the benefits of our method become more significant in both HMER and sketch recognition tasks.

HMER on Hard Cases. A strong motivation for Art4Math is to better cope with hard cases by multimodal signals, such as symbol confusion. Therefore, we customize a testing set of hard cases for comparing *Art4Math* with *CAN* [26], a popular image-based method. HMER decoders are standardized as ABM decoders [3] for fair comparison. The testing set contains 667 samples, more details are provided in the Appendix. The results are shown in Table 6. We also sample several combinations of typically confusing symbols and let them appear in an HME example, the statistical results are shown in Figure 5. It can be observed that Art4Math significantly improves recognition accuracy on challenging cases (9.79% ExpRate), particularly for easily confused symbols such as b vs. 6 and α vs. d shown in Figure 4. However, when both the stroke order and image representation introduce ambiguity, such as an unfinished 8 vs. δ — Art4Math isn't immune to errors either. A promising direction to tackle such cases lies in leveraging contextual information, which could help resolve these ambiguities.

Justifying design components. We evaluate our models (ViT-B/16 as the image encoder), dropping one component at a time (Table 7) for both sketch recognition and HMER. While not using

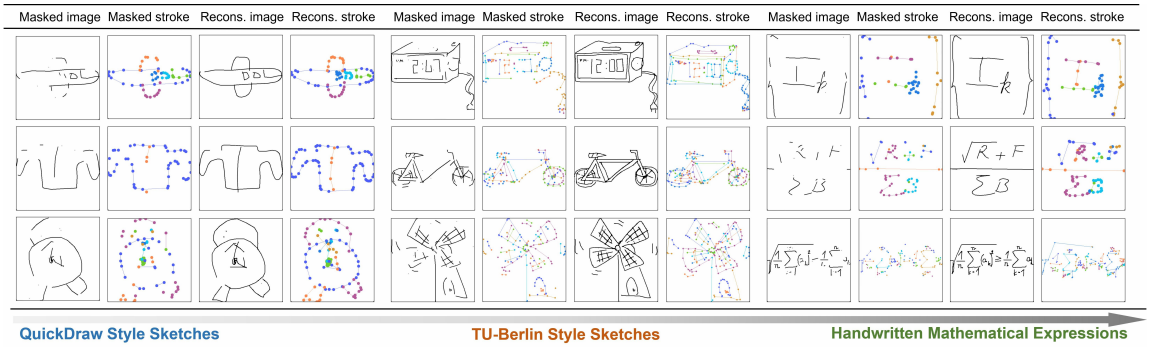

Figure 6: Masked reconstruction results of Art4Math from human sketches to handwritten mathematical expressions.

Table 6: Comparison of Art4Math and CAN network on the hard HMER dataset.

Method	Arch.	ExpRate
CAN [26]	DenseNet-ABM	66.12%
Art4Math	DenseNet-ABM	**75.91%**

Table 7: Ablation study of end-to-end fine-tuning on QuickDraw and CROHME 2014 datasets.

Methods	Sketch Rec.	HMER
	Top-1 Acc.	ExpRate
w/o Image encoder ($\mathcal{F}_i$)	70.41%	42.83%
w/o Stroke encoder ($\mathcal{F}_s$)	79.93%	53.56%
w/o Global alignment ($\mathcal{L}_{its}$)	84.26%	54.99%
w/o Masked image modeling ($\mathcal{L}_{mim}$)	82.09%	53.17%
w/o Masked Stroke modeling ($\mathcal{L}_{msm}$)	84.55%	54.21%
Ours (ViT-B/16 image encoder)	**87.70%**	58.11%
Ours (DenseNet image encoder)	85.63%	**63.18%**

fine-grained alignment slightly degrades the performance of the sketch recognition paradigm, it can severely affect HMER as it loses the ability to discriminate symbol ambiguity. As global alignment favors fine-grained alignment, removing $\mathcal{L}_{its}$ reduces the accuracy of sketch recognition and HMER by 6.66% and 3.12%, respectively, due to the difference in the model's sensitivity to the alignment of global semantic relations caused by the abstractness of sketches and the rigor of HME, where sketches are more semantically oriented as a whole, while HME relies more on local details. Furthermore, using DenseNet instead of ViT in HMER improves the ExpRate by 5.07%, thus being optimal, as it is more stable during fine-tuning. For fixed-size sketches, ViT improves by 2.29% compared to DenseNet.

