Handwritten Chemical Structure Image to Structure-Specific Markup Using Random Conditional Guided Decoder

Jinshui Hu
jshu@mail.ustc.edu.cn
University of Science and Technology of China

Hao Wu
haowu16@iflytek.com
iFLYTEK Research

Mingjun Chen
mjchen5@iflytek.com
iFLYTEK Research

Chenyu Liu
cyliu7@iflytek.com
iFLYTEK Research

Jiajia Wu
jjwu@iflytek.com
iFLYTEK Research

Shi Yin
shiyin@iflytek.com
iFLYTEK Research

Baocai Yin
bcyin@iflytek.com
iFLYTEK Research

Bing Yin
bingyin@iflytek.com
iFLYTEK Research

Cong Liu
congliu2@iflytek.com
iFLYTEK Research

Jun Du*
jundu@ustc.edu.cn
University of Science and Technology of China

Lirong Dai*
lrdai@ustc.edu.cn
University of Science and Technology of China

ABSTRACT

Satisfactory recognition performance has been achieved for simple and controllable printed molecular images. However, recognizing handwritten chemical structure images remains unresolved due to the inherent ambiguities in handwritten atoms and bonds, as well as the signifcant challenge of converting projected 2D molecular layouts into markup strings. Target to address these problems, this paper proposes an end-to-end framework for handwritten chemical structure images recognition, with novel structure-specific markup language (SSML) and random conditional guided decoder (RCGD). SSML alleviates ambiguity and complexity in Chemfig syntax by designing an innovative markup language to accurately depict molecular structures. Besides, we propose RCGD to address the issue of multiple path decoding of molecular structures, which is composed of conditional attention guidance, memory classification and path selection mechanisms. In order to fully confirm the effectiveness of the end-to-end method, a new database containing 50,000 handwritten chemical structure images (EDU-CHEMC) has been established. Experimental results demonstrate that compared to traditional SMILES sequences, our SSML can significantly reduces the semantic gap between chemical images and markup strings. It is worth noting that our method can also recognize invalid or non-existent organic molecular structures, making it highly applicable for tasks related to teaching evaluations in the fields of chemistry and biology education. The EDU-CHEMC will be released soon in https://github.com/iFLYTEK-CV/EDU-CHEMC.

CCS CONCEPTS

• **Human-centered computing** → **Text input**; • **Computing methodologies** → **Object recognition**; • **Information systems** → **Multimedia information systems**.

KEYWORDS

Optical Chemical Structure Recognition, Handwritten OCR, Attention-based Encoder-Decoder Neural networks

ACM Reference Format:
Jinshui Hu, Hao Wu, Mingjun Chen, Chenyu Liu, Jiajia Wu, Shi Yin, Baocai Yin, Bing Yin, Cong Liu, Jun Du, and Lirong Dai. 2023. Handwritten Chemical Structure Image to Structure-Specific Markup Using Random Conditional Guided Decoder. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3581783.3612573

1 INTRODUCTION

Recognition systems for chemical molecular structures have been widely utilized in various fields, including pharmaceutical research and development, human-computer interaction, biochemistry, education, and organic synthesis [13, 16, 24]. These systems are particularly valuable for pharmaceutical enterprises, as they enable the gathering and organization of millions of chemical molecular structures from academic journals and patents spanning several decades. By incorporating visualization technology, large-scale chemical molecular structure databases [5] can be enhanced, improving efficiency and convenience in analyzing drug molecular structures. Therefore, it is crucial to develop user-friendly tools that facilitate effective human-computer interaction.

Thanks to advancements in computer vision techniques [8, 12, 15, 20], recent years have witnessed significant advancements in

*co-corresponding author

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612573

chemical molecular structure recognition for both the commercial and academic communities [4, 5, 22, 23, 25, 35, 40]. Despite the great success, most existing methods overlook recognizing handwritten molecular images and oversimplify the complexity of molecular structure recognition, relying excessively on rule-based post-processing methods. Recognizing large and complex molecular structures still remains a great challenge due to the complexity of 2D projections from 3D molecular structures, such as Natta projection and Fischer projection.

To model complex molecular structures and generalize to variant handwritten images, in this paper, we propose a novel solution which has two prominent components. First, a new structure-specific markup language (SSML) is designed to minimize the semantic gap between molecular structure images and sequence-styled markup annotations. The proposed SSML is derived from the visual characteristics of molecular structures, making it more suitable for handling complex molecular structures. Second, we propose a random conditional guided decoder (RCGD), which can be seen as a graph traversal, to address the issue of multiple path decoding of molecular structures (Fig. 1).

As shown in Fig. 1(a), the process of RCGD starts with a branch point which generates two candidate branch angle units (BAUs), with BAU-1 pointing upwards and BAU-2 to the lower right. Both of them are sent to the memory and tagged as unexplored. Subsequently, in Fig. 1(b), RCGD selects BAU-1, and marks it as explored with light yellow. The decoding process continues by traversing upwards along BAU-1. At the moment of Fig. 1(e), the red path completes the traversal of one branch, and then RCGD selects BAU-6 to continue traversing (the blue arrow). However, the atom it encounters is already connected with a visited atom (two red solid circles), resulting in a "re-connection" relationship that needs to be predicted by the model. The corresponding candidate BAU-8 is explored, and its states are marked as light yellow in Fig. 1(f). The decoding process continues in this way and concludes until all bonds and atoms on the graph are traversed.

To validate the effectiveness of our method, we establish a new database named EDU-CHEMC which contains 50,000 handwritten molecular structure images as well as corresponding SSML-styled annotations. Extensive experiments are conducted on both the public Mini-CASIA-CSDB data set [5] and our EDU-CHEMC. With the proposed structure-specific markup language, our end-to-end decoder RCGD achieves state-of-the-art performances with an exact match (EM) score of 95.01% and 62.86% on the Mini-CASIA-CSDB test set and EDU-CHEMC test set using a DenseNet [10] backbone, outperforming other methods by a large margin.

Our main contributions can be summarized as follows:

- We establish a new benchmark named EDU-CHEMC, comprising 50,000 handwritten molecular structure images gathered from diverse devices such as cameras, scanners, and electronic screens. SSML-styled annotations are also available and will be made public soon.
- A new molecular structure markup language (SSML) is designed. SSML demonstrates greater consistency with images and is not restricted by chemical knowledge. This flexibility allows it to represent erroneous or non-existent molecular structures, making it particularly suitable for learning.

