Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MolTabReco: Table-Coordinate-Grounded Chemical Table Recognition with SMILES Recovery for Molecular-Structure Cells

Wei Hu
Hunan University of Chinese Medicine
Wei Liu
weiliu@hnucm.edu.cn
Hunan University of Chinese Medicine
Yuanjie Xiang
Hunan University of Chinese Medicine
Jiawei Huang
Hunan University of Chinese Medicine
Mei Ouyang
Hunan University of Chinese Medicine
Jiajun Tao
Hunan University of Chinese Medicine
Yao Song
Hunan University of Chinese Medicine

Research Article

Keywords: chemical table recognition, molecular-structure cell, SMILES, heterogeneous cell discrimination, MolScribe, ChemTable dataset, chemical information extraction, MolTabReco

Posted Date: August 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9870580/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MolTabReco: Table-Coordinate-Grounded Chemical Table Recognition with SMILES Recovery for Molecular-Structure Cells

Wei Hu, Wei Liu, Yuanjie Xiang, Jiawei Huang, Mei Ouyang, Jiajun Tao, Yao Song*
School of Informatics, Hunan University of Chinese Medicine, Changsha, China
*Correspondence: Wei Liu, School of Informatics, Hunan University of Chinese Medicine, Changsha, China. Email: weiliu@hnucm.edu.cn

Abstract

Chemical literature tables are an important source for chemical knowledge extraction and structured data. They contain heterogeneous text, numerical values, chemical formulas, footnotes, and molecular-structure depictions. Existing optical character recognition (OCR) models, general table recognition models, and vision-language models (VLMs) can recover part of row-column structure and text; for molecular-structure cells, however, they often return placeholders, empty contents, or weak descriptions, making it difficult to obtain computable Simplified Molecular Input Line Entry System (SMILES) representations. This study addresses row-column recognition, cell-text recovery, and molecular-structure depiction recognition by developing a heterogeneous cell recovery framework under table-coordinate constraints. We propose MolTabReco, a multi-stage framework that localizes the table region, uses VLM-predicted row-column structure and text results, and combines them with a structure detection model to generate cell locations and a unified table coordinate system. A molecular-structure cell discriminator routes each cell to a text recovery branch or an optical chemical structure recognition (OCSR)-SMILES branch. Non-structure cells use conservative VLM-first fusion; structure cells use a MolScribe-led OCSR-SMILES strategy with crop expansion, image enhancement, foreground segmentation, RDKit validity checking, canonicalization, and fragment filtering. Large language model (LLM)-based correction is restricted to controlled review for low-confidence or abnormal cells. On ChemTable under a shared VLM-predicted grid, MolTabReco achieved 44.05% canonical exact match (EM) on 1,682 molecular-structure cells. On dimension-matched pages, 1,159 structure cells reached 54.79% canonical EM, 60.31% no-stereochemistry EM, and 94.48% valid SMILES recognition. Among 1,095 parseable predictions, 65.8% had Tanimoto similarity ≥0.9 and 77.5% had similarity ≥0.5. The VLM baseline could not convert depictions into SMILES; its 8.09% score corresponded to placeholder or weak-content matching. Because structure cells are rare, Overall Value Accuracy increased from 82.53% to 83.29%, and Tree Edit Distance based Similarity from 0.8537 to 0.8595. MolTabReco converts structure cells that would otherwise remain at the placeholder level into canonical SMILES linked to row-column positions.

Scientific Contribution

MolTabReco extends chemical table recognition beyond layout and text extraction by jointly recovering molecular-structure cells as canonical SMILES under explicit table-coordinate constraints. The framework integrates table-region localization, VLM-predicted row-column grounding, structure-cell routing, and OCSR-based SMILES recovery so that chemical structures are linked to their original row-column context. This provides a reproducible pathway for converting molecular-structure cells that would otherwise remain empty or as image placeholders into computable chemical data.

Keywords: chemical table recognition; molecular-structure cell; SMILES; heterogeneous cell discrimination; MolScribe; ChemTable dataset; chemical information extraction; MolTabReco

1 Introduction

Chemical literature tables contain information on reaction conditions, molecular structures, catalyst performance, and related experimental variables. Unlike ordinary document tables, chemical tables often contain text, numerical values, superscripts and subscripts, footnotes, and two-dimensional molecular-structure depictions within the same layout. For cheminformatics tasks, recovering only the table layout or ordinary text is insufficient. Molecular-structure depictions in tables must also be recovered as computable representations such as SMILES to support molecular retrieval, structural similarity analysis, reaction-condition organization, and chemical database construction.

In recent years, general table recognition methods have advanced in table detection [1,2], structure recovery [3], and cell relationship parsing [4]. OCR models [5] and document parsing models [6,7] have also improved the ability to recognize document text and generate structured representations. At the same time, VLMs have shown strong layout understanding and contextual reasoning capabilities in complex document understanding and table question answering [8,9], and multimodal models in chemistry have begun to be applied to chemical image-text understanding tasks [10]. However, molecular-structure cells in chemical tables are not ordinary text cells. OCR models often recognize structure depictions as empty text or image tags, and VLMs also struggle to directly recover canonical SMILES from structure depictions. A VLM may identify that a molecular structure is likely present at a given position, but it usually cannot convert that depiction into a computable molecular representation, nor can it reliably determine the precise coordinate position of the recognized content.

Research in optical chemical structure recognition (OCSR) has produced models such as MolScribe [11], MolGrapher [12], MolNexTR [13], and DECIMER [14,15], which can recover molecular representations from two-dimensional structure depictions [16]. These methods provide an important foundation for recognizing molecular-structure depictions, but their default inputs are usually independent molecular images with complete boundaries and relatively low noise. In contrast, structure depictions in chemical tables are often smaller and may be affected by table ruling lines, adjacent text, low resolution, and crop displacement. Moreover, a standalone OCSR model can only predict the molecule corresponding to an image and cannot predict the specific row and column in which the molecule is located. Therefore, molecular-structure recovery in chemical tables cannot be treated simply as an isolated OCSR task. It should instead be regarded as a hybrid table task involving table localization, cell-type discrimination, and molecular-structure recovery.

The ChemTable dataset provides a data foundation for studying real chemical literature tables. Its page-level tables, cell layouts, and domain annotations make it possible to evaluate the full workflow from page image to table coordinates, cell content, molecular-structure depictions, and SMILES [17]. Existing studies, however, have mainly focused on table structure recovery, text recognition, or standalone structure depiction recognition, while the computable recovery of structure cells inside tables has received limited attention. To address this problem, we propose MolTabReco, a multi-stage heterogeneous cell recovery framework for chemical literature tables. Starting from a page image, the framework automatically performs table-region localization, VLM-based row-column grid prediction, table-cell position detection, cell cropping, molecular-structure cell discrimination, text recovery, and molecular-structure-to-SMILES recognition. In this pipeline, the VLM serves as the source of the table skeleton and initial text recognition. MolTabReco further introduces a structure-cell discriminator and a MolScribe-led OCSR-SMILES recovery branch to convert structure placeholders into canonical SMILES linked to row-column positions. In the post-processing stage, the LLM is used only as a restricted review tool for low-confidence, format-

abnormal, or contextually inconsistent cells, and the risk of erroneous free-form correction is specifically evaluated.

The main contributions of this work are as follows.

First, we propose a multi-stage framework for heterogeneous cell recovery in chemical tables. This work extends chemical literature table recognition from table structure and cell text recognition to heterogeneous cell recovery under table-coordinate constraints. Given a chemical paper page image, MolTabReco performs table-region localization, row-column grid prediction, cell bounding-box generation, cell-type discrimination, ordinary text recovery, and SMILES backfilling for structure cells. Unlike general table recognition methods that output only text, MolTabReco maps heterogeneous content types, including text, numerical values, and molecular-structure depictions, to specific row-column positions. The final structured table object preserves row-column position indices, cell bounding boxes, and content types. This design not only recovers the table content but also determines the position and index of each cell, thereby providing a table-coordinate foundation for downstream chemical structure retrieval, reaction-condition organization, and structured database construction.

Second, we develop a SMILES recovery method for small molecular-structure cells in chemical tables. To address the low proportion, small image size, thin line width, and susceptibility to interference from table lines and adjacent text in molecular-structure cells, we first use a structure-cell discriminator to determine whether a cell contains a molecular-structure depiction. Structure cells are then processed by a MolScribe-led OCSR-SMILES recovery branch combined with crop expansion, image enhancement, foreground segmentation, RDKit validity checking, SMILES canonicalization, and abnormal fragment filtering to improve the stability of small structure depiction recovery in table scenarios.

Third, we construct a VLM-first table-content recovery strategy and analyze the safety boundary of LLM-assisted review. We fully exploit the general image recognition capability of VLMs and use the VLM output as the primary source for table row-column structure and ordinary cell text. Because numerical values, units, superscripts and subscripts, footnotes, chemical abbreviations, and compound identifiers in chemical tables have explicit experimental meanings, we further analyze the restricted role of LLMs in low-confidence cells, multi-source output conflicts, format abnormalities, and contextual inconsistencies. The experimental results show that free-form generative whole-table correction can introduce correct-to-wrong errors. Therefore, this study does not use free-form LLM correction as a source of main-process performance. Instead, it positions LLM correction as an auxiliary review mechanism that must be rule-validated, traceable, and reversible.

2 Background and related work

2.1 Table detection and structure recognition in scientific documents

Table detection and table structure recognition are fundamental tasks in visual document understanding. They typically include table-region localization, row-column structure recovery, cell content recognition, and reconstruction of structured representations such as HTML. Earlier approaches relied mainly on line detection, connected-component analysis, and rule-based templates. More recently, convolutional neural networks (CNNs), Faster R-CNN [18], DEtection TRansformer (DETR) [19], Transformers, and sequence generation methods have been widely used for this task. Datasets and frameworks such as PubTables-1M [1], TableBank [2], PubTabNet [3], and Global Table Extractor (GTE) [4] have promoted the development of table structure recognition models. These methods are suitable for regular tables or ordinary document tables, but chemical literature tables still present challenges arising from molecular-structure depictions, complex headers, footnotes, and non-text cells.

2.2 OCR and vision-language models

OCR and end-to-end document parsing methods provide important tools for scientific document text recognition. TrOCR [5] improves character recognition by combining a visual Transformer with a text decoder. DONUT [6] performs document understanding in an end-to-end OCR-free manner, while Nougat [7] explores the conversion of academic document images into markup text. These methods are suitable for ordinary text or structured markup, but cell content in chemical tables is more position-dependent and domain-specific. The same numerical value may represent temperature, time, yield, or selectivity depending on its column. Chemical formulas, superscripts and subscripts, footnotes, units, and abbreviations can also introduce format noise. More importantly, the target output of a molecular-structure cell is not an ordinary text string but a computable chemical representation such as SMILES or a molecular graph. In recent years, VLMs have demonstrated strong layout understanding and contextual reasoning in image-text understanding, complex document parsing, and table question answering [8,9], and multimodal models in chemistry have begun to be used for chemical image-text understanding [10]. In chemical tables, VLMs are well suited as a source of table skeletons and initial text, and they can handle complex headers, row and column spans, and ordinary text relatively well. However, a VLM is not a complete chemical structure recognizer. It usually only detects the presence of a structure depiction or outputs placeholders such as <img> or an empty string, and it cannot stably generate canonical SMILES. In this work, the VLM is treated as a source of table structure and context rather than as a molecular-structure recovery module. Generative models are suitable for text polishing and contextual completion, but numerical values, units, footnotes, superscripts and subscripts, and compound identifiers in chemical tables have explicit experimental meanings and cannot be freely rewritten on the basis of linguistic plausibility alone. Therefore, in chemical table recognition, an LLM is more appropriate as an auxiliary correction tool for low-confidence or format-abnormal cells than as a full-table rewriting module.