Did the alignment really happen? Art4Math is trained on the QuickDraw dataset. Figure 6 shows some visualized examples of masked sketch joint reconstruction, and we also use two different handwriting style domains to demonstrate its adaptability, including TU-Berlin style sketches and handwritten math expressions. TU-Berlin sketches are more detailed and realistic due to the wider drawing time, while HMEs are more structurally regular. It can be seen that the inductive bias of A-Grd for abstract human sketches can adapt to new handwriting types, providing a promising basis for diverse downstream tasks. To gain more insight into whether

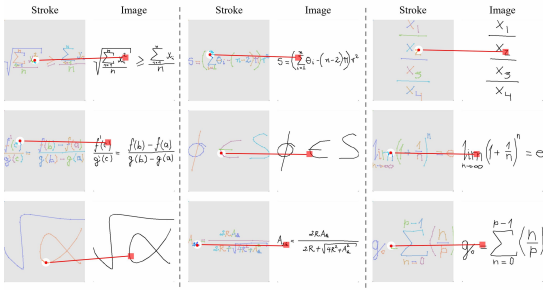

Figure 7: Visual correspondence across two modalities. Given a query HME with a selected key point, we show the images with the corresponding matched tokens (color-coded in red).

our model learns fine-grained alignment in human sketches, we use Art4Math (ViT-B/16) for visualization. We take the stroke data of HME in the CROHME dataset, randomly select a coordinate token as a query, and further find the highest-responsive token in the image modality by calculating the cosine similarity. As shown in Figure 7, such local matches do exist, which suggests that modality alignment knowledge learned from human sketches can be transferred to HME data.

5 Conclusion and Future Work

We present Art4Math, a framework that leverages human sketches to improve handwritten mathematical expression recognition (HMER) through fine-grained cross-modal alignment. In the Art Grounding stage, a co-masking strategy enables structured interaction between stroke and image modalities. The learned representations are transferred to the HMER task via Math Decoding, achieving state-of-the-art performance on both HMER and sketch-related tasks. Art4Math not only achieves SOTA performance on HMER and sketch-related benchmarks but also establishes a novel connection between human sketches and the structured semantics of mathematical notation for the first time.

A potential limitation of our approach lies in the fixed image resolution during Art Grounding. When transferred to HME data, this constraint may limit dense layout reconstruction. As future work, we plan to explore adaptive resolution strategies to further improve modality alignment and extend Art4Math to more handwriting and sketch-related tasks, such as text recognition and sketch-based image retrieval.

Acknowledgments

This work is supported by the National Key R&D Program of China (No.2022YFB3303102), Robotics Institute of Zhejiang University under Grant K11808 and K11811.

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254* (2021).

[2] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. 2021. Vectorization and rasterization: Self-supervised learning for sketch and handwriting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5672–5681.

[3] Xiaohang Bian, Bo Qin, Xiaozhe Xin, Jianwu Li, Xuefeng Su, and Yanfeng Wang. 2022. Handwritten mathematical expression recognition via attention aggregation based bi-directional mutual learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 113–121.

[4] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. Generative pretraining from pixels. In *International conference on machine learning*. PMLR, 1691–1703.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PmLR, 1597–1607.

[6] Xinlei Chen, Saining Xie, and Kaiming He. 2021. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9640–9649.

[7] Yuntian Deng, Anssi Kanervisto, Jeffrey Ling, and Alexander M Rush. 2017. Image-to-markup generation with coarse-to-fine attention. In *International Conference on Machine Learning*. PMLR, 980–989.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[11] Mathias Eitz, James Hays, and Marc Alexa. 2012. How do humans sketch objects? *ACM Transactions on graphics (TOG)* 31, 4 (2012), 1–10.

[12] Pengbin Fu, Ganyun Xiao, and Huirong Yang. 2024. SATD: syntax-aware handwritten mathematical expression recognition based on tree-structured transformer decoder. *The Visual Computer* (2024), 1–18.

[13] Philippe Gervais, Asya Fadeeva, and Andrii Maksai. 2024. Mathwriting: A dataset for handwritten mathematical expression recognition, 2024. *URL https://arxiv.org/abs/2404.10690* (2024).

[14] Trishita Ghosh, Shibaprasad Sen, Sk Md Obaidullah, KC Santosh, Kaushik Roy, and Umapada Pal. 2022. Advances in online handwritten recognition in the last decades. *Computer Science Review* 46 (2022), 100515.

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems* 33 (2020), 21271–21284.

[16] Tongkun Guan, Chengyu Lin, Wei Shen, and Xiaokang Yang. 2024. PosFormer: recognizing complex handwritten mathematical expression with position forest transformer. In *European Conference on Computer Vision*. Springer, 130–147.

[17] David Ha and Douglas Eck. 2017. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477* (2017).