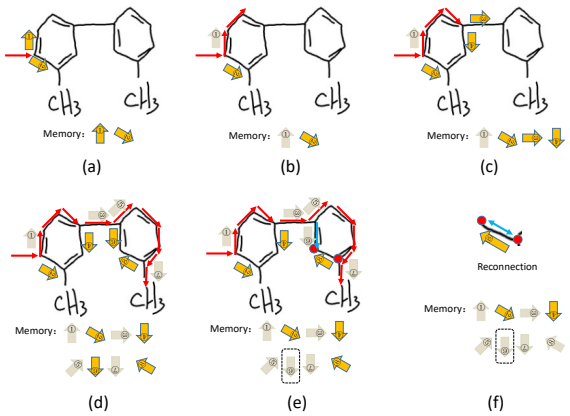

Figure 1: Illustration of the "graph traversal process" in RCGD. The "Memory" stores the current states of all branch angle units (BAUs). (a) to (e) show partial traversal paths in temporal order, where the red arrows cover the visited bonds and atoms, and the numbered arrows represent the candidate BAUs with their IDs. The deep yellow arrow indicates unexplored BAUs, while the light yellow arrow denotes explored or exhausted ones. (f) In (e) denotes a "re-connection" relationship (two red solid circles) where an atom hass already connected to a previously visited atom.

- We introduce a novel molecular structure recognition method called Random Conditional Guided Decoder (RCGD). By incorporating conditional attention guidance, memory classification and path selection mechanisms, RCGD surpasses conventional string decoders by a large margin.

2 RELATED WORK

2.1 Molecular structure representation

The representation format for chemical molecular structures has always been a crucial concern in academic writing related to chemical and biological technologies. Traditionally, molecular structures were predominantly conveyed through images. However, the emergence of big data and knowledge discovery has led to the increasing use of SMILES [30] or inChl [9] notation in recent chemistry papers to represent chemical formulas. Furthermore, RDKit, an open-source package, provides support for diverse chemical informatics operations, such as 2D and 3D molecular structural manipulations, molecular visualization, and other functionalities. The RDKit toolkit facilitates seamless conversion between SMILES and Chemfig [21]. In this paper, we utilize Chemfig for annotating molecular formulas. Chemfig is specifically designed for drawing molecular structure images, thereby reducing the semantic gap between the visual representation and the corresponding annotation.

2.2 Handcrafted rules based modeling

In the field of molecular structure recognition, a combination of traditional image analysis, and recognition with rule-based post-processing has been the predominant method [23, 35, 40] between

1990 and 2017. This approach involves a sequence of mandatory procedures. Firstly, image pre-processing techniques are applied such as noise reduction, thinning, and enhancing images. Afterwards, the corresponding handcrafted features, including SIFT or HOG features, are extracted. Finally, a perceptron or SVM classifier is employed for recognition. To recognize text in structural formulas image, either template matching or existing OCR engines are usually utilized. Precise reconstruction of molecular structural formula characteristics is essential for higher performance. Accurate and clear identification of each component's location and properties is essential. The recognition accuracy is significantly affected if the position precision is insufficient.

Recently, advancements in computer vision techniques for object detection [2, 27, 28, 36] and semantic segmentation [11, 12, 17, 19, 29, 37, 43, 46] have significantly improved the detection of molecular elements and chemical bonds. Combining extracted content and location of elements, together with rule-based post-processing [35] has led to higher accuracy. To further improve effectiveness, several studies focus on identifying recognition error patterns in the model. Additionally, machine translation models [26, 31] have been used to post-process and correct identified structural errors.

2.3 End-to-End modeling

Scholars in the field of hand-written mathematical expression recognition have proposed end-to-end recognition methods [14, 33, 38, 40, 41, 45] based on sequence modeling strategies, in addition to chemical molecular structure recognition tasks. These methods have significantly improved recognition accuracy and expanded their application scope, including automatic math problem-solving robots. Further research on TreeDecoder [41, 45], structured string decoder [33], and other methods [32, 34, 39, 42] has led to the realization that reducing the semantic gap between images and annotations is an effective means of completing image to markup generation. Drawing inspiration from this idea, this paper proposes a structure-specific sequence modeling method designed explicitly for molecular structure recognition. By doing so, this approach reduces the semantic gap between images and annotations while also more effectively tagging the image structure for direct and efficient modeling. Currently, several related end-to-end molecular structure recognition methods [4, 23, 31] are available. However, using SMILES format for chemical structure representation has limitations. It requires domain-specific knowledge and does not provide complete information regarding chemical structures, resulting in less effective modeling and recognition efficiency.

3 METHODOLOGY

In this section, we first introduce the details of Structure-Specific Markup Language (SSML) and then present the Random Conditional Guided Decoder (RCGD).

3.1 The Structure-Specific Markup Language

Our SSML is most closely-related to Chemfig, a LaTeX package utilizing TikZis and serves as a markup language for chemical structural formulas. Compared to the most commonly used SMILES, Chemfig has advantages on keeping structural appearance of chemical images and requires minimal abstract chemical knowledge.

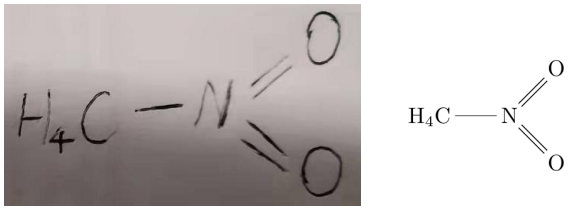

Figure 2: (a) The handwritten chemical molecular image. (b) The rendered image from the corresponding Chemfig string H_{4}C-N(=[1]O)=[-1]O.

Fig. 2(a) presents an handwritten chemical molecular image to illustrate the Chemfig syntax. In Fig. 2(a), the atom "N" is connected to two double bonds, one pointing towards the upper-right direction and the other towards the lower-right direction. Chemfig provides a description of "angle" to specify the orientation of bonds in a molecule, such as "[1]" and "[-1]" to express the approximate orientation of these two double bonds. The angle information enables Chemfig to stick to the molecular structure and build a more comprehensive visual representation.

Despite its advantages, employing Chemfig markup strings as the target for Encoder-Decoder training encounters following challenges: a) Ambiguity in Chemfig syntax. Different starting points and traversal orders result in multiple correct Chemfig sequences that can represent the same molecule image. This ambiguity is unavoidable and increases the difficulty of model learning. b) Complexity of Chemfig syntax. Prior rules and domain knowledge are necessary for Chemfig to ensure correct labeling of compound structures, which adds to the complexity. To address the above issues, we extend the Chemfig and propose a structure-specific markup language (SSML), as presented in Fig. 3. The structure-specific SSML is unambiguous and more effectively incorporates visual information, akin to following step-by-step instructions for drawing the molecular structure.

3.1.1 Structure Analysis. Fig. 3(a) displays three different Chemfig markup strings, all representing the same molecular formula. To resolve this ambiguity, SSML extends Chemfig's syntax and present a graph representation for molecular structure. Parsing with hand-crafted rules, we can recover all "atomic groups" and "chemical bonds" from the SSML-styled strings and then connect them to produce a graph. Fig. 3(b) displays a graph example where strings such as "HO", "COOH" or vertices of benzene rings are denoted as "atomic groups". For benzene ring, we treat the central ring as a special atomic group. The lines between atomic groups indicate "chemical bonds", which can be single "-", double "=", or triple "~" bonds. An atomic group can be connected to another atomic group through a single or multiple chemical bonds. By using atomic groups as vertices and chemical bonds as edges, we obtain a graph representation of the molecular structure (Fig. 3(c). The graph representations of the same molecule are identical even with various Chemfig markup strings, eliminating ambiguity in annotation.