2.3 Optical chemical structure recognition

Optical chemical structure recognition aims to convert two-dimensional molecular-structure depictions into SMILES or molecular graphs and is a key technology for chemical literature mining and database construction. Earlier OCSR methods relied mainly on image processing, line detection, character recognition, and rule-based parsing. In recent years, deep learning methods such as MolScribe [11], MolGrapher [12], MolNexTR [13], and DECIMER [14,15] have improved molecular-structure recovery from the perspectives of image-to-graph prediction, sequence generation, and structure modeling. These studies demonstrate the feasibility of recovering computable chemical representations from two-dimensional structure depictions, but their inputs are typically independent structure depictions with complete boundaries and low noise. Structure cells in real chemical tables are often small, have thin lines, and may be affected by table ruling lines, adjacent text, compression noise, and crop displacement. More importantly, standalone OCSR models can identify the molecule represented by a chemical image, but they cannot provide the molecule's row-column position in the table. Therefore, structure recovery inside tables must be integrated with the table coordinate system.

2.4 Chemical information extraction and chemical table datasets

Chemical information extraction covers entity recognition, reaction extraction, experimental-condition extraction, table extraction, structure depiction recognition, and database construction. Resources such as PubChem and ChEMBL illustrate the value of structured chemical data for molecular retrieval, property analysis, and drug discovery [20,21]. However, a large amount of chemical knowledge initially appears in article text, tables, diagrams, reaction schemes, and supplementary materials. Converting this knowledge into computable data requires the joint processing of text, layout, images, and molecular

representations. ChemTables focuses on semantic classification of tables in chemical patents [22], whereas the ChemTable dataset focuses on real chemical literature tables and provides annotations for cell layout, logical structure, domain labels, molecular-structure depictions, and table question answering [17]. Unlike datasets that focus only on table classification or ordinary text extraction, ChemTable enables researchers to evaluate whether structure cells are localized, whether they are bound to the correct row-column positions, and whether they are further recovered as SMILES. This study builds MolTabReco on ChemTable with the goal of connecting general table recognition, VLM-based document understanding, and dedicated OCSR models into a system framework for computable recovery of chemical tables.

3 Materials and methods

This section introduces the MolTabReco framework from the perspectives of method design, dataset and preprocessing, evaluation methodology, and experimental settings. Unlike pure OCR or HTML reconstruction, MolTabReco must recover not only the table structure and ordinary text, but also identify molecular-structure cells in the table and convert them into computable canonical SMILES.

3.1 Overall workflow of MolTabReco

MolTabReco is a multi-stage heterogeneous cell recovery framework under coordinate constraints. Given a page image, the framework automatically performs table-region localization, table row-column grid prediction, cell position prediction, cell cropping, molecular-structure cell discrimination, text recognition, and molecular-structure recognition. The overall workflow of MolTabReco is shown in Figure 1. The framework first crops the main table region from the page to reduce interference from titles, captions, footnotes, and other page elements. It then uses a VLM to obtain an initial row-column grid and text results, and combines these results with Table Transformer to recognize row and column bounding boxes and generate a crop for each cell. On this basis, the structure-cell discriminator determines whether each cell contains a molecular-structure depiction. Non-structure cells enter the text recovery branch, which adopts a conservative VLM-first strategy with TrOCR as a fallback. Structure cells enter the OCSR-SMILES recovery branch, where crop expansion, image enhancement, MolScribe recognition, and RDKit canonicalization are used to obtain canonical SMILES. For cells with low confidence, multi-source conflicts, or format abnormalities, the system records the reason for triggering post-processing and uses the LLM only as a restricted review tool. Its output must pass rule checks for numerical values, units, symbols, and chemical formats before it can be accepted. Finally, the system outputs a recovered structured table in which ordinary text, numerical values, and SMILES all have explicit cell coordinates.

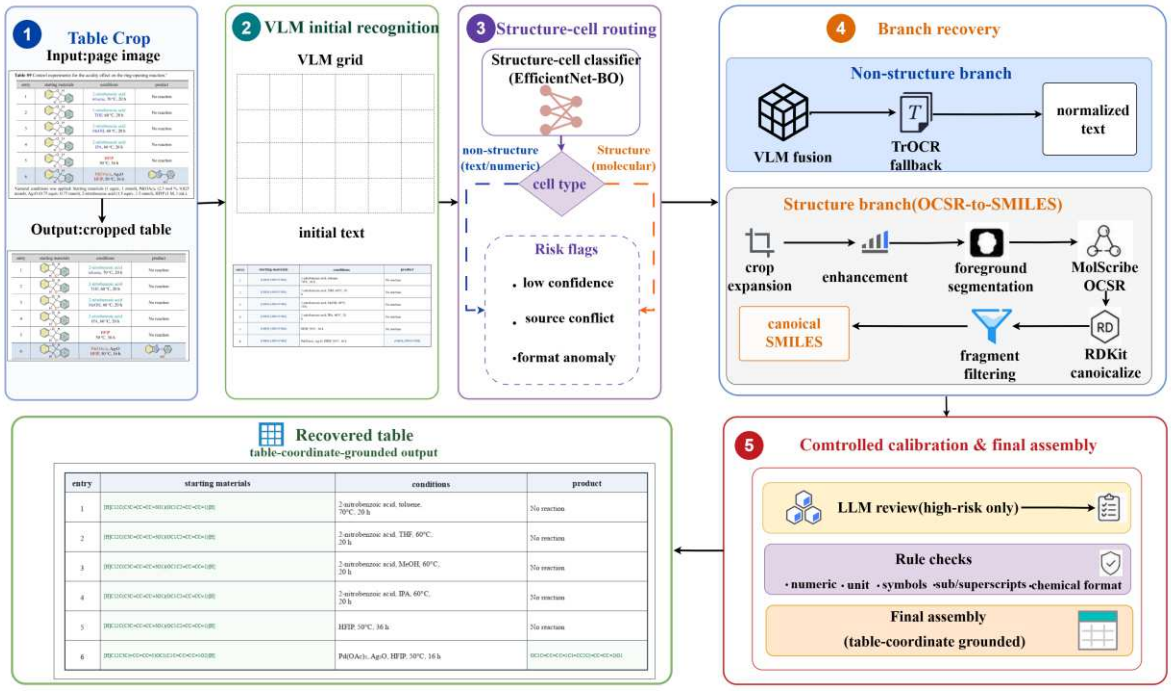


Figure 1. Overview of the MolTabReco framework for table-grounded molecular-structure cell recognition and SMILES recovery.

3.2 Table-coordinate localization and cell generation

The goal of the table-coordinate localization layer is to establish a unified coordinate system for subsequent cell discrimination and SMILES backfilling. For chemical tables, the impact of coordinate errors differs from that of ordinary OCR errors. Even if the text or SMILES itself is correctly recognized, a row-column position error will bind downstream text and molecular-structure recognition results to the wrong table context. Therefore, coordinate localization is the first constraint layer of MolTabReco.

3.2.1 Main-table region detection

Main-table region detection is the first step in the table-coordinate localization layer of MolTabReco. Its goal is to localize the main table region in a complete chemical paper page image and restrict subsequent structure parsing, cell cropping, and content recovery to the visual region associated with the target table. Compared with ordinary document tables, chemical paper pages may contain titles, reaction schemes, figure captions, footnotes, and paragraph text. Reaction schemes and molecular-structure depictions also visually contain regular lines, chemical symbols, and local grid-like arrangements, which can be mistakenly identified by structure parsers as table rows, columns, or cell boundaries. If the full page image is directly input into subsequent structure parsers, page-level interfering elements may cause errors such as incorrect row or column counts, over-extended bounding boxes, footnotes being merged into the table, or reaction schemes being treated as table content. These errors then affect structure-cell discrimination and SMILES backfilling. MolTabReco performs table-region detection on the page image and outputs the table bounding box. Table-region detection not only reduces interference from titles, reaction schemes, footnotes, and background page elements in row-column structure parsing, but also provides a stable local image coordinate system for later pixel-level row and column bounding-box decoding. In the implementation, this study uses a Table Transformer-based table detection model as the main-table region localization module and performs domain adaptation on ChemTable pages. The DETR-

style object detection framework of Table Transformer is suitable for document table boundary localization and can directly predict table regions from page-level inputs.

3.2.2 Row-column grid parsing

After the main table image is cropped, MolTabReco further recovers the logical row-column grid to establish the correspondence between cell contents and table coordinates. This step determines the number of rows and columns and the ordering of cells, forming the coordinate basis for subsequent cell cropping, type discrimination, and content backfilling. Even if the text or SMILES itself is correctly recognized, a row-column coordinate shift will bind the result to the wrong experimental condition or chemical entity. In this study, the row-column grid predicted by a VLM is used as the reference row-column structure. The VLM reads the main table image and outputs the row-column structure, cell order, and initial text content. MolTabReco then introduces molecular-structure cell discrimination, VLM text fusion, and the OCSR-SMILES recovery branch on the same predicted grid. Because both the VLM baseline and MolTabReco share the same automatically predicted row-column grid, row-column matching metrics in the system-level comparison mainly reflect the shared structure parsing quality, while the differences between MolTabReco and the VLM baseline mainly arise from subsequent content recovery, especially whether structure cells can be further recovered from empty contents or placeholders into canonical SMILES. It should be noted that the VLM output mainly provides the logical structure and contextual text but does not include precise pixel boundaries for each row and column, and therefore cannot be used directly for structure depiction cropping or OCSR input. To address this, MolTabReco combines a fine-tuned Table Transformer Structure Recognition model, which predicts row and column candidate boxes, with the target row and column counts provided by the VLM for count-constrained selection and spatial sorting. In this way, the logical grid is projected onto image coordinates. The VLM provides stable row-column counts and the table skeleton, whereas the structure detection model provides crop-able pixel-level boundaries. By combining the two outputs, the system generates reliable cell images and binds text, numerical values, and molecular-structure depictions to specific row-column positions, ultimately performing discrimination, recognition, and SMILES backfilling under table-coordinate constraints.

3.2.3 Row-column bounding-box decoding and cell cropping

The table structure output by the VLM can usually provide the number of rows, the number of columns, and the textual content, but it does not provide the image coordinates of each row and column; therefore, it cannot be directly used for cell cropping. To obtain cell images for subsequent structure-cell discrimination and OCSR-SMILES recovery, MolTabReco uses a fine-tuned Table Transformer Structure Recognition model to predict row bounding boxes and column bounding boxes, and combines these predictions with the row-column count constraints provided by the VLM to generate cell bounding boxes. Specifically, the structure detection model first outputs the row candidate set and column candidate set:

$$\mathcal{R} = \{(b_i^r, p_i^r)\}_{i=1}^{n_r}, \quad \mathcal{C} = \left\{\left(b_j^c, p_j^c\right)\right\}_{j=1}^{n_c} \qquad (1)$$

where $\mathcal{R}$ and $\mathcal{C}$ denote the row candidate set and the column candidate set, respectively; b_i^r denotes the i-th row candidate box and p_i^r its confidence score; b_j^c denotes the j-th column candidate box and p_j^c its confidence score; and n_r and n_c denote the numbers of row and column candidate boxes predicted by the model, respectively. The system then filters and sorts the candidate boxes in Equation (1) according to the target row and column counts given by the VLM. Row candidates are sorted by vertical position and column candidates by horizontal position. When the number of candidates exceeds the target number of rows or columns, candidates with higher confidence and reasonable spatial ordering are preferentially

retained. After row-column alignment, the bounding box of the cell in row i and column j is obtained from the intersection of the corresponding row and column regions:

$$b_{ij}^{cell} = b_i^r \cap b_j^c \qquad (2)$$

Equation (2) represents the generation of the cell bounding box: the overlapping image region between row i and column j is treated as the pixel extent of the cell. In this way, the row-column count and global structure information provided by the VLM are combined with the pixel-level boundary information provided by the structure detection model. The former constrains the number and spatial order of rows and columns, while the latter determines the exact crop position. In practice, the system also performs image-boundary clipping, minimum-size filtering, and controlled boundary expansion on the cell bounding boxes obtained from Equation (2), thereby avoiding crop regions that are out of bounds, too small, or include excessive content from adjacent cells. The resulting cell crops are used as common inputs for structure-cell discrimination, supplementary recognition of non-structure text, and OCSR-SMILES recovery.

