Appendix of RxnCaption: Reformulating Reaction Diagram Parsing as Visual Prompt Guided Captioning

1. Related Work

1.1. Chemical Reaction Mining

Existing research on extracting chemical reactions from literature primarily employs text mining to identify entities such as reactants and products in patents [18]. Tools like LeadMine perform entity recognition, while fingerprint-based methods enhance role assignment in noisy data [19]. Rule-based and machine learning approaches, including Naive Bayes, are used for section identification and dataset construction [18]. Recent initiatives like the ORD offer structured schemas for reaction data sharing [11]. Reaction Miner [30] and automated methods using Transformer-based ChemBERT improve product identification and role labeling [9]. ReactIE [29] employs weak supervision with linguistic cues to enhance extraction. LLMs facilitate zero-shot NER, IUPAC-to-SMILES conversion, and atom mapping, adding 26% new reactions from patents [22].

Chemical reaction diagram parsing has advanced recently. ReactionDataExtractor 2.0 [24] combines molecular and arrow detectors with text recognition, but its rule-based system struggles with generalization due to cumulative errors. RxnScribe [17] redefines diagram parsing as a sequence generation task using Pix2Seq [4], creating a RxnDP dataset from real papers. RxnIM [5] introduces LVLMs, using synthetic data to address scarcity and employing a three-stage training strategy: 1) Pre-localization for molecular detection; 2) Full parsing with synthetic data; 3) Fine-tuning with real data. Despite high computational demands, it only marginally improves parsing accuracy over RxnScribe and struggles with out-of-distribution samples (as detailed in § 4.2 of the Main paper).

1.2. LVLM and Visual Prompt

Recently, Large Vision-Language Models (LVLMs) [2, 6, 10] built on Large Language Models (LLMs) [1, 20, 21] have made significant strides. These models excel in visual perception [14], visual question answering (VQA) [27], and multimodal reasoning [8, 28], with applications in medicine [13], autonomous driving [7], remote sensing [15, 16], and OCR [23].

Visual prompts [25] provide input in visual forms (e.g., images, selections, markings) in LVLMs, enhancing fine-grained understanding and task adaptability. Unlike text prompts, visual prompts directly affect the input image, improving visual attention and reducing hallucinations. Set-of-Mark (SoM) [26] demonstrated the effectiveness of "discrete markings" by overlaying symbols on images to guide GPT-4V in tasks like directional question-answering without fine-tuning, establishing a "marking as prompt" zero-shot paradigm. Vip-llava [3] allows models to understand arbitrary visual prompts, enabling region-specific question-answering without additional fine-tuning, and introduces the ViP-Bench for evaluation. Image-of-Thought (IoT) [31] suggests models plan a multi-step visual-text reasoning chain, dynamically generating necessary masks or text markings using external tools (e.g., SAM [12], OCR), achieving an automated "prompt as program" process.

2. Details for Pilot Study

2.1. Task Definition of RxnDP

We define the Reaction Diagram Parsing (RxnDP) task as $R = \mathcal{F}(I)$, where I represents the input chemical reaction diagram, $\mathcal{F}$ is the RxnDP model, and $R = \{R_i\}$ is the set of all chemical reactions in the diagram.

Each reaction R_i consists of three roles:

$$R_i = \left(\underset{\text{reactants}}{\mathcal{C}_{\text{react}}}, \underset{\text{conditions}}{\mathcal{C}_{\text{cond}}}, \underset{\text{products}}{\mathcal{C}_{\text{prod}}} \right)$$

Each role $\mathcal{C}_{\text{role}} = \{c_j\}$ (role $\in \{\text{react}, \text{cond}, \text{prod}\}$) is composed of several components. The components can be in two modalities: molecular structure diagrams and text. A molecular structure diagram component (referred to as molecular component) is represented by the bounding box coordinates of the molecule in the diagram, while a text component is represented directly by its textual content.

2.2. Prompts for