3.1.2 Structure-Specific Markup Generation. The graph produced during the structure parsing step is a complex data structure that must be converted into a form suitable for model training, for Encoder-Decoder models, this typically means a one-dimensional string, or standardized label. Here we just briefly introduce the steps for generating the SSML:

- Step 1: Start with a graph representation of the molecule or chemical structure and traverse from a designated starting point (typically the left-most atom).
- Step 2: As you traverse, add the atoms and chemical bonds to the output string. For the atoms, include their written representation while ensuring consistency with their visual appearance. As for the chemical bonds, use the format "<bond>[:<angle value>]" to indicate the bond type and drawing angle. The angle value is calculated during graph construction and output according to the format.
- Step 3: When confronted with a branch point containing multiple branches, we follow an ascending traversing order according to chemical bond angles. For the selected branch, a pair of "phantom" symbols named "(" and ")" are employed to enclose the resulting markup strings.
- Step 4: If a reconnection is detected, the notations "?[<tag>]" and "?[<tag>,<bond>]" will be added to the output sequence following the relevant atoms. A pair of notations with the same tag denotes the starting and ending atoms of the reconnection. For instance, as illustrated in Fig. 3(c), "?[a]" and "?[a,-]" represent a reconnection with a single bond.
- Step 5: For the circle in benzene ring, we consider it as a special atom, defined as "\circle", which is connected to a certain atom on the benzene ring via a virtual bond "- -".
- Step 6: Once the traversal is finished, we obtain the SSML label, as depicted in Fig. 3(c), where modeling units are separated by spaces. The SSML comprises five types of elements, namely "atomic group," "bond", "angle", "phantom," and "reconnection mark".

The SSML mentioned above can be directly used as the training target for the String Decoder widely used in SMILES-based methods. For convenience, we refer to it as SD-SSML. For the RCGD method proposed in this paper, we need to make slight modifications. We refer to the modified markup language as RCGD-SSML, which will be discussed in next section together with the RCGD method.

3.2 Random Conditional Guided Decoder

Among all the attention-based Encoder-Decoder frameworks used for identifying handwritten mathematical expressions, the String Decoder and Tree Decoder are the most frequently used decoder modules. This paper employs the widely-used LaTeX-based string decoding paradigm, as it is a widely-used model for translating chemical structure images into SMILES strings. In light of this, we briefly review the RNN-based string decoder, followed by a discussion of its differences compared with the proposed RCGD.

3.2.1 String Decoder. We feed the image $\mathcal{I}$ into the encoder to extract visual features $x \in \mathbb{R}^{d_x \times h \times w}$,where d_x represents the dimension of the output features, and h and w represent the height and width of the feature map outputted by the encoder, which can

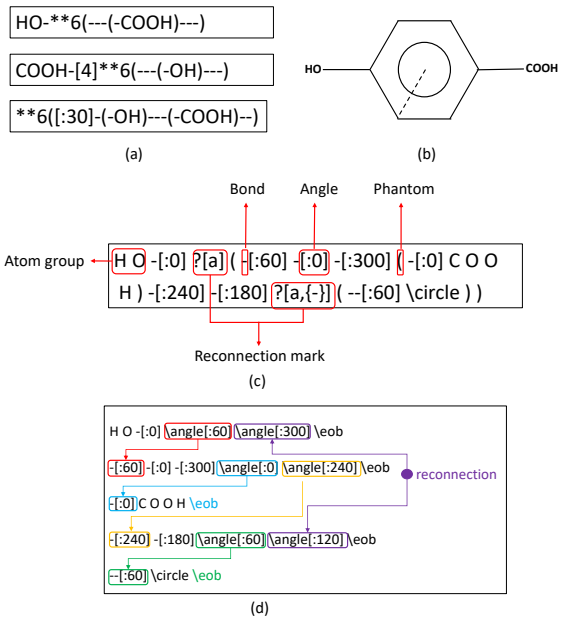

Figure 3: (a) Original equivalent Chemfig strings. (b) Molecular structure image. (c) The SSML label for String Decoder (SD-SSML). (d) The SSML label for RCGD (RCGD-SSML).

be a CNN or ViT [6, 18] structure. Once the encoded features are obtained, the model decodes each character in an autoregressive manner. In each step of the decoding process, attention modules are used to query the visual context features in x that are relevant to the current decoding state. The calculation process is as follows:

$$e_{t,i} = w^T \tanh(W^x x_i + W^y E y_{t-1} + W^s s_{t-1} + (W^\alpha * [\alpha_{t-1}; \sum_j^{t-1} \alpha_j])_i) \quad (1)$$

$$\alpha_{t,i} = \frac{e_{t,i}}{\sum_{j=1}^{h \times w} e_{t,j}} \quad (2)$$

$$c_t = \sum_{i=1}^{h \times w} \alpha_{t,i} x_i \quad (3)$$

Here, W^x, W^y, W^s, W^α and w are projection parameters for the attention module. $E \in \mathbb{R}^{d_y \times V}$ represents word embedding, with d_y being the dimension of the word embedding, and V representing the number of modeled unit characters. $y_{t-1} \in \mathbb{R}^V$ is the one-hot vector corresponding to the output of the previous step, and $s_{t-1} \in \mathbb{R}^{d_s}$ is the output state of the previous decoding step in the recurrent neural network, with d_s being the state dimension. $*$ denotes a convolution operation for modeling historical attention weight information. $e_{t,i}$ is the energy of x_i at the t-th decoding step. All the energy values are fed into the softmax function to obtain the attention weight α_t at all positions at the current decoding step. The visual context information is obtained by a weighted sum of the feature map. Finally, we combine the recurrent neural network to model the language model and complete the final classification decision:

$$
\begin{aligned}
s_t &= GRU(c_t, Ey_{t-1}; s_{t-1}) \\
p_t &= \text{softmax}(W^c s_t) \\
L_{ce} &= -\sum_{i=1}^{V} y_{ti} \log(p_{ti})
\end{aligned} \tag{4}
$$

Here, we use a GRU [3] as the computational unit for the recurrent neural network. $W^c \in \mathbb{R}^{V \times d_s}$ are learnable weights of the classification layer, and p_t represents the classification probability distribution at the t-th decoding step. We adopt cross-entropy loss L_{ce} as the training loss, while decoding can be performed during the inference phase using the beam search algorithm.

Although a String Decoder can fit training data well, it has difficulty in understanding meanings of different modeled units. This results in poor generalization performance on complex, incorrect, and reconnection structures. Hence, we propose a Random Conditional Guided Decoder (RCGD) that differs from String Decoder and other methods by incorporating three mechanisms to address these issues: conditional attention guidance, memory classification and path selection.