3.3 Molecular-structure cell discrimination

Molecular-structure cell discrimination determines whether a given cell contains a two-dimensional molecular-structure depiction. This step is a key distinction between MolTabReco and ordinary table OCR because the target output of a structure cell is not a normal string but SMILES. This study focuses on binary classification between molecular-structure and non-molecular-structure cells, and this decision directly determines whether a cell enters the SMILES recovery branch. The structure-cell discriminator uses EfficientNet-B0 [23] as the visual backbone. Because structure cells in ChemTable are a minority class with high information value, Class-Balanced Focal Loss [24,25] is used during training to alleviate class imbalance. The loss function is defined in Equation (3):

$$L_{CB} = -\alpha_y(1-p_y)^{\gamma}\log(p_y), \quad \alpha_y = \frac{1-\beta}{1-\beta^{n_y}} \qquad (3)$$

where n_y denotes the number of samples in class y, β controls the smoothing degree of the effective number of samples, γ is the Focal Loss modulation factor, and p_y denotes the predicted probability assigned by the model to the true class y. During inference, the structure-cell discriminator outputs the probability that the cell in row i and column j is a structure cell. If the probability is not lower than a threshold, the cell enters the OCSR-SMILES recovery branch; otherwise, it enters the non-structure text fusion branch.

3.4 Heterogeneous cell content recovery

The cell content recovery layer separately processes non-structure cells and structure cells. This layer does not use a single model to decode all cells uniformly. Instead, it selects the output pathway according to the cell type: non-structure cells mainly output text, numerical values, or mixed chemical text, whereas structure cells output computable SMILES.

3.4.1 Content recovery for non-structure cells

The non-structure text fusion branch processes ordinary text, numerical, and mixed chemical-text cells. In the main pipeline, these cells use a conservative VLM-first fusion strategy: when the VLM output is valid, the system preserves it; only when the VLM output is empty or parsing fails does TrOCR provide supplementary recognition. For the cell in row i and column j, let the VLM output be v_{ij}^V and the TrOCR output be v_{ij}^T. The fused result is defined as:

$$v_{ij} = \begin{cases} v_{ij}^{V}, & v_{ij}^{V} \neq \varnothing, \\ v_{ij}^{T}, & v_{ij}^{V} = \varnothing. \end{cases} \quad (4)$$

Equation (4) reflects the minimum-modification principle adopted for non-structure cells. Local OCR supplementation is activated only when the VLM result is missing or unavailable, thereby avoiding unnecessary replacement of stable VLM text. Both the raw text and normalized text are retained in the output to allow subsequent differentiation between true recognition errors and format differences.

3.4.2 Content recovery for structure cells

OCSR-SMILES recovery is the core module of MolTabReco. For cells classified as containing molecular-structure depictions, the system outputs SMILES rather than ordinary text or image placeholders. This study uses MolScribe [11] as the primary recognizer and combines crop expansion, image enhancement, foreground segmentation, RDKit canonicalization, and fragment filtering to improve recovery stability for small structure depictions in table contexts. In practice, the system first crops the structure depiction region according to the cell bounding box and applies moderate boundary expansion to reduce the risk of edge truncation. It then enhances low-resolution or thin-line structure depictions and performs foreground segmentation. MolScribe generates candidate SMILES, and RDKit [26] is used for validity checking, canonicalization, fragment splitting, and abnormal fragment filtering. MolNexTR and MolGrapher are used only in diagnostic OCSR candidate-complementarity experiments [12,13] to analyze consistency across different OCSR models and potential upper bounds; they are not required for the system-level main conclusion. For the k-th structure cell, the candidate SMILES set is defined as:

$$S_k = \{s_k^{MS}, s_k^{MN}, s_k^{MG}\} \quad (5)$$

where s_k^{MS}, s_k^{MN}, and s_k^{MG} denote the candidate SMILES provided by MolScribe, MolNexTR, and MolGrapher, respectively. The candidates in Equation (5) first undergo RDKit [26] validity checking, and valid candidates are converted into Morgan fingerprints [27]. After valid candidates are obtained, the system performs consistency selection according to the Tanimoto similarity [28] between candidates and preferentially selects the SMILES that is more chemically consistent with other valid candidates in fingerprint space. It should be emphasized that multi-OCSR candidate complementarity and Tanimoto-based consistency selection are used only to explain the diagnostic upper bound and error sources of the OCSR-SMILES recovery branch. The system-level main conclusions are still based on the output under the shared VLM-predicted coordinate benchmark.

3.5 Output fusion and post-processing

MolTabReco follows a minimum-modification principle in output fusion. The system first preserves the table structure, row-column indices, and cell bounding boxes. It then selects either the text fusion branch or the OCSR-SMILES recovery branch according to the discrimination result, and finally writes text, SMILES, confidence, and source information into a unified table object. Non-structure cells use VLM text first and TrOCR as a fallback for missing values; structure cells are filled with SMILES obtained from the OCSR-SMILES branch. Post-processing of chemical tables differs from ordinary text polishing. Numerical values, units, superscripts, subscripts, compound identifiers, and yields all have explicit experimental meanings, and modifications made by a generative model on the basis of linguistic plausibility may appear fluent but fail to remain faithful to the source. For cells with low confidence, multi-source result conflicts, format abnormalities, or inconsistent table context, the system may trigger LLM-assisted review. However, LLM suggestions are not written directly into the final table; they must pass rule checks for numerical formats, units, special symbols, superscripts and subscripts, and verifiable chemical

formats before being accepted. To make the recovered table object traceable and diagnostically interpretable, the main cell-level output fields retained by MolTabReco are summarized in Table 1.

Table 1. Cell-level output fields.

Field	Meaning
cell_id	Unique cell identifier, for example, page_0001_r02_c03
bbox	Cell bounding box in the format [x1, y1, x2, y2]
row_start / row_end	Start and end row indices of the cell
col_start / col_end	Start and end column indices of the cell
cell_type	Cell type, such as text, numeric, mixed, structure, or empty
text_raw	Raw text output
text_norm	Normalized text output
smiles	SMILES output corresponding to a structure cell
confidence	Confidence score or quality flag for the current output
source	Output source, such as VLM, TrOCR, MolScribe, OCSR candidate-consistency selection, or rule-based repair

This output format retains raw text, normalized text, SMILES, and source information, which facilitates diagnosis of whether an error arises from coordinate localization, structure-cell discrimination, text recognition, OCSR-SMILES recovery, or post-processing. It also facilitates importing recognition results into manual validation or chemical database construction workflows. Given a chemical paper page image containing a table, the goal of MolTabReco is to output a traceable structured table object. This object includes the table region, row-column structure, ordinary cell text, pixel positions, row-column indices, cell types, recognition sources, and quality flags. For molecular-structure cells, the system must also output canonical SMILES parseable by cheminformatics tools and bind the SMILES to the correct row-column position. For high-risk cells that trigger LLM-assisted review, the source and post-processing fields record the reason for correction triggering, correction status, and rule-validation results.

3.6 ChemTable dataset and preprocessing

Experiments were conducted on the ChemTable dataset [17]. The experimental set contains 1,382 chemical table page images from 473 chemistry papers, with publication years covering 2015-2024. Under the original annotation protocol, the dataset contains 892 chemical reaction annotations, 1,523 table-cell SMILES mappings, 2,802 reaction-related SMILES mappings, and 1,019 table footnotes or notes. The total number of cell-level samples is 68,931, including 1,507 molecular-structure cells (2.19%) and 67,424 non-structure cells (97.81%). The OCR subset used for training and evaluating the non-structure text recognition branch contains 67,419 samples and is split into 47,451 training, 9,448 validation, and 10,520 test samples. The difference of five samples between the 67,424 cells in the classification statistics and the 67,419 cells in the OCR subset mainly results from text cleaning or abnormal sample filtering.

The difficulties of the ChemTable dataset mainly arise from page layout, structural layout, and cell content. Page titles, reaction schemes, and footnotes may interfere with main-table localization. Inside tables, multi-row headers, row and column spans, empty cells, and non-uniform row heights and column widths are common. Ordinary cells may also contain superscripts and subscripts, units, chemical formulas,

footnotes, and LaTeX remnants. Although molecular-structure cells account for a small proportion of the dataset, they usually correspond to substrates, products, catalysts, or compound entities and are high-value minority-class cells for downstream cheminformatics tasks. Text preprocessing applies only lightweight normalization, including handling of spaces, HTML escapes, LaTeX remnants, superscript/subscript markers, reference symbols, percent signs, and common chemical format noise. This step reduces pseudo-errors caused by typographic differences without changing the original chemical meaning. Molecular-structure preprocessing uses RDKit to perform validity checking and canonicalization on both predicted and ground-truth SMILES. For multi-fragment prediction results, the OCSR-SMILES recovery branch performs main-fragment selection and abnormal fragment filtering. Because original cell annotations, training subsets, post-discriminator crop samples, and the system-level automatically predicted grid use different coordinate protocols, denominators are reported separately in each experiment. The original 1,507 structure cells are used for class distribution and discriminator training statistics, whereas the system-level main conclusion uses 1,682 structure cells aligned under the shared automatically predicted grid as the denominator; these two denominators are not directly equivalent.

3.7 Evaluation metrics

This study distinguishes among full-page images, ground-truth table crops, detection-based crops, system-level evaluation under a shared VLM-predicted table grid, and structure-cell crops after filtering by the structure-cell discriminator. Full-page images are used to observe the influence of page-level interfering factors such as titles, footnotes, and reaction schemes. Ground-truth table crops are used to estimate module upper bounds under ideal cropping conditions. Detection-based crops are used to evaluate the impact of table-region detection error on subsequent processing. The shared VLM-predicted grid is used to compare the VLM baseline and MolTabReco under the same coordinate benchmark. Structure-cell crops after discriminator filtering are used to analyze the diagnostic performance of the OCSR-SMILES recovery branch. The main evaluation metrics used for different subtasks in the MolTabReco pipeline are summarized in Table 2.

Table 2. Main evaluation metrics and their uses.