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009.

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[21] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.

[22] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849* (2020).

[23] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*. PMLR, 5583–5594.

[24] Subhadeep Koley, Ayan Kumar Bhunia, Deeptanshu Sekhri, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. 2024. It's All About Your Sketch: Democratising Sketch Control in Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7204–7214.

[25] Anh Duc Le. 2020. Recognizing handwritten mathematical expressions via paired dual loss attention network and printed mathematical expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 566–567.

[26] Bohan Li, Ye Yuan, Dingkang Liang, Xiao Liu, Zhilong Ji, Jinfeng Bai, Wenyu Liu, and Xiang Bai. 2022. When counting meets HMER: counting-aware network for handwritten mathematical expression recognition. In *European conference on computer vision*. Springer, 197–214.

[27] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.

[28] Lei Li, Changqing Zou, Youyi Zheng, Qingkun Su, Hongbo Fu, and Chiew-Lan Tai. 2020. Sketch-R2CNN: an RNN-rasterization-CNN architecture for vector sketch recognition. *IEEE transactions on visualization and computer graphics* 27, 9 (2020), 3745–3754.

[29] Zhe Li, Xinyu Wang, Yuliang Liu, Lianwen Jin, Yichao Huang, and Kai Ding. 2023. Improving handwritten mathematical expression recognition via similar symbol distinguishing. *IEEE Transactions on Multimedia* 26 (2023), 90–102.

[30] Fengyin Lin, Mingkang Li, Da Li, Timothy Hospedales, Yi-Zhe Song, and Yonggang Qi. 2023. Zero-Shot Everything Sketch-Based Image Retrieval, and in Explainable Style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23349–23358.

[31] Hangyu Lin, Yanwei Fu, Xiangyang Xue, and Yu-Gang Jiang. 2020. Sketch-bert: Learning sketch bidirectional encoder representation from transformers by self-supervised learning of sketch gestalt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6758–6767.

[32] Chenyu Liu, Jia Pan, Jinshui Hu, Baocai Yin, Bing Yin, Mingjun Chen, Cong Liu, Jun Du, and Qingfeng Liu. 2024. NAMER: Non-Autoregressive Modeling for Handwritten Mathematical Expression Recognition. In *European Conference on Computer Vision*. Springer, 273–291.

[33] Qing Liu, Lingxi Xie, Huiyu Wang, and Alan L Yuille. 2019. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3662–3671.

[34] Zhuang Liu, Ye Yuan, Zhilong Ji, Jinfeng Bai, and Xiang Bai. 2023. Semantic graph representation learning for handwritten mathematical expression recognition. In *International conference on document analysis and recognition*. Springer, 152–166.

[35] Harold Mouchere, Christian Viard-Gaudin, Richard Zanibbi, and Utpal Garain. 2014. ICFHR 2014 competition on recognition of on-line handwritten mathematical expressions (CROHME 2014). In *2014 14th International Conference on Frontiers in Handwriting Recognition*. IEEE, 791–796.

[36] Harold Mouchère, Christian Viard-Gaudin, Richard Zanibbi, and Utpal Garain. 2016. ICFHR2016 CROHME: Competition on recognition of online handwritten mathematical expressions. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 607–612.

[37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

[38] Zhiyu Qu, Yulia Gryaditskaya, Ke Li, Kaiyue Pang, Tao Xiang, and Yi-Zhe Song. 2023. Sketchxai: A first look at explainability for human sketches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 23327–23337.

[39] Zhiyu Qu, Tao Xiang, and Yi-Zhe Song. 2023. Sketchdreamer: Interactive text-augmented creative sketch ideation. *arXiv preprint arXiv:2308.14191* (2023).

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[41] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[42] Leo Sampaio Ferraz Ribeiro, Tu Bui, John Collomosse, and Moacir Ponti. 2020. Sketchformer: Transformer-based representation for sketched structure. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14153–14162.

[43] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. 2023. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2765–2775.

[44] Aneeshan Sain, Ayan Kumar Bhunia, Vaishnav Potlapalli, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. 2022. Sketch3t: Test-time training for zero-shot sbir. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

Recognition. 7462–7471.

[45] Jialin Tian, Xing Xu, Fumin Shen, Yang Yang, and Heng Tao Shen. 2022. Tvt: Three-way vision transformer through multi-modal hypersphere learning for zero-shot sketch-based image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 2370–2378.

[46] Thanh-Nghia Truong, Cuong Tuan Nguyen, Khanh Minh Phan, and Masaki Nakagawa. 2020. Improvement of end-to-end offline handwritten mathematical expression recognition by weakly supervised learning. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 181–186.