3.2.2 RCGD-SSML. For RGCD, some modifications need to be made to the SSML, referred to as RCGD-SSML. Here are the key changes in RCGD-SSML compared to SD-SSML:

(1) Deletion of branch and reconnection expressions: the original syntax for branch and reconnection "(", ")", "?[<tag>]", "?[<tag>, <bond>]" in SD-SSML are removed in RCGD-SSML.
(2) Addition of a branch ending symbol: In RCGD-SSML, a special symbol "\eob" is added to indicate the end of a branch.
(3) Introduction of a branch angle set M: RCGD-SSML includes a branch angle set M, which changes with each decoding step. The content of M is synchronized at each time step.
(4) New processing approach for branches:
    - When encountering a branch, sequential outputs are generated as "\angle[:<angle value 1>]", "\angle[:<angle value 2>]", ..., "\angle[:<angle value n>]", followed by "\eob". Each "\angle[:<angle n>]" represents a branch angle token. Each output branch angle modeling unit is synchronously added to M.
    - After the "\eob" output, a candidate angle is selected from M (removed from M, which we implement with a 0/1 mask), and the traversal continues along the new branch guided by the selected candidate angle. When outputting the first token of the new branch, the information of the selected candidate angle is synchronized.
    - If M is empty after the output of "\eob", the traversal ends.
(5) New processing approach for reconnection: When an atom A that is connected to a previously traversed atom B (reconnection), there must be branch angles attached to both of them. The bond between A and B is identified, and the corresponding branch angle attached to them is determined. The determined branch angle is removed from M, and the information of the this branch angle and the identified bond type is synchronized.

Please refer to Fig. 3(c) and (d) for a specific comparison between SD-SSML and RCDG-SSML.

3.2.3 The Conditional Attention Guidance Mechanism. The Conditional Attention Guidance Mechanism leverages the natural graph structure of molecular structural formulas and treats their recognition process as a graph traversal problem. As the model traverses the graph, it encounters multiple branch angle units, and the order of these angles in the proposed modeling units follows a fixed counterclockwise direction. However, if decoding is done solely based on the fixed angle order, the model may "forget" which angle units have not yet been decoded due to the prolonged decoding step in later stages. To address this issue, we propose to use angle directions as conditional information to guide the decoding process. When the model encounters branch, it continues to decode along the specified angle direction. The updated approach eliminates the use of "(" and ")" to indicate the start and end of a branch. Instead, we first predict "\angle[:<angle value>]" for each branch, individually storing the context and attention weight information computed during the prediction in the memory module. When "\eob" is decoded, indicating that there are no additional branch angles to predict, we select the angle state information from the memory as the condition to continue the decoding process with Attention. The following calculation process is employed:

$$
\begin{aligned}
e_{t,i} &= w^T \tanh(W^x x_i + W^y E y_{t-1} \\
&+ W^s s_{t-1} + (W^\alpha * [\alpha_{t-1}; \sum_{j}^{t-1} \alpha_j])_i \\
&+ W^{sp} s_b + (W^{\alpha p} * \alpha_b)_i)
\end{aligned} \tag{5}
$$

Here, α_b and s_b represent the attention weight information and state features of the corresponding decoding branch. It is worth emphasizing that when there is no branch angle to decode, α_b and s_b are both zero vectors, and the calculation method of the original attention structure is maintained. After decoding each branch angle, the corresponding angle state information in the memory will be consumed. When "\eob" is decoded and the memory is empty, the decoding process terminates.

3.2.4 Memory Classification Mechanism. The main difference between the graph structure and the tree structure is that the graph structure has reconnection characteristics. In this scheme, the angle corresponding to the reconnection has already been stored in the memory and has not yet been decoded. Therefore, we propose to build a simple multi-label classification module to determine the corresponding direction of the reconnection angle and to simultaneously classify the type of the reconnection bond (such as single bond, double bond, etc.). If N represents the number of bond types, then the calculation formula is as follows:

$$
\begin{aligned}
q_{tb} &= \text{softmax}(W^m s_b + W^o s_t) \\
L_{bc} &= -\sum_{b=1}^{B} \sum_{i=1}^{N+1} z_{tbi} \log(q_{tbi})
\end{aligned} \tag{6}
$$

In this context, s_t is the state of the decoding process when branch angles are decoded. W^m and W^o parameters for bond classification are stored in $\mathbb{R}^{(N+1) \times d_s}$. The probability distribution of bond classification between the state feature stored in the memory and the t-th decoding step is represented as $q_{tb} \in \mathbb{R}^{N+1}$. We also add an additional category to represent the option of no bond

connection for the stored state feature, and z_{tb} represents the corresponding one-hot classification label. B denotes the number of remaining branch angles in the memory that have not been decoded yet. L_{bc} represents the loss that results from reconnection. Generally speaking, if a branch angle is decoded in the current decoding step, the current step's state information and all remaining decoding branch angles in memory need to be classified together. The classification result is either a certain type of chemical bond or empty. If it is empty, it indicates that the branch angle does not connect with the remaining decoding branch angles.

3.2.5 Path Selection Mechanism. The recognition process of RCGD can be regarded as a graph traversal problem, where different traversals of a graph can yield different sequences. Although the conditional attention guidance mechanism decodes according to a fixed counterclockwise order, it can cause overfitting and recognition errors in complex or uncommon structures. Thus, we have proposed a path selection mechanism that randomly samples different paths during the training process to improve the alignment between visual information and decoded characters. During the inference process, we enable the model to attempt decoding all candidate branch angles stored in memory and participate in calculating the beam search path score. This allows automatic selection of the path with the highest score for continued decoding.

In summary, our RCGD has two loss functions: modeling unit classification loss L_{ce} and memory classification loss L_{bc}. The overall loss function is the sum of their individual components.

4 DATA SETS

We evaluate the performance of our proposed method on handwritten and printed data sets. Most available printed data sets consist of synthetic images with clean backgrounds. For the printed scenario, we use the publicly accessible CASIA-CSDB [5] data set as a benchmark. Unfortunately, in the case of handwritten scenarios, no publicly accessible handwritten molecular structure data sets currently exist. Therefore, we establish our own handwritten data set obtained mainly from real-world handwritten molecular structures in an educational setting. This data set includes numerous instances with writing errors and nonexistent structural data. A detailed introduction to both of these data sets can be found below.

The CASIA-CSDB data set is presently the largest publicly accessible printed molecular structure image data based on the chemical database ChEMBL [7], which stores data in SMILES string format. The RDKit software is employed to produce 480,668 samples which called CASIA-CSDB. The Mini-CASIA-CSDB subset includes 97,309 samples, and specific information on the data partitions is presented in Table 1. However, the images provided with the CASIA-CSDB data set have low rendering resolution (300×300), causing significant blurring, many samples even incomprehensible to humans. To overcome these limitations, we increase the rendering resolution to 500×500 and train our model using these high resolution RGB images as shown in Fig. 4. Due to limited computational resources, we conduct our experiments only the Mini-CASIA-CSDB.