Subtask	Metric	Main use
Table detection	Mean Intersection over Union (IoU), IoU@0.9	Evaluate main-table region localization quality
Table structure recovery	TEDS, TEDS-Struct, Row Acc, Col Acc, Both Acc	Evaluate logical structure and row-column dimension matching
Cell bounding box	Cell IoU, Row Acc, Col Acc, Both Acc	Evaluate reliability of cell cropping
Text recognition	Exact Match	Evaluate text recovery for non-structure cells
Structure-cell discrimination	Precision, Recall, F1, Accuracy	Evaluate binary classification between structure and non-structure cells
OCSR-SMILES recovery	Canonical EM, NoStereo EM, SMILES recognition rate, Tanimoto	Evaluate SMILES-level recovery and structural similarity

The system-level conclusions are based on evaluation under the shared VLM-predicted grid. This evaluation takes page images as input and structured table objects as output. The row-column grid predicted by the VLM serves as the automatic intermediate result of the table structure parsing stage and is used as the common coordinate benchmark for constructing both the VLM baseline and MolTabReco. This setting does not use ground-truth row-column structures, ground-truth cell types, or ground-truth text as

intermediate inputs; ground-truth annotations are used only for final metric calculation. For structure cells, this study does not match ground-truth SMILES as ordinary text, but instead separately computes SMILES canonical EM, no-stereochemistry EM, and Tanimoto similarity [28]. Placeholder or weak-content matching by the VLM baseline on structure cells is not equivalent to OCSR-to-SMILES recognition capability and therefore must be interpreted separately from SMILES-level recovery by MolTabReco. If the same metric differs across experiments, the difference is explained according to denominator, input type, use of ground-truth crops, whether the experiment is a diagnostic OCSR candidate-complementarity experiment, and whether it is a system-level shared-prediction-coordinate evaluation, thereby avoiding the mixing of evaluation protocols.

3.8 Experimental configuration and comparison settings

The experiments are divided into module-level diagnostic experiments and system-level evaluation. Module-level experiments explain the rationale for module selection and failure modes, including table-region detection, row-column structure parsing, cell bounding-box generation, structure-cell discrimination, text fusion, OCSR-SMILES recovery, and post-processing. System-level evaluation combines structure-cell discrimination, VLM-first text fusion, and OCSR-SMILES recovery under the shared VLM-predicted grid and recomputes the performance of the full pipeline. Module-level experiments are used to justify design choices, while system-level evaluation supports the main conclusions of this study.

Model training and inference were mainly performed on a single V100 GPU with 16 GB of memory. Table-region detection and row-column bounding-box generation are based on Table Transformer [1]. The structure-cell discriminator is based on EfficientNet-B0 [23]. The non-structure text branch uses TrOCR [5]. The OCSR-SMILES recovery branch uses MolScribe [11] as the primary model, and MolNexTR and MolGrapher [12,13] are introduced in diagnostic candidate-complementarity experiments. RDKit [26] is used for SMILES validity checking, canonicalization, stereochemistry removal, fragment filtering, and Tanimoto similarity calculation.

4 Results

This section evaluates the performance of MolTabReco on the ChemTable dataset and further analyzes the SMILES recovery capability for molecular-structure cells, error propagation paths, and the influence of key modules on the final results. To avoid mixing module-level results with system-level results, the experimental results are divided into two categories. The first consists of multi-stage comparisons under the shared VLM-predicted grid and is used to support the main conclusions. The second consists of diagnostic experiments, including table detection, row-column parsing, cell bounding-box localization, structure-cell discrimination, text recovery, OCSR-SMILES recovery, and post-processing, and is used to explain the sources of performance and the main bottlenecks.

4.1 Full-pipeline evaluation of MolTabReco

On all 1,382 ChemTable pages, both the VLM baseline and MolTabReco use the page image as input and share the same VLM-predicted row-column grid. Therefore, the Row Accuracy, Column Accuracy, and Both Accuracy of the two methods are identical, at 91.97%, 97.68%, and 90.30%, respectively. This setting makes the multi-stage comparison mainly reflect cell content recovery rather than differences in row-column structure prediction. As shown in Figure 2, MolTabReco increases Overall Value ACC from 82.53% to 83.29% and TEDS from 0.8537 to 0.8595. The overall improvement is limited mainly because molecular-structure cells account for a small proportion of all cells, whereas ordinary text, numerical, and mixed-text cells form the major denominator of the overall metrics. Therefore, Overall Value ACC and TEDS are not sensitive to improvements in structure cells. More representative results are

obtained from structure cells. The 8.09% score of the VLM baseline on this cell type mainly corresponds to placeholders, empty contents, or weak semantic matches, and does not indicate that the VLM completed structure-to-SMILES conversion. In contrast, MolTabReco achieves a SMILES canonical EM of 44.05% on 1,682 system-level structure cells, showing that the explicit structure-cell discriminator and OCSR-SMILES branch can recover originally non-computable image cells as canonical SMILES. At the same time, MolTabReco reaches an ACC of 84.26% on non-structure cells, slightly lower than the VLM baseline at 84.36%. This indicates that the main contribution of MolTabReco is not to replace the VLM for ordinary text recognition, but to complement the VLM by providing missing structure-to-SMILES recovery under table-coordinate constraints.

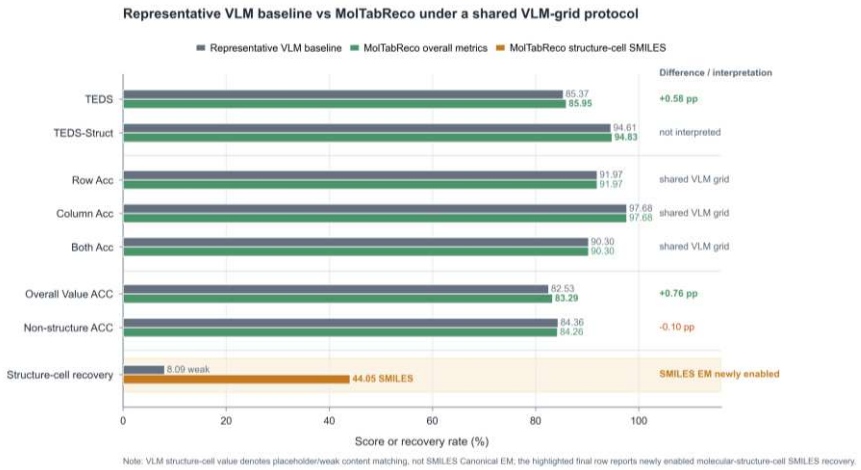


Figure 2. System-level comparison between the representative VLM baseline and MolTabReco.

4.2 SMILES-level recovery of molecular-structure cells

Figure 3 further presents the SMILES-level recovery results for molecular-structure cells. In the full system-level evaluation, there were 1,682 structure cells under the system-level alignment protocol, and MolTabReco obtained 741 canonical exact matches, corresponding to a canonical EM of 44.05%. Among the 1,248 pages where the VLM-predicted row and column dimensions matched the ground-truth annotation, there were 1,159 structure cells with ground-truth SMILES annotations. In this subset, MolTabReco produced 1,095 parseable SMILES, corresponding to a recognition rate of 94.48%. Among them, 699 samples matched the ground truth after stereochemistry was ignored, corresponding to a no-stereochemistry EM of 60.31%; 635 samples achieved strict canonical exact matching, corresponding to a canonical EM of 54.79%. Canonical EM, no-stereochemistry EM, and Tanimoto similarity reflect different levels of structure recovery quality. Canonical EM requires the predicted SMILES to be identical to the ground-truth annotation after canonicalization and is the strictest database-level metric. No-stereochemistry EM focuses more on whether the molecular scaffold is correct. Tanimoto similarity measures the structural similarity between non-exact predictions and the ground-truth structure. The 5.52-percentage-point gap between canonical EM and no-stereochemistry EM indicates that some errors mainly arise from stereochemical expression rather than from failure to recognize the main molecular scaffold.

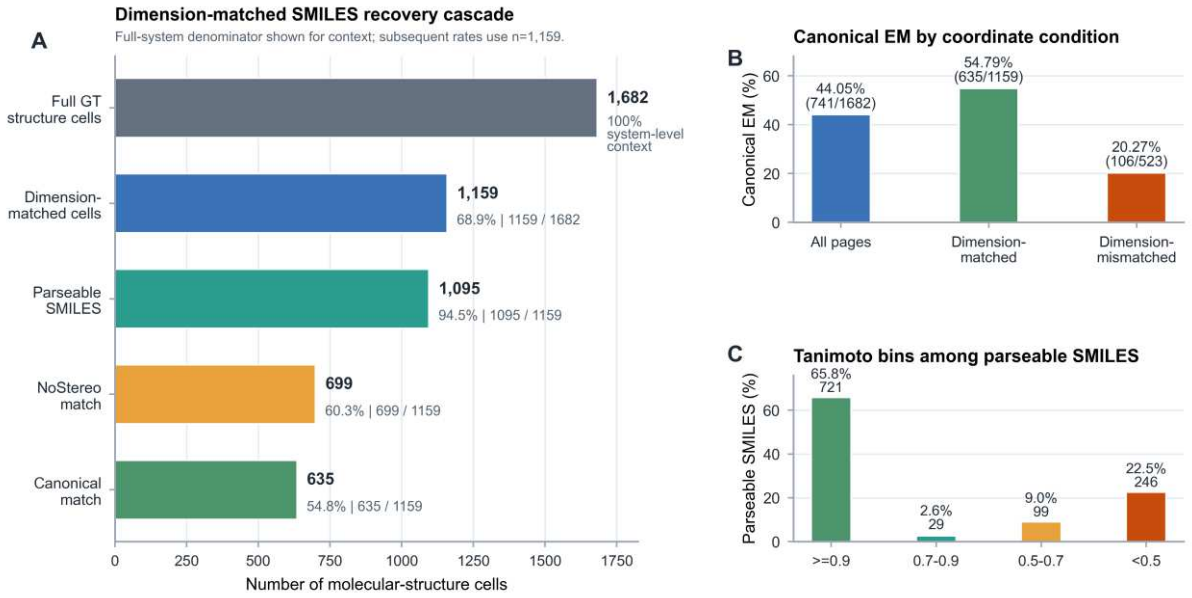


Figure 3. SMILES recovery cascade and coordinate-dependent evaluation for molecular-structure cells. (A) Recovery cascade from all ground-truth structure cells to canonical matches. (B) Canonical EM under different coordinate-matching conditions. (C) Tanimoto similarity distribution among parseable SMILES predictions.

The coordinate condition has a clear impact on SMILES recovery. Canonical EM is 44.05% on all pages, increases to 54.79% on pages with matched row-column dimensions, and decreases to 20.27% on pages with unmatched row-column dimensions. This shows that structure recovery in chemical tables cannot be regarded as an independent OCSR task. Even if an OCSR model can recognize the molecule in an image region, the final table-level evaluation still marks the result as incorrect if the structure is not cropped from the correct cell or is not bound to the correct row-column position. Among 1,095 parseable SMILES predictions, 721 samples had a Tanimoto similarity of at least 0.9, accounting for 65.8%, and 849 samples had Tanimoto similarity of at least 0.5, accounting for 77.5%. This indicates that some predictions that fail to meet strict canonical EM still retain high structural similarity and can provide useful candidates for manual validation, candidate screening, or similarity search.

4.3 Error propagation and bottlenecks in full-pipeline performance

To explain the coexistence of small overall metric changes and substantial improvement in structure cells, this study decomposes multi-stage error propagation into three coupled bottlenecks: row-column dimension matching, structure-cell discrimination, and OCSR-SMILES recovery. Figure 4 shows the error propagation path. It should be noted that different stages do not use the same denominator. Page-level dimension matching, structure-cell discrimination, and SMILES recovery correspond to denominators at the page level, system-level structure-cell level, and dimension-matched structure-cell subset level, respectively. Therefore, these proportions should not be simply added together; instead, they should be treated as diagnostic indicators for locating error sources and estimating system upper bounds.