[47] Jianqiang Wan, Sibo Song, Wenwen Yu, Yuliang Liu, Wenqing Cheng, Fei Huang, Xiang Bai, Cong Yao, and Zhibo Yang. 2024. Omniparser: A unified framework for text spotting key information extraction and table recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15641–15653.

[48] Jiaming Wang, Jun Du, Jianshu Zhang, and Zi-Rui Wang. 2019. Multi-modal attention network for handwritten mathematical expression recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1181–1186.

[49] Changjie Wu, Jun Du, Yunqing Li, Jianshu Zhang, Chen Yang, Bo Ren, and Yiqing Hu. 2022. Tdv2: A novel tree-structured decoder for offline mathematical expression recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 2694–2702.

[50] Jin-Wen Wu, Fei Yin, Yan-Ming Zhang, Xu-Yao Zhang, and Cheng-Lin Liu. 2019. 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*. Springer, 18–34.

[51] Jin-Wen Wu, Fei Yin, Yan-Ming Zhang, Xu-Yao Zhang, and Cheng-Lin Liu. 2020. Handwritten mathematical expression recognition via paired adversarial learning. *International Journal of Computer Vision* 128 (2020), 2386–2401.

[52] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9653–9663.

[53] Kelvin Xu. 2015. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044* (2015).

[54] Peng Xu, Yongye Huang, Tongtong Yuan, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, Zhanyu Ma, and Jun Guo. 2018. Sketchmate: Deep hashing for million-scale human sketch retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8090–8098.

[55] Lan Yang, Kaiyue Pang, Honggang Zhang, and Yi-Zhe Song. 2021. Sketchaa: Abstract representation for abstract sketches. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10097–10106.

[56] Lan Yang, Kaiyue Pang, Honggang Zhang, and Yi-Zhe Song. 2024. Annotation-Free Human Sketch Quality Assessment. *International Journal of Computer Vision* 132, 8 (2024), 2743–2764.

[57] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. 2017. Sketch-a-net: A deep neural network that beats humans. *International journal of computer vision* 122 (2017), 411–425.

[58] Ye Yuan, Xiao Liu, Wondimu Dikubab, Hui Liu, Zhilong Ji, Zhongqin Wu, and Xiang Bai. 2022. Syntax-aware network for handwritten mathematical expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4553–4562.

[59] Jianshu Zhang, Jun Du, and Lirong Dai. 2018. Multi-scale attention with dense encoder for handwritten mathematical expression recognition. In *2018 24th international conference on pattern recognition (ICPR)*. IEEE, 2245–2250.

[60] Jianshu Zhang, Jun Du, and Lirong Dai. 2018. Track, attend, and parse (tap): An end-to-end framework for online handwritten mathematical expression recognition. *IEEE Transactions on Multimedia* 21, 1 (2018), 221–233.

[61] Jianshu Zhang, Jun Du, Yongxin Yang, Yi-Zhe Song, Si Wei, and Lirong Dai. 2020. A tree-structured decoder for image-to-markup generation. In *International Conference on Machine Learning*. PMLR, 11076–11085.

[62] Jianshu Zhang, Jun Du, Shiliang Zhang, Dan Liu, Yulong Hu, Jinshui Hu, Si Wei, and Lirong Dai. 2017. Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition. *Pattern Recognition* 71 (2017), 196–206.

[63] Xinyu Zhang, Han Ying, Ye Tao, Youlu Xing, and Guihuan Feng. 2023. General category network: Handwritten mathematical expression recognition with coarse-grained recognition task. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[64] Wenqi Zhao and Liangcai Gao. 2022. Comer: Modeling coverage for transformer-based handwritten mathematical expression recognition. In *European conference on computer vision*. Springer, 392–408.

[65] Wenqi Zhao, Liangcai Gao, Zuoyu Yan, Shuai Peng, Lin Du, and Ziyin Zhang. 2021. 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*. Springer, 570–584.

[66] Zhen Zhao, Jingqun Tang, Binghong Wu, Chunhui Lin, Shu Wei, Hao Liu, Xin Tan, Zhizhong Zhang, Can Huang, and Yuan Xie. 2024. Harmonizing visual text comprehension and generation. *arXiv preprint arXiv:2407.16364* (2024).

[67] Jianhua Zhu, Wenqi Zhao, Yu Li, Xingjian Hu, and Liangcai Gao. 2024. TAMER: Tree-Aware Transformer for Handwritten Mathematical Expression Recognition. *arXiv preprint arXiv:2408.08578* (2024).