We establish a handwritten data set named EDU-CHEMC, which consists of totally 52,987 handwritten molecular structure images collected in educational scenarios. The images were obtained using various devices such as cameras, scanners, and screens and are labeled as native Chemfig strings. To promote research related to the recognition of handwritten chemical structures, we plan to release this dataset in the near future. The key characteristics of this data set are as follows:

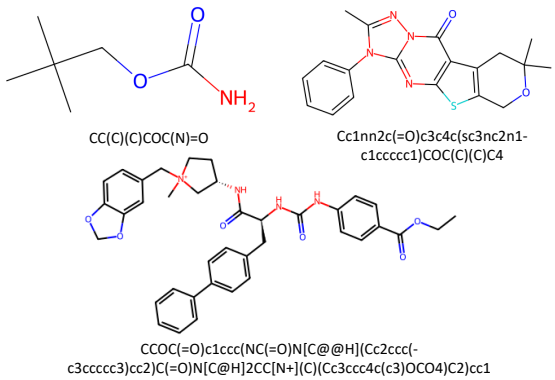

Figure 4: The high-resolution CASIA-CSDB images and its SMILES strings.

- Real-world educational scenarios. The data consists of handwritten molecular structures from primary and secondary education scenarios, which to our knowledge has never been made public before. Additionally, a small proportion of data is people copying and photographing ChEMBL molecular structures, also under real settings.
- Mixture of molecular structures and regular formulas. In addition to isolated isolated molecular structures, many instances of the data contain combinations of formulas and molecular structures, such as organic reaction equations as shown in Fig. 5.
- Writing diversity. The data exhibits various styles of structure writing, such as the use/non-use of abbreviations, the type of Kekule ring notation for benzene, and the inclusion/exclusion of hydrogen atoms. Moreover, the data contains numerous instances of artificially-written erroneous and even nonexistent structures that violate chemical principles as shown in Fig. 6, and the recognition of such structures can potentially be applied in correcting and revising handwritten answers.
- Complex structures. To test the model's generalization performance, the molecular structure complexity level is defined as the number summation of atoms and bonds, with around 10% of the test set featuring a complexity level exceeding that of the most complex sample in the training set. The complexity distribution of our dataset is shown in Fig. 7. The data partitions of our EDU-CHEMC data set are shown in Table 1.

Table 1: The partition details of Mini-CASIA-CSDB and our proposed EDU-CHEMC.

Data sets	Training set	validation set	Test set
Mini-CASIA-CSDB	80781	8242	8286
EDU-CHEMC	48998	999	2992

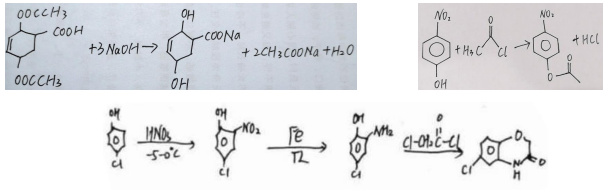

Figure 5: Examples of images that contain mixture of molecular structures and regular formulas in our handwritten data set.

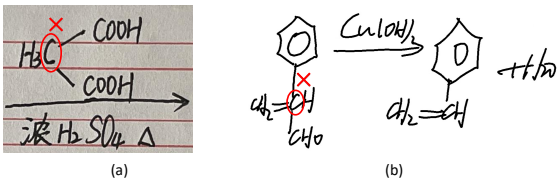

Figure 6: Examples of images that encounter chemical erroneous in our handwritten data set. In both (a) and (b), the chemical bond connections of the carbon atoms marked by the red circle violate the valence principle.

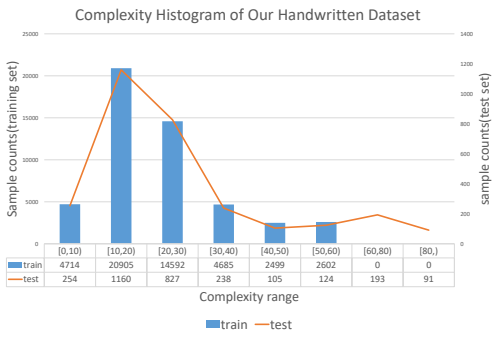

Figure 7: Complexity histogram of our handwritten data set. We can see that our testing set contain samples whose complexity level is never seen in training set.

5 EXPERIMENTS

5.1 Evaluation Metrics

Exact Match (EM). Similar to handwritten mathematical expression recognition tasks, we use the EM score as the primary evaluation criterion. Specifically, EM represents a full match between the predicted and labeled strings. For the SMILES modeling approach, since the original labels in the CASIA-CSDB data set have been standardized and the SMILES strings are concise enough, we directly compare whether the predicted and labeled strings are the same. For the Chemfig modeling approach, due to the polysemous nature of the labels, we first convert both the labels and recognition strings into Graphs and then compare if their Graph forms match. Let T denote the number of samples and R denote the number of predicted results that match the labeled results, then EM can be calculated using the following formula:

$$EM = \frac{R}{T} \tag{7}$$

In addition, for our EDU-CHEMC data, since it contains mixtures of formulas and molecular structures (see Fig. 5), a single image may contain multiple molecular structures. Therefore, we define two auxiliary metrics as follows:

Structure Exact Match (Structure EM). For samples that contain mixed molecular structure and regular formulas, when all the molecular structure recognition results match the labeled Graphs, we consider the sample to be "correctly recognized for the structure". Let T denote the number of samples and R_{struct} denote the number of samples with correctly recognized structures, then:

$$EM = \frac{R_{struct}}{T} \tag{8}$$

Structure EM and Single EM measure the model's recognition performance for the only molecular structures in the mixed-mode data.

5.2 Implementation Details

The base approach used in this paper is an end-to-end recognition framework based on the attention mechanism, which only uses conventional cross-entropy loss to optimize the model and does not include any additional loss functions. In RCGD, an extra memory classification loss L_{bc} is introduced.

The Encoder network used in this paper is DenseNet, which includes three dense blocks that convert the input RGB three-channel image into high-dimensional features. The growth rate and depth in each dense block are set to 24 and 32, respectively, which is exactly the same as the Encoder configuration in DenseWAP.

The String Decoder(SD) and Random Conditional Guided Decoder(RCGD) used in this paper both employ a GRU with a hidden state dimension of 256 as the recurrent unit of the RNN, and the attention projection dimension is set to 128. In addition, the embedding dimension is set to 256 and a dropout rate of 15% is applied. For the RCGD Decoder, the projection dimension for memory classification is set to 256.

The optimizer used in this paper is Adam, with an initial learning rate of 2e-4 and a learning rate decay strategy of multi-step decay, using Pytorch's MultiStepLR to adjust the learning rate, and a decay factor of gamma set to 0.5. In the Mini-CASIA-CSDB data, we use milestones of [40,60,70,75,80,85,...], and in our EDU-CHEMC data, we use milestones of [60,90,105,115,120,125,...]. We use teacher-forcing to calculate the character accuracy on the validation set and select the model with the highest character ACC for testing.