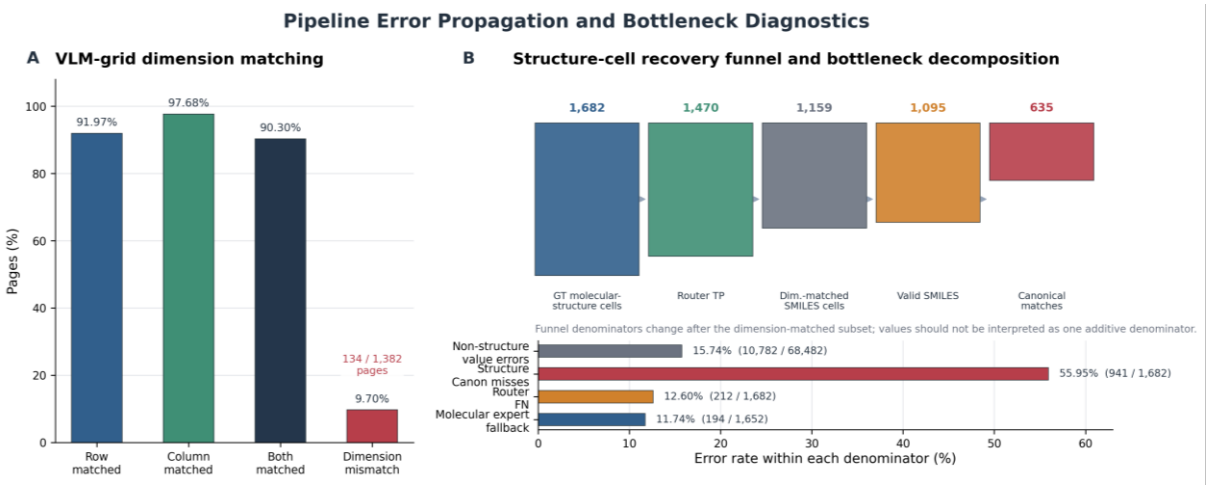


Figure 4. Pipeline error propagation and bottleneck diagnostics for table-grounded SMILES recovery.

There are still 134 pages in the VLM-predicted grid with row-column dimension mismatches, which simultaneously affect content matching, the crop input to the structure-cell discriminator, and SMILES position binding. Among 1,682 ground-truth structure cells, 1,470 were correctly classified as structure cells by the discriminator, while 212 missed samples directly lost the opportunity for SMILES recovery. On dimension-matched pages, 1,095 samples produced parseable SMILES, and 635 achieved canonical exact matching. These results show that the main errors in MolTabReco do not arise from a single module, but propagate jointly through coordinates, structure-cell discrimination, and OCSR-SMILES recovery. Upstream dimension mismatches may cause a structure depiction to be cropped from the wrong position or backfilled into the wrong cell. Discriminator false negatives interrupt the OCSR-SMILES recovery path. The OCSR-SMILES branch remains affected by complex scaffolds, stereochemistry, small structure depictions, table lines, and bounding-box displacement. Therefore, improving system-level canonical EM for structure cells requires simultaneous improvements in coordinate localization, minority-class discrimination, and OCSR under table contexts, rather than relying only on a stronger single-image structure recognition model. This also explains why MolTabReco improves structure cells substantially while Overall Value ACC increases only slightly: structure cells are rare among all cells but have higher information density for chemical database construction.

4.4 Influence of table localization on cell recovery

SMILES recovery in MolTabReco depends on a stable table-coordinate layer. This layer includes main-table region detection, row-column grid parsing, and pixel-level cell bounding-box generation. Only when the table region is correctly cropped, the logical row-column structure is correctly recovered, and the cell bounding boxes are sufficiently accurate can structure cells be correctly cropped, classified, and passed to the OCSR-SMILES branch. Therefore, the coordinate localization experiments in this section do not pursue detection metrics in isolation; instead, they analyze how coordinate quality affects structure recovery.

4.4.1 Table-region detection and main-table cropping

Figure 5 compares the performance of different table detection models under zero-shot and ChemTable-finetuned settings. Under zero-shot conditions, different detection models vary substantially, indicating that titles, footnotes, reaction schemes, molecular-structure depictions, and surrounding

paragraphs in chemical literature pages interfere with general object detection models. After ChemTable fine-tuning, Table Transformer [1] reaches a mean IoU of 0.9528 and an IoU@0.9 of 94.26%; YOLOv8x [29] reaches a mean IoU of 0.9507 and an IoU@0.9 of 91.39%; and YOLOv11x [30], YOLOv12x [31], and RT-DETR-L [32] also show clear improvements after domain adaptation.

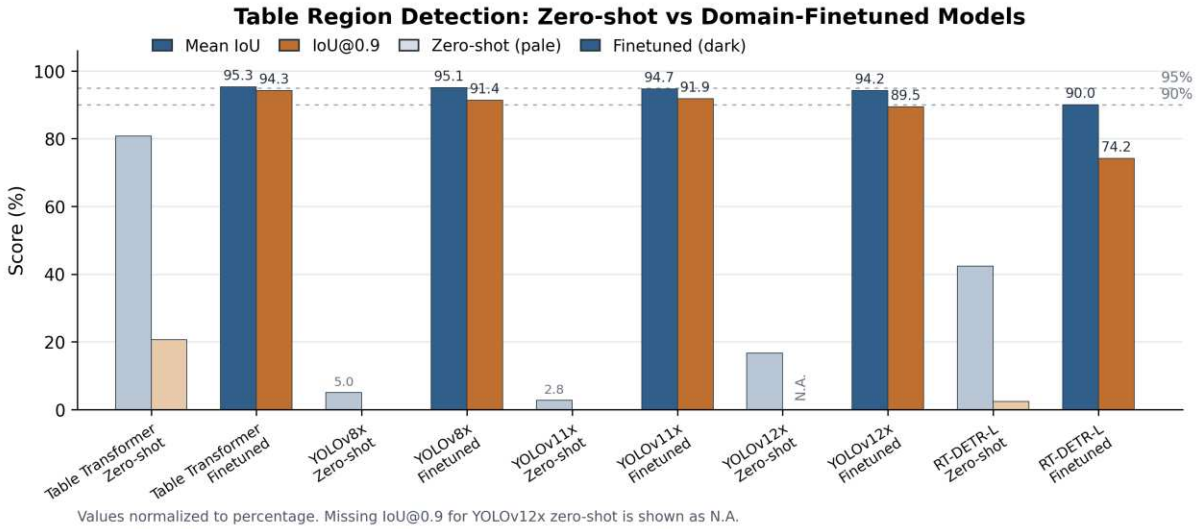


Figure 5. Table-region detection performance under zero-shot and ChemTable-finetuned settings.

These results indicate that, after fine-tuning, main-table region detection is no longer the main bottleneck of MolTabReco, but it remains a necessary preliminary step for the entire multi-stage pipeline. Chemical paper pages usually contain titles, reaction schemes, molecular-structure depictions, footnotes, and explanatory text. If row-column parsing is performed directly on the full page image, non-table elements may be misrecognized as rows, columns, or cell contents. Main-table cropping reduces page-level interference and provides a more stable spatial input for subsequent row-column grid parsing and cell cropping.

4.4.2 Row-column grid parsing and the influence of crop inputs

Figure 6 shows TEDS-Struct performance of different models for chemical table row-column structure recovery. Traditional table structure recognition (TSR) models perform weakly overall on complex chemical tables. LORE, DETR, TRUST, ETT, VAST, Swin-Tiny, and GridFormer all have TEDS-Struct values below approximately 0.40 on original full-page input. GridFormer is the strongest among the traditional models but still does not meet the coordinate requirements of multi-stage chemical table recovery. In contrast, recent multimodal parsers and VLM-type methods perform more strongly. For example, PP-TableMagic reaches 0.847 [33], DeepSeek-OCR-2 reaches 0.862 [34], Qwen3-VL-8B reaches 0.887 [35], PaddleOCR-VL-1.5 reaches 0.897 [36], GLM-OCR reaches 0.926 [37], Logics-Parsing-v2 reaches 0.928 [38], Qwen2.5-VL-7B reaches 0.943 [39], and the representative VLM pipeline based on Kimi 2.5 API calls reaches 0.946. These results support the use of the VLM-predicted grid as the shared coordinate benchmark for system-level evaluation. Chemical tables often contain multi-row headers, row and column spans, molecular-structure depictions, non-uniform row heights and column widths, and dense chemical symbols; traditional TSR models struggle to recover a complete logical grid stably. VLMs or multimodal document parsers show more stable logical structure understanding.

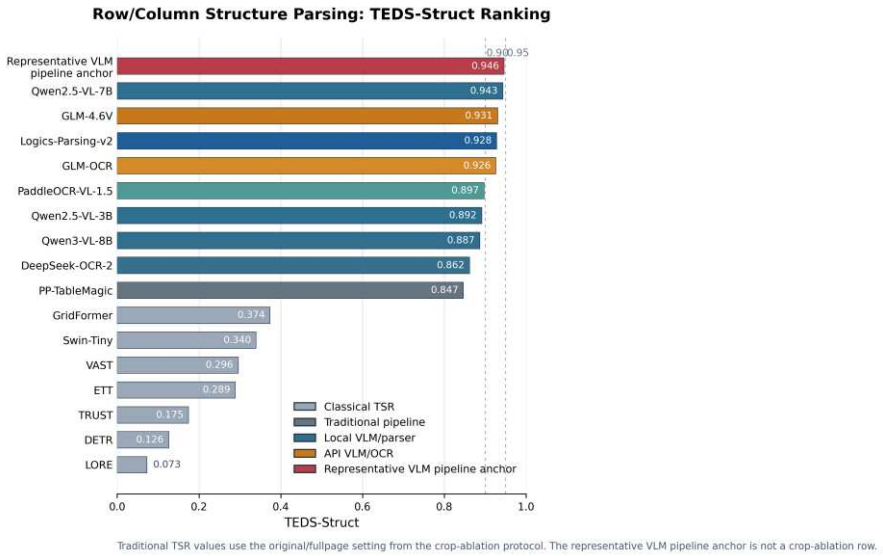


Figure 6. Row-column grid recovery across model families for table-coordinate grounding.

Figure 7 further compares the influence of original full-page images, ground-truth main-table crops, and detected main-table crops on TEDS-Struct. For many traditional TSR models, main-table cropping significantly improves structure recovery. For example, Swin-Tiny increases from 0.340 on original full pages to 0.495 on ground-truth crops and 0.466 on detected crops; VAST increases from 0.296 to 0.381 and 0.366; and GridFormer increases from 0.374 to 0.456 and 0.425. Some VLMs or parsers also benefit from cropping. For example, DeepSeek-OCR-2 increases from 0.862 to 0.926 [34], Logics-Parsing-v2 increases from 0.928 to 0.946 [38], and GLM-OCR reaches 0.959 after cropping [37]. Detected-crop results are usually close to ground-truth-crop results, indicating that the fine-tuned table detector can provide sufficiently reliable main-table region inputs.



Figure 7. Main-table crop ablation for row-column grid recovery in chemical tables.

4.4.3 Cell bounding-box localization

A logical row-column grid alone is insufficient for structure recovery because the OCSR-SMILES branch requires pixel-level cell crops. Figure 8 evaluates the influence of count constraints on cell bounding-box generation. Without count constraints, the fine-tuned structure detection model has a Row Acc of 79.2% and a Col Acc of 85.3%. After introducing ground-truth row and column count constraints, Row Acc increases to 94.4% and Col Acc to 98.5%. When representative VLM-predicted row and column counts are used as constraints, Row Acc still reaches 90.7% and Col Acc reaches 98.5%. Notably, count constraints improve the correctness of row and column counts without substantially damaging localization quality. Under the unconstrained, ground-truth-count-constrained, and VLM-count-constrained settings, Cell IoU values are 76.86%, 76.67%, and 76.67%, respectively, with only minor differences. In contrast, the uniform-grid baseline has a Cell IoU of only 27.24%, indicating that chemical tables cannot be approximated by equally spaced grids. In real chemical tables, row heights and column widths are often highly non-uniform because of molecular-structure depictions, multi-row headers, merged cells, long text, and footnotes.