As for the comparison methods, since the data sets and evaluation criteria in this field are not fully standardized and the data sets is relatively new, there are limited comparison methods that can be used. We choose the two latest works in formula recognition, BTTR

Table 2: Evaluation on Mini-CASIA-CSDB data set. SD means String Decoder, RCGD means Random Conditional Guided Decoder, SSML means Structure-specific Markup Language, * means our implementation

Method	Resolution	Markup	Struct	EM
SD (DenseWAP)	300*300	SMILES	/	75.64%
SD (BTTR)	300*300	SMILES	/	78.22%
SD (WYGIWYS)	300*300	SMILES	/	78.55%
SD (DenseWAP*)	500*500	SMILES	/	81.89%
SD (DenseWAP*)	500*500	SSML	92.47%	92.09%
RCGD	500*500	SSML	95.38%	95.01%

[44] and ABM [1], as the comparison methods. The main idea of these two methods is to supervise training with R2L reverse decoding loss, which enables the decoder to have more powerful language modeling capabilities. We re-implement these two methods based on their ideas to fit the chemical structural formula recognition task in this paper. For BTTR, we only use its L2R decoding result for testing, and for ABM, consistent with the original paper, we only use L2R decoder for inference.

5.3 Experiments Results

Experiments on Mini-CASIA-CSDB. We first compare our proposed approach with the SMILES-based approach on the Mini-CASIA-CSDB data set, as shown in Table 2. It can be seen from the experimental results that our Chemfig-based SSML improve recognition performance significantly compared to the SMILES-based approach. This is mainly due to the reduced polysemy of our proposed modeling units, as well as the stronger consistency between the images and labels. Additionally, the RCGD proposed in this paper has a significant advantage over the String Decoder. It should be emphasized that the computational and parametric costs of the RCGD are almost the same as those of the String Decoder.

Table 3: Evaluation on rotated images of Mini-CASIA-CSDB. SD means String Decoder, RCGD means Random Conditional Guided Decoder, SSML means Structure-specific Markup Language

Method	Markup	Normal		Rotated	
		Struct	EM	Struct	EM
SD	SMILES	/	81.89%	/	71.53%
SD	SSML	92.47%	92.09%	90.27%	89.90%
RCGD	SSML	95.38%	95.01%	94.53%	94.16%

Rotation evaluation on Mini-CASIA-CSDB. As we know, the molecular structure expressed by the structural formula remains unchanged regardless of how it is rotated on a 2D plane. To further demonstrate the superiority of the proposed modeling units and modeling approach in terms of image-text consistency, we perform tests on the Mini-CASIA-CSDB test dataset with the images rotated 180 degrees. To avoid the characters being flipped and affecting the recognition results, we rerender the SMILES using the RDKit tool, controlling its output orientation to be 180 degrees different from the standard orientation. The test results are shown in Table 3. When using SMILES markup string the recognition performance drops by up to 10.36% on the rotated test set. In contrast, when using the proposed SSML, the string decode(SD) performance drop is only 2.19%, and the RCGD drop is only 0.85%.

Table 4: Evaluation on EDU-CHEMC. SD means String Decoder, RCGD means Random Conditional Guided Decoder, SSML means Structure-specific Markup Language, * means our implementation

Method	Markup	Struct	EM
SD (BTTR)	SSML	66.83%	58.21%
SD (ABM)	SSML	67.24%	58.78%
SD (DenseWAP*)	SSML	69.68%	61.35%
RCGD	SSML	71.88%	62.86%

Experiments on EDU-CHEMC. Furthermore, we also validate the superiority of the Random Conditional Graph Decoder proposed in this paper on the Handwritten Molecular Structure data set (EDU-CHEMC). As shown in Table 4, the Random Conditional Graph Decoder also outperforms the String Decoder on this handwritten task. It is worth noting that both ABM and BTTR employ reverse training modes to assist forward model training, benefiting greatly in regular formula recognition tasks but not as much in molecular structure recognition tasks as unidirectional training with DenseWAP. In fact, the main reason for this difference is that regular formulas are consistent in their monotonicity, making it easier for the model to determine the starting and ending positions. In contrast, molecular structures are naturally graph-structured, which makes it difficult to locate their ending positions. Thus, the reverse pattern of BTTR and ABM cannot be well-trained and may have a negative impact on the forward decoding mode. However, the RCGD employed in this paper adopts a conditional guided mechanism and a path selection mechanism to make full use of multiple traverse paths for training, making it easier to align decoding paths with visual features.

Due to the page limit, we just put additional contents including more analysis and visualization results in the supplementary materials.

6 CONCLUSION

The End-to-end recognition methods have shown high efficiency in handwritten mathematical expression recognition, but have not been fully utilized in the field of handwritten chemical structure recognition due to the lack of suitable handwritten datasets and markup annotations. To address these challenges, we proposed a structure-specific markup language and RCGD algorithm in this paper. Through experiments on multiple publicly available datasets we achieved a huge improvement (95.01% vs 81.89%) on Mini-CASIA-CSDB data compared to the image-to-SMILES method. Our model currently supports some common chemical molecular image projections, such as Natta projection, Fischer projection, and Sawhorse projection, while we aim to expand its capabilities to support more projection layouts in the future. We hope this work will inspire further research on this important topic.

REFERENCES

[1] 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. *Proceedings of the AAAI Conference on Artificial Intelligence* (Jul 2022), 113–121. https://doi.org/10.1609/aaai.v36i1.19885

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 213–229.

[3] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling.

[4] Djork-Arné Clevert, Tuan Le, Robin Winter, and Floriane Montanari. 2021. Img2Mol–accurate SMILES recognition from molecular graphical depictions. *Chemical science* 12, 42 (2021), 14174–14181.

[5] Longfei Ding, Mengbiao Zhao, Fei Yin, Shuiling Zeng, and Cheng-Lin Liu. 2022. A Large-Scale Database for Chemical Structure Recognition and Preliminary Evaluation. In *ICPR*. 1464–1470.

[6] 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).

[7] Anna Gaulton, Anne Hersey, Michal Nowotka, A.Patrícia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, LouisaJ. Bellis, Elena Cibrian-Uhalte, and et al. 2017. The ChEMBL database in 2017. *Nucleic Acids Research* (Jan 2017).

[8] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. 2023. A Survey on Vision Transformer. *TPAMI* 45, 1 (2023), 87–110.

[9] Stephen Heller, Alan McNaught, Stephen Stein, Dmitrii Tchekhovskoi, and Igor Pletnev. 2013. InChI - the worldwide chemical structure identifier standard. *Journal of Cheminformatics* 5, 1 (Jan 2013). https://doi.org/10.1186/1758-2946-5-7

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[11] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. 2022. Restr: Convolution-free referring image segmentation using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18145–18154.

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643* (2023).

[13] Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, and et al. 2011. DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Research* (2011).

[14] 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 *ECCV*, Vol. 13688. 197–214.

[15] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. 2022. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13619–13627.

[16] Qingliang Li, Tiejun Cheng, Yanli Wang, and StephenH. Bryant. 2010. PubChem as a public resource for drug discovery. *Drug Discovery Today* (Dec 2010).

[17] Sun-Ao Liu, Hongtao Xie, Hai Xu, Yongdong Zhang, and Qi Tian. 2022. Partial Class Activation Attention for Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16836–16845.

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[20] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. 2021. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3651–3660.

[21] Eugenia Namiot. 2019. Using LaTeX for chemical formulas.

[22] Martijn Oldenhof, Adam Arany, Yves Moreau, and Jaak Simm. 2020. ChemGrapher: Optical Graph Recognition of Chemical Compounds by Deep Learning. *Journal of Chemical Information and Modeling* 60, 10 (2020), 4506–4517.

[23] Yujie Qian, Jiang Guo, Zhengkai Tu, Zhening Li, ConnorW. Coley, and Regina Barzilay. 2023. MolScribe: Robust Molecular Structure Recognition with Image-To-Graph Generation. *Journal of Chemical Information and Modeling* 63, 7 (2023), 1925–1934.

[24] Kohulan Rajan, HenningOtto Brinkhaus, Achim Zielesny, and Christoph Steinbeck. 2020. A review of optical chemical structure recognition tools. *Journal of Cheminformatics* 12, 1 (2020), 60.

[25] Joshua Staker, Kyle Marshall, Robert Abel, and CarolynM. McQuaw. 2019. Molecular Structure Extraction From Documents Using Deep Learning. *Journal of Chemical Information and Modeling* 59, 3 (2019), 1017–1029.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NeurIPS*. 5998–6008.

[27] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. 2022. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*. PMLR, 1475–1485.

[28] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. 2021. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 913–922.

[29] Xuehui Wang, Kai Zhao, Ruixin Zhang, Shouhong Ding, Yan Wang, and Wei Shen. 2022. Contrastmask: Contrastive learning to segment every thing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11604–11613.

[30] David Weininger. 1988. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling* (Feb 1988).

[31] Robin Winter, Floriane Montanari, Frank Noe, and Djork-Arne Clevert. 2019. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical Science* 10, 6 (2019), 1692–1701.

[32] 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.

[33] Jiajia Wu, Jinshui Hu, Mingjun Chen, Lirong Dai, Xuejing Niu, and Ning Wang. 2022. Structural String Decoder for Handwritten Mathematical Expression Recognition. In *ICPR*. 3246–3251.

[34] Jin-Wen Wu, Fei Yin, Yan-Ming Zhang, Xu-Yao Zhang, and Cheng-Lin Liu. 2020. Handwritten Mathematical Expression Recognition via Paired Adversarial Learning. *IJCV* (2020).

[35] Youjun Xu, Jinchuan Xiao, Chia-Han Chou, Jianhang Zhang, Jintao Zhu, Qiwan Hu, Hemin Li, Ningsheng Han, Bingyu Liu, Shuaipeng Zhang, Jinyu Han, Zhen Zhang, Shuhao Zhang, Weilin Zhang, Luhua Lai, and Jianfeng Pei. 2022. MolMiner: You only look once for chemical structure recognition. *Journal of Chemical Information and Modeling* 62, 22 (2022), 5321–5328.

[36] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. 2021. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11784–11793.

[37] Yuhui Yuan, Xilin Chen, and Jingdong Wang. 2019. Object-Contextual Representations for Semantic Segmentation. *arXiv preprint arXiv:1909.11065* (2019).

[38] 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 *CVPR*. 4543–4552.

[39] 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.

[40] Xiangxiang Zeng, Hongxin Xiang, Linhui Yu, Jianmin Wang, Kenli Li, Ruth Nussinov, and Feixiong Cheng. 2022. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence* 4, 11 (2022), 1004–1016.

[41] 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.

[42] 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. *PR* (2017).

[43] Jiaming Zhang, Kailun Yang, Chaoxiang Ma, Simon Reiß, Kunyu Peng, and Rainer Stiefelhagen. 2022. Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16917–16927.

[44] Wenqi Zhao, Liangcai Gao, Zuoyu Yan, Shuai Peng, Lin Du, and Ziyin Zhang. 2021. Handwritten Mathematical Expression Recognition with Bidirectionally Trained Transformer. *Springer International Publishing eBooks* (Sep 2021).

[45] Shuhan Zhong, Sizhe Song, Guanyao Li, and S-H Gary Chan. 2022. A tree-based structure-aware transformer decoder for image-to-markup generation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 5751–5760.

[46] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. 2022. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2582–2593.

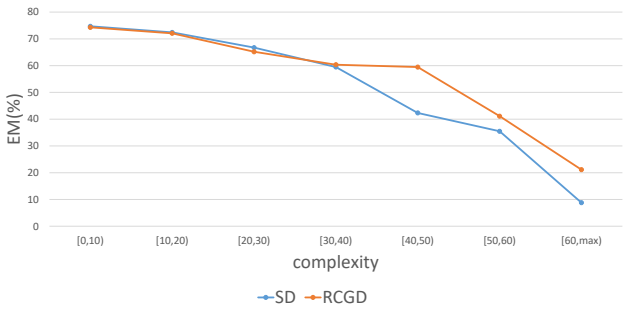

Figure 8: Illustration of the comparison of performance between SD and RCGD under different Complexity levels.

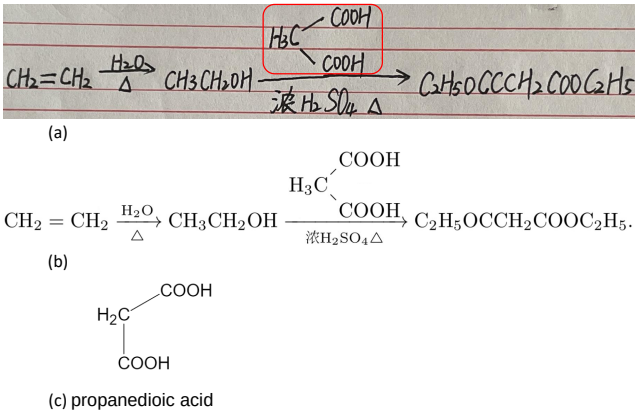

Figure 9: Illustration of a example for Invalid or Non-Existent Organic Molecular Structures. Figure (a) is the original image, the handwritten parts in red boxes contains errors in chemical principles. Figure (b) is the rendered images of model outputs. Figure (c) is the corresponding correct forms and molecule names.

A APPENDIX

In this appendix material, we present additional results on our method, along with visual comparisons. In the ablation study section, we will provide further details and results on the efficacy of our three mechanisms of RCGD.

A.1 The Generalization ability to Structure Complexity Levels

We define molecular complexity as the number of atoms and bonds in a chemical molecular structure. On EDU-CHEMC dataset, as Fig. 8 illustrated, the performance of SD and RCGD is comparable when the complexity is below 40. However, when the molecular complexity exceeds 40, the performance of RCGD is significantly better than that of SD.

A.2 The Advantages of Structure-Specific Markup

Our method can also recognize invalid or non-existent organic molecular structures, which makes it highly applicable for tasks related to teaching evaluations in the fields of chemistry and biology education. As Fig. 9(a-c) the handwritten version of propanedioic acid illustrated, the "C" atom in the "H3C" group is connected to 5 single bonds, which violates the principle that the valence of "C" atoms is 4. In the correct structure of propanedioic acid, the fragment should be "H2C". However, our recognition result is "H3C", which is consistent with the handwritten image. So when encountering incorrect handwritten molecules, our model is able to identify errors, thus enabling the possibility of correction and revision.