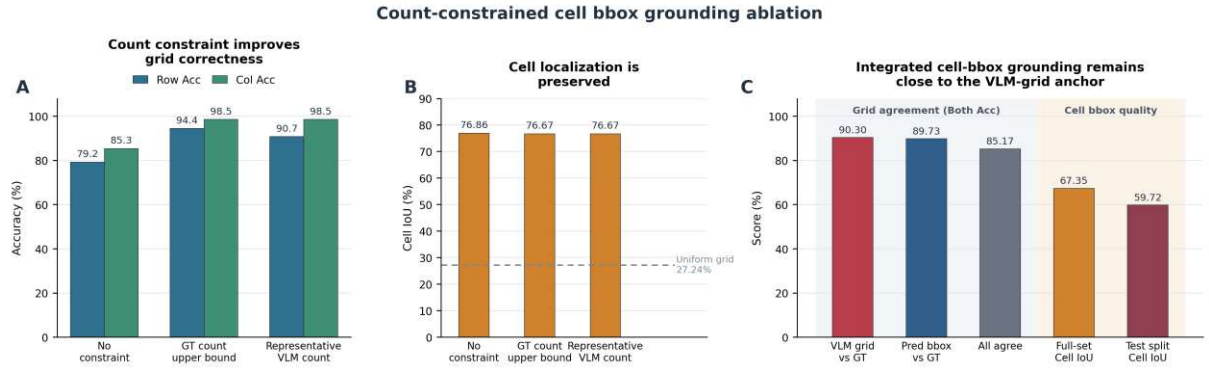


Figure 8. Count-constrained cell bounding-box grounding for table-coordinate construction.

The combined cell localization results show that the Both Acc between the VLM grid and ground-truth annotation is 90.30%, the matching rate between predicted bounding boxes and ground-truth annotation is 89.73%, and the proportion for which all three are simultaneously consistent is 85.17%. The Cell IoU is 67.35% on the full set and 59.72% on the test set. These results indicate that the coordinate layer is generally close to the VLM grid anchor but still has boundary shifts and crop errors. For structure cells, even when the row-column dimensions are correct, bounding-box displacement may truncate structure edges, introduce table lines, or mix in adjacent content, thereby reducing SMILES recovery performance.

4.5 Molecular-structure cell discrimination

The molecular-structure cell discriminator determines whether a cell enters the text recovery branch or the OCSR-SMILES recovery branch and is an important component that distinguishes MolTabReco from ordinary table OCR. Because structure cells are few in number but high in information value, using only overall Accuracy can obscure minority-class errors. Therefore, this study focuses on Precision, Recall, and F1.

Figure 9 presents diagnostic results for different structure-cell discrimination methods. At the default threshold, the DECIMER image classifier [14,15] has a Precision of 0.315, Recall of 0.857, and F1 of

0.460. After threshold tuning, its Precision increases to 0.851, Recall is 0.798, and F1 is 0.823. In contrast, the EfficientNet-B0 structure-cell discriminator [23] trained on clean ChemTable cell crops reaches a Precision of 0.979, Recall of 0.987, and F1 of 0.983, indicating a high module-level upper bound.

When this discriminator is placed into the full automatic pipeline, the input changes from clean annotated crops to automatically predicted cell crops, and its performance decreases to a Precision of 0.8898, Recall of 0.8740, and F1 of 0.8818. This decrease indicates that upstream row-column dimension mismatches, bounding-box displacement, and crop noise change the input distribution of the discriminator. In system-level evaluation, 212 true structure cells are missed and therefore cannot enter the OCSR-SMILES recovery branch, directly losing SMILES recovery opportunities.

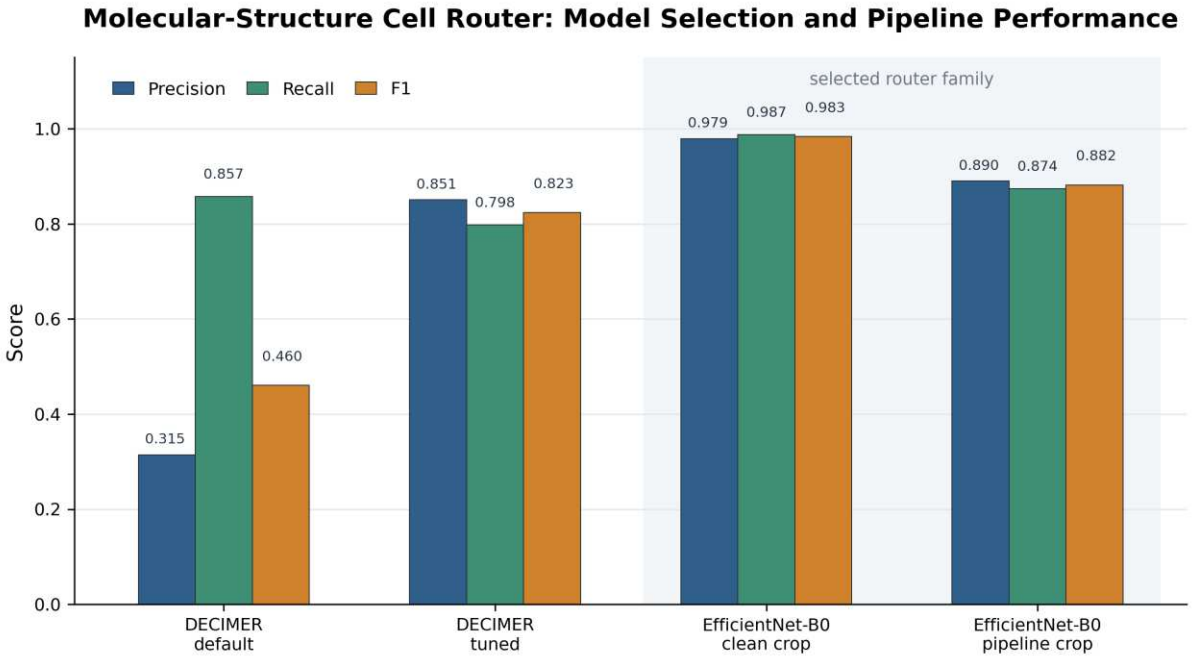


Figure 9. Molecular-structure cell router diagnostics under clean-crop and pipeline settings.

From the perspective of multi-stage effects, false negatives for structure cells are usually more serious than false positives. A false negative causes a true structure cell to bypass the OCSR-SMILES branch entirely. Although a false positive sends a non-structure cell into the molecular branch, its impact can be limited by SMILES validity checking, confidence filtering, and fallback strategies. Therefore, the optimization objective of the structure-cell discriminator should not be only overall Accuracy. Instead, it should reduce false negatives for high-value structure cells while maintaining high Precision.

4.6 Cell content recognition

MolTabReco includes text content recognition for non-molecular-structure cells and OCSR-SMILES content recognition for molecular-structure cells. The main improvement of the framework is concentrated in OCSR-SMILES content recognition for molecular-structure cells.

4.6.1 Text content recognition for non-molecular-structure cells

Figure 10 analyzes the non-structure text recovery branch. On clean cell crops, different TrOCR versions [5] improve gradually after label normalization and post-processing: V1 reaches 84.24%, V2 reaches 84.17%, V3 reaches 87.96%, and V4 reaches 88.59%. This indicates that TrOCR has strong cell

text recognition capability under clean cropping and normalized label conditions. However, clean-crop performance cannot be directly extrapolated to the full pipeline. Figure 10B shows that on automatically predicted cell crops, the exact match of the TrOCR pipeline crop is only 70.63%; VLM text output reaches 85.50%; and VLM+TrOCR fallback reaches 84.76%, slightly lower than VLM-only output. This suggests that bounding-box displacement, neighboring-content interference, and lack of context in the real pipeline substantially reduce local OCR performance, whereas the VLM can use table context to recover ordinary text and numerical values. Therefore, indiscriminately replacing VLM text output with TrOCR may introduce additional noise.

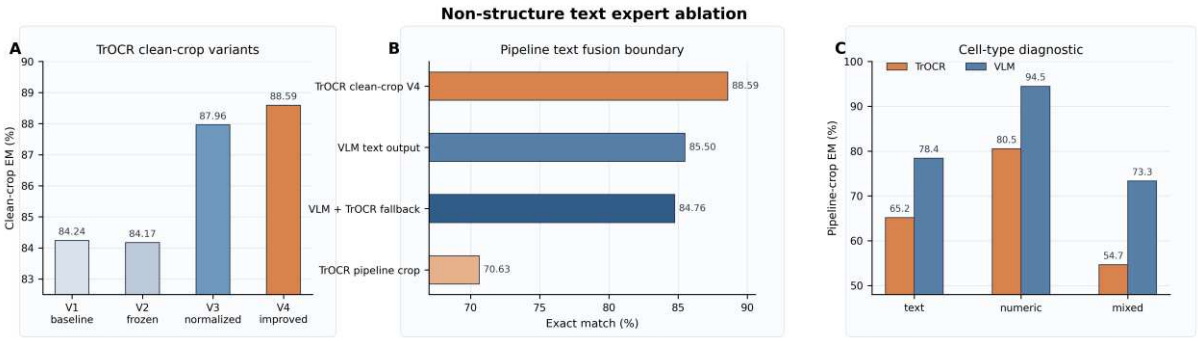


Figure 10. Non-structure text recovery ablation and VLM-first fusion boundary.

Analysis by cell type also supports this conclusion. Under pipeline cropping, TrOCR reaches EM values of 65.2%, 80.5%, and 54.7% for text, numeric, and mixed cells, respectively, whereas the corresponding VLM results are 78.4%, 94.5%, and 73.3%. The advantage of the VLM is especially clear for numeric cells, likely because table context helps identify values, units, and column semantics. For mixed chemical text, local OCR is also susceptible to superscripts and subscripts, chemical abbreviations, footnote markers, and special symbols. Therefore, MolTabReco adopts a conservative VLM-first fusion strategy: when VLM text is available, the VLM output is retained, and TrOCR is used only as a fallback when the VLM output is missing or parsing fails.

4.6.2 SMILES recovery for molecular-structure cells

Figure 11 summarizes the main ablation results of the OCSR-SMILES recovery branch and explains the source of MolTabReco's SMILES recovery performance on structure cells. In the single-model OCSR comparison, AtomLenz [40] has a canonical EM of 6.45%, MolGrapher [12] 25.74%, MolNexTR [13] 38.05%, and MolScribe [11] 43.78%. MolScribe is the most stable and is therefore selected as the primary recognizer in the OCSR-SMILES recovery branch. After fragment repair is added, canonical EM increases to 59.18%. GT-style diagnostic fusion further increases it to 63.46%. It should be emphasized that GT-style diagnostic fusion is a diagnostic upper-bound analysis used to illustrate complementarity among candidate models, and it is not equivalent to a strict multi-stage system setting.

Figure 11B shows that the bottleneck of the OCSR-SMILES recovery branch does not originate only from the OCSR model itself. When row-column coordinates match, canonical EM is 52.9%; when row-column coordinates do not match, it is only 27.8%. When GT-style auxiliary candidates are absent, canonical EM is 13.1%; when GT-style candidates are present, it increases to 51.7%. In addition, canonical EM is 47.9% when initial pipeline bounding boxes are used and increases to 63.5% when ground-truth structure depiction bounding boxes are used. These results indicate that molecular recognition strongly depends on upstream coordinate localization and crop quality. Even when the OCSR model itself has

strong recognition capability, erroneous cropping, edge truncation, table-line contamination, and adjacent-cell interference can still lead to SMILES errors.