A.3 More Ablation Studys

The Table 5 shows that the removal of any of these mechanisms results in a decrease in performance to varying degrees. Notably, the removal of the memory classification mechanism leads to a more significant decline in performance. This is because, as the molecular structures become more complex, the reconnection processing becomes more complicated.

A.3.1 The Conditional Attention Guidance Mechanism. The role of the conditional attention guidance mechanism is to avoid the model's inability to accurately determine the location of the undecoded branch due to excessively long decoding steps. The Fig. 11 illustrates this concept. From the red area in Fig. 11(b), it can be seen that SD made a mistake in judging the corresponding angle branch to be decoded when decoding "= O", resulting in a dislocation problem. However, as illustrated in Fig. 11(c) the RCGD can find the correct branch to be decoded by using the state and attention position information stored in the memory module.

Table 5: Results of RCGD without Path Selection (PS) and Memory Classification (MC).

Method	Struct	EM
RCGD	71.88%	62.86%
RCGD w/o PS	70.85%	62.15%
RCGD w/o (PS&MC)	68.84%	60.31%

A.3.2 Memory Classification Mechanism. In native Chemfig, "?[a]" and "?[a,<bond>]" are two modeling units used to represent reconnection. This makes it difficult for the model to accurately determine the "bond angle" unit. The memory classification mechanism is more reasonably model the reconnection in the graph structure as illustrated in Fig. 10.

A.3.3 Path Selection Mechanism. The path selection mechanism may randomly choose the next decoding condition to continue decoding during training, which prevent overfitting and the generalization of the model is improved.

As shown in the Fig. 12(c), we can observe that after removing the path selection mechanism, the attention map may produce multiple peaks during decoding. Fig. 12(d) shows the accumulated attention visualization results without the path selection mechanism. In Fig. 12(e), the bold arrow represents the current decoded unit. Although the bond type predicted by the current decoding

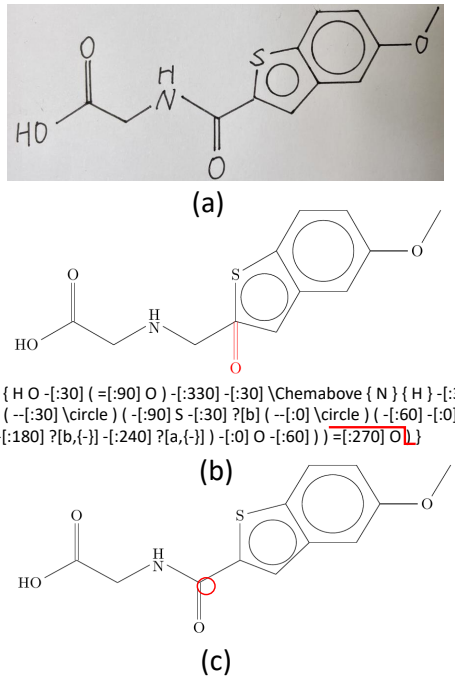

Figure 11: Illustration of result of conditional attention guidance mechanism. (a) Image of the molecular structure to be recognized. (b) The decoding string of SD and its corresponding rendered result image. (c) The rendered image of RCGD recognition result.

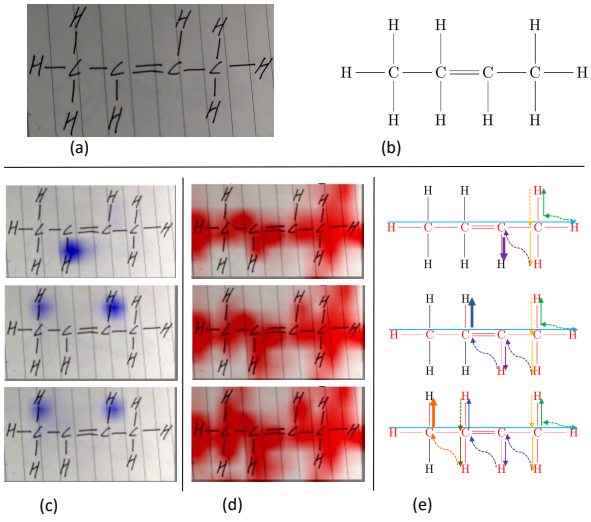

Figure 12: Illustration of decoding when without the path selection mechanism. (a) The image of a molecular structure image to be recognized. (b) displays the recognition result obtained after removing the path selection mechanism.

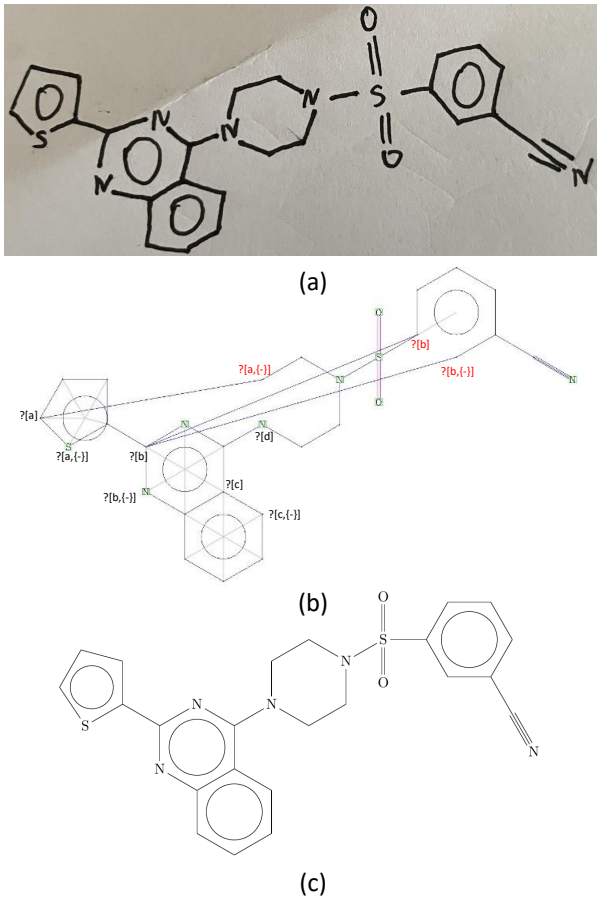

Figure 10: Illustration of result of memory classification mechanism . (a) The image of a molecular structure to be recognized. (b) The recognition result after removing both the path selection mechanism and the memory classification mechanism. (c) The render image of the recognition result without path selection mechanism.

step is correct, the position of the connecting "C" atom is incorrect, leading to severe location confusion when decoding hydrogen atoms and their connecting bonds in the subsequent steps. This issue mainly stems from the existence of numerous similar branching structures in the molecular structure. During training, the model did not perform random path selection, leading the model to fail to fully utilize the position information contained in the unexplored branch angles and instead relied only on the semantic information and local visual information. As a result, when encountering these similar branching structures, the attention probability distribution was more dispersed, leading to decoding errors.