Figure 11C compares the effects of crop expansion and Hybrid Attention Transformer (HAT) super-resolution [41]. Moderate expansion helps reduce the risk of truncating structure depiction edges, but excessive expansion may introduce table lines, adjacent text, or content from other cells. The benefit of HAT super-resolution in this task is unstable. Under 30% expansion, canonical EM is 48.54% without HAT and 48.81% with HAT 2x. Under 35% expansion, HAT 2x reaches 48.88%, but the improvement relative to 30% expansion is only 0.07 percentage points. This indicates that the main bottleneck of structure recovery is not only image resolution, but also crop alignment, line noise, atom-symbol recognition, bond-topology recovery, and stereochemical expression.

Figure 11D shows the stepwise ablation of the molecular expert branch. Using 30% margin + HAT as a stable baseline, canonical EM is 48.81%. Adding U-Net foreground processing [42] increases it to 49.83%; adding candidate voting increases it to 50.24%; further adding Tanimoto >= 0.7 consistency selection increases it to 50.37%; and relaxing the Tanimoto threshold to >= 0.5 increases it to 50.44%. From the 30% margin + HAT baseline to the final setting, canonical EM increases by 1.63 percentage points. The improvement is limited but cumulative. It should be noted that the 50.44% in Figure 11D corresponds to the diagnostic denominator of 741/1,469 structure-cell crops after discriminator filtering, whereas the system-level 44.05% is obtained by placing the same 741 canonical exact matches back into the denominator of 1,682 system-level structure cells. This denominator difference shows that the goal of MolTabReco is not to maximize OCSR scores on ideal single-structure crops, but to recover computable SMILES in the real table-coordinate pipeline and backfill them into the correct cell positions.

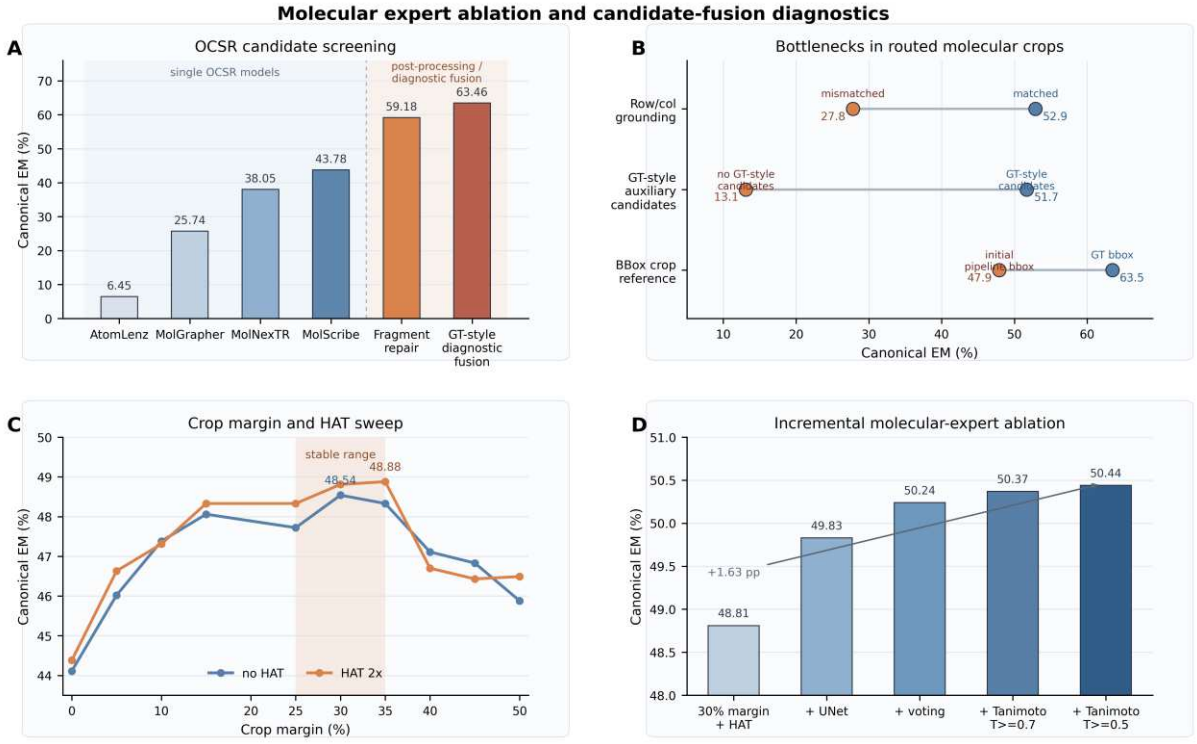


Figure 11. Molecular-expert ablation and candidate-fusion diagnostics for table-grounded SMILES recovery.

Overall, the OCSR-SMILES recovery branch remains affected by small structure depiction size, bounding-box displacement, complex ring systems, spurious fragments, multi-fragment outputs, and stereochemical differences. Multi-OCSR candidates and Tanimoto-based consistency selection are mainly used to explain model complementarity and potential upper bounds, rather than to introduce uncontrolled generative correction. All candidates must pass RDKit validity checking, canonicalization, and fragment filtering, and the final system-level conclusions remain based on outputs under the shared VLM-predicted grid.

4.7 Evaluation of LLM-assisted post-processing

Figure 12 evaluates the effectiveness and risk of post-processing strategies. In chemical tables, numerical values, units, superscripts and subscripts, footnotes, compound identifiers, and structure labels usually have explicit experimental meanings and cannot be freely rewritten only according to linguistic fluency. The experimental results show that unconstrained LLM correction does not provide stable benefits. Among four sets of free-form generative correction experiments, three reduced Value ACC, and only one produced a very small improvement of 0.03 percentage points. Among 239 cells modified by the LLM, wrong-to-correct edits accounted for only 13.4%, wrong-to-wrong edits accounted for 55.6%, and correct-to-wrong edits accounted for 31.0%. In other words, the LLM can indeed correct a small number of errors, but ineffective modification or erroneous modification of correct content is more common. In contrast, among categorized errors, 34.2% belonged to rule-repairable patterns, mainly including format cleaning, symbol normalization, fragment filtering, and verifiable string repair.

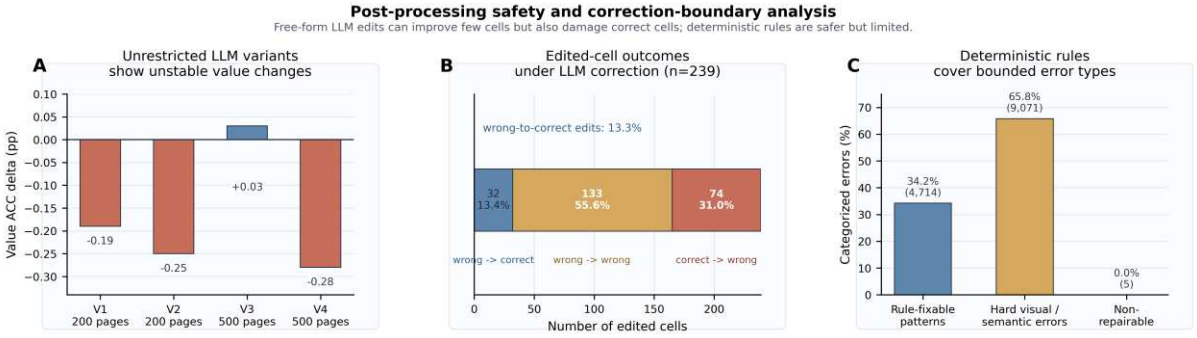


Figure 12. Post-processing safety and correction boundaries in chemical table recovery.

Therefore, MolTabReco should not adopt free-form generative correction over the whole table. A more appropriate approach is to restrict LLM use to high-risk cells with low confidence, multi-source conflicts, format abnormalities, or contextual inconsistencies, and to accept modifications only after they pass checks for numerical fidelity, unit consistency, special-symbol handling, superscript/subscript constraints, and chemical-format validity. When verifiable evidence is absent, the system should retain the original prediction or mark it as low-confidence rather than automatically overwrite the result.

5 Discussion

The core question addressed in this study is not whether ordinary table text can be recognized, but whether two-dimensional molecular-structure depictions in chemical tables can be recovered as computable SMILES representations under table-coordinate constraints. The experimental results show that MolTabReco can recover a substantial proportion of molecular-structure depictions in chemical tables as computable SMILES under table-coordinate constraints.

5.1 Molecular-structure cells are a key bottleneck in chemical table recognition

VLMs can recover table skeletons and ordinary text content relatively well, but they lack molecular-structure recognition capability. The 8.09% score of the VLM baseline on structure cells mainly arises from placeholders or weak-content matching and should not be regarded as structure-to-SMILES recovery. In contrast, MolTabReco reaches a SMILES canonical EM of 44.05% in the multi-stage pipeline, showing that explicit structure-cell discrimination and the OCSR-SMILES branch can compensate for this missing capability in general table recognition workflows. It should be noted that structure cells account for a small proportion of all cells, and therefore Overall Value ACC and TEDS are not sensitive to improvements in these cells. The small improvement in overall metrics does not imply that structure recovery has limited value. For chemical data organization, structure retrieval, and reaction database construction, structure cells often correspond to substrates, products, catalysts, or key compound entities and directly determine whether table content can be converted into computable chemical data.

5.2 Heterogeneous recovery under table-coordinate constraints is necessary

Cells in chemical tables have different content forms. Text, numerical values, mixed chemical text, and molecular-structure depictions differ in visual appearance, semantic function, and target output. If all cells are processed with a uniform text decoding workflow, ordinary text and numerical values may be recovered reasonably well, but structure cells are often recognized only as empty content or image tags. MolTabReco first establishes table coordinates and then selects different recovery branches according to cell type. In this framework, the VLM mainly provides the table skeleton, row-column relationships, and contextual information. The OCSR model can recover molecular representations from structure depictions, but it cannot independently determine which row and column the structure cell belongs to, nor can it establish the relationship between the molecule and adjacent experimental conditions or performance metrics. Therefore, structure recovery in chemical tables should neither rely solely on VLMs nor be treated simply as standalone OCSR. It should instead perform localization, discrimination, and SMILES backfilling under table-coordinate constraints.

5.3 Post-processing should be bounded by verifiability

The post-processing experiments show that chemical tables are not suitable for free-form generative whole-table correction. Although an unconstrained LLM can correct a small number of errors, it can also change correct cells into incorrect ones or transform an erroneous content into a different erroneous content. For experimental values, units, compound identifiers, superscripts and subscripts, and structure labels, such modifications directly alter scientific meaning. Therefore, an LLM is more suitable as a targeted review tool than as a full post-processing module. A safer strategy is to trigger the LLM only for cells with low confidence, multi-source conflicts, format abnormalities, or contextual inconsistencies, and to accept modifications only after they pass checks for numerical fidelity, unit consistency, special-symbol handling, and chemical-format validity. In contrast, deterministic rules have limited coverage but are verifiable, reversible, and reproducible, and are therefore more appropriate as the basic layer of chemical table post-processing.

5.4 Limitations and future work

The MolTabReco framework still has several limitations. First, the system-level workflow depends on a representative VLM to provide the row-column grid and ordinary text output; therefore, different

VLM versions, call stability, and closed-source interfaces may affect reproducibility. Future work should evaluate more open-source VLMs or lightweight document parsing models to reduce dependence on closed-source interfaces. Second, the current results show that table coordinates, structure-cell discrimination, and OCSR-SMILES recovery all affect final performance, but finer-grained system-level ablations are still needed to quantify the contribution of each component. In particular, row-column misalignment, cell-boundary displacement, structure-cell false negatives, and OCSR recognition errors are linked through cascading relationships. Future work can improve system interpretability through stricter error attribution and confidence modeling. Finally, current SMILES recovery still does not meet the requirements of fully automated database construction. Complex molecules, low-resolution structure depictions, table-line interference, crop displacement, multi-fragment outputs, and stereochemical differences remain major challenges. Future work should further validate generalization in scenarios such as patent literature, scanned PDFs, historical literature, supplementary materials, and materials-science tables, and should combine manual review or active learning mechanisms to improve the recovery reliability of high-value structure cells.

6 Conclusions

This study proposes MolTabReco, a table-coordinate-constrained heterogeneous cell recovery framework for chemical literature tables. Given a page image, the method combines VLM-predicted table grids, cell cropping, structure-cell discrimination, VLM-first text fusion, and OCSR-SMILES backfilling to output a structured table object containing row-column coordinates, ordinary text, numerical content, and SMILES for structure cells. The experimental results show that VLMs can provide good table skeletons and ordinary text recovery, but cannot stably convert two-dimensional molecular-structure depictions inside tables into canonical SMILES. Compared with the 8.09% weak-content matching achieved by the VLM baseline on structure cells, MolTabReco reaches a SMILES canonical EM of 44.05% in system-level evaluation under the shared automatically predicted grid. On pages with matched row-column dimensions, canonical EM further reaches 54.79% and no-stereochemistry EM reaches 60.31%. In terms of overall metrics, Overall Value ACC increases from 82.53% to 83.29%, and TEDS increases from 0.8537 to 0.8595. These results show that the key challenge in chemical table recognition is not only table structure parsing or ordinary OCR, but also the handling of chemical cells with different output forms under table-coordinate constraints. Only when molecular-structure depictions are bound to the correct row-column positions and recovered as SMILES or molecular graphs can molecular information in tables enter downstream structure retrieval, similarity computation, and database construction workflows. The post-processing experiments further show that LLMs are more appropriate as restricted review tools for low-confidence or format-abnormal cells and should not be used for free-form whole-table rewriting. Overall, MolTabReco provides a multi-stage implementation pathway for traceable and verifiable recovery of molecular-structure cells in chemical literature tables.

Availability of data and materials

The ChemTable dataset used in this study is the publicly available research resource described in Ref. [17]. The original ChemTable data should be obtained from the original dataset release and used according to its terms and conditions.

The source code and publicly shareable evaluation materials for MolTabReco are available at: https://github.com/lucashu1234/MolTabReco.

The individual models, software packages, and public resources used in MolTabReco can be found under the following URLs:

1. Table Transformer: https://github.com/microsoft/table-transformer
2. TrOCR: https://huggingface.co/microsoft/trocr-base-printed
3. MolScribe: https://github.com/thomas0809/MolScribe
4. RDKit: https://github.com/rdkit/rdkit
5. Logics-Parsing-v2: https://huggingface.co/Logics-MLLM/Logics-Parsing-v2

Declarations

Ethics approval and consent to participate

This study is based on publicly available literature table data and model-output evaluation. It does not involve human participants, animal experiments, or personal privacy data.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific funding.

Authors' contributions

WH and WL conceived the study. WH designed and implemented the MolTabReco pipeline. WH, YX and JH prepared the ChemTable preprocessing and evaluation workflow. MO, JT and YS performed experiments, error analysis, and result validation. WH drafted the manuscript. WL supervised the study and revised the manuscript. All authors reviewed and approved the final manuscript.

Acknowledgements

The authors express their gratitude to the students and teachers of Hunan University of Chinese Medicine for their support in preparing experimental materials, organizing chemical table examples, and assisting with the evaluation workflow. The authors also thank the developers and maintainers of the ChemTable dataset and the open-source models and software used in this study.

References

1. Smock B, Pesala R, Abraham R (2022) PubTables-1M: towards comprehensive table extraction from unstructured documents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 4634-4642. https://doi.org/10.1109/CVPR52688.2022.00459.
2. Li M, Cui L, Huang S, Wei F, Zhou M, Li Z (2020) TableBank: table benchmark for image-based table detection and recognition. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp 1918-1925.

3. Zhong X, ShafieiBavani E, Jimeno Yepes A (2020) Image-based table recognition: data, model, and evaluation. In: European Conference on Computer Vision, pp 564-580. https://doi.org/10.1007/978-3-030-58589-1_34.
4. Zheng X, Burdick D, Popa L, Zhong X, Wang NXR (2021) Global Table Extractor (GTE): a framework for joint table identification and cell structure recognition using visual context. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp 697-706. https://doi.org/10.1109/WACV48630.2021.00074.
5. Li M, Lv T, Chen J, Cui L, Lu Y, Florencio D, et al (2023) TrOCR: transformer-based optical character recognition with pre-trained models. Proc AAAI Conf Artif Intell 37:13094-13102. https://doi.org/10.1609/aaai.v37i11.26538.
6. Kim G, Hong T, Yim M, Nam J, Park J, Yim J, et al (2022) OCR-free document understanding transformer. In: European Conference on Computer Vision, pp 498-517. https://doi.org/10.1007/978-3-031-19815-1_29.
7. Blecher L, Cucurull G, Scialom T, Stojnic R (2024) Nougat: neural optical understanding for academic documents. In: International Conference on Learning Representations (ICLR).
8. Li J, Li D, Savarese S, Hoi S (2023) BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the 40th International Conference on Machine Learning, PMLR 202:19730-19742.
9. Liu H, Li C, Wu Q, Lee YJ (2023) Visual instruction tuning. Adv Neural Inf Process Syst 36:34892-34916.
10. Li J, Zhang D, Wang X, Hao Z, Lei J, Tan Q, et al (2025) ChemVLM: exploring the power of multimodal large language models in chemistry area. Proc AAAI Conf Artif Intell 39:415-423. https://doi.org/10.1609/aaai.v39i1.32020.
11. Qian Y, Guo J, Tu Z, Li Z, Coley CW, Barzilay R (2023) MolScribe: robust molecular structure recognition with image-to-graph generation. J Chem Inf Model 63:1925-1934. https://doi.org/10.1021/acs.jcim.2c01480.
12. Morin L, Danelljan M, Agea MI, Nassar A, Weber V, Meijer I, et al (2023) MolGrapher: graph-based visual recognition of chemical structures. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 19552-19561.
13. Chen Y, Leung CT, Huang Y, Sun J, Chen H, Gao H (2024) MolNexTR: a generalized deep learning model for molecular image recognition. J Cheminform 16:141. https://doi.org/10.1186/s13321-024-00926-w.
14. Rajan K, Zielesny A, Steinbeck C (2021) DECIMER 1.0: deep learning for chemical image recognition using transformers. J Cheminform 13:61. https://doi.org/10.1186/s13321-021-00538-8.
15. Rajan K, Brinkhaus HO, Sorokina M, Zielesny A, Steinbeck C (2021) DECIMER-Segmentation: automated extraction of chemical structure depictions from scientific literature. J Cheminform 13:20. https://doi.org/10.1186/s13321-021-00496-1.
16. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28:31-36. https://doi.org/10.1021/ci00057a005.
17. Zhou Y, Cheng M, Mao Q, Luo Y, Liu Q, Li Y, Zhang X, Liu D, Li X, Chen E (2025) Benchmarking multimodal LLMs on recognition and understanding over chemical tables. arXiv preprint arXiv:2506.11375. https://doi.org/10.48550/arXiv.2506.11375.
18. Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39:1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031.
19. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European Conference on Computer Vision, pp 213-229. https://doi.org/10.1007/978-3-030-58452-8_13.

20. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al (2019) PubChem 2019 update: improved access to chemical data. Nucleic Acids Res 47:D1102-D1109. https://doi.org/10.1093/nar/gky1033.
21. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Felix E, et al (2019) ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 47:D930-D940. https://doi.org/10.1093/nar/gky1075.
22. Zhai Z, Druckenbrodt C, Thorne C, Akhondi SA, Nguyen DQ, Cohn T, et al (2021) ChemTables: a dataset for semantic classification on tables in chemical patents. J Cheminform 13:97. https://doi.org/10.1186/s13321-021-00568-2.
23. Tan M, Le QV (2019) EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning (ICML), pp 6105-6114.
24. Cui Y, Jia M, Lin TY, Song Y, Belongie S (2019) Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 9268-9277. https://doi.org/10.1109/CVPR.2019.00949.
25. Lin TY, Goyal P, Girshick R, He K, Dollar P (2020) Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 42:318-327. https://doi.org/10.1109/TPAMI.2018.2858826.
26. Landrum G, Tosco P, Kelley B, Rodriguez R, Cosgrove D, Vianello R, et al (2025) RDKit: open-source cheminformatics software. Zenodo. Version 2024.09.6. https://doi.org/10.5281/zenodo.14943932.
27. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50:742-754. https://doi.org/10.1021/ci100050t.
28. Bajusz D, Racz A, Heberger K (2015) Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminform 7:20. https://doi.org/10.1186/s13321-015-0069-3.
29. Jocher G, Chaurasia A, Qiu J (2023) Ultralytics YOLOv8. GitHub repository. https://github.com/ultralytics/ultralytics. Accessed 27 May 2026.
30. Khanam R, Hussain M (2024) YOLOv11: an overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725. https://doi.org/10.48550/arXiv.2410.17725.
31. Tian Y, Ye Q, Doermann D (2025) YOLOv12: attention-centric real-time object detectors. arXiv preprint arXiv:2502.12524. https://doi.org/10.48550/arXiv.2502.12524.
32. Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, Liu Y, Chen J (2024) DETRs beat YOLOs on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 16965-16974. https://doi.org/10.1109/CVPR52733.2024.01605.
33. PaddleOCR Developers (2025) General table recognition v2 pipeline (PP-TableMagic). PaddleOCR Documentation. https://paddlepaddle.github.io/PaddleOCR/v3.0.2/en/version3.x/pipeline_usage/table_recognition_v2.html. Accessed 27 May 2026.
34. Wei H, Sun Y, Li Y (2026) DeepSeek-OCR 2: Visual Causal Flow. arXiv preprint arXiv:2601.20552. https://doi.org/10.48550/arXiv.2601.20552.
35. Bai S, Cai Y, Chen R, Chen K, Chen XH, Cheng Z, Deng L, Ding W, Gao C, Ge C, et al (2025) Qwen3-VL technical report. arXiv preprint arXiv:2511.21631. https://doi.org/10.48550/arXiv.2511.21631.
36. Cui C, Sun T, Liang S, Gao T, Zhang Z, Liu J, et al (2026) PaddleOCR-VL-1.5: towards a multi-task 0.9B VLM for robust in-the-wild document parsing. arXiv preprint arXiv:2601.21957. https://doi.org/10.48550/arXiv.2601.21957.
37. Duan S, Xue Y, Wang W, Su Z, Liu H, Yang S, et al (2026) GLM-OCR technical report. arXiv preprint arXiv:2603.10910. https://doi.org/10.48550/arXiv.2603.10910.
38. Logics-MLLM (2026) Logics-Parsing-v2. Hugging Face model repository. https://huggingface.co/Logics-MLLM/Logics-Parsing-v2. Accessed 27 May 2026.
39. Bai S, Chen K, Liu X, Wang J, Ge W, Song S, Dang K, et al (2025) Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923. https://doi.org/10.48550/arXiv.2502.13923.

40. Oldenhof M, De Brouwer E, Arany A, Moreau Y (2024) Atom-level optical chemical structure recognition with limited supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 17669–17678. https://doi.org/10.1109/CVPR52733.2024.01673.

41. Chen X, Wang X, Zhou J, Qiao Y, Dong C (2023) Activating more pixels in image super-resolution transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 22367-22377. https://doi.org/10.1109/CVPR52729.2023.02142.

42. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015, Lecture Notes in Computer Science, vol 9351, pp 234-241. https://doi.org/10.1007/978-3-319-24574-4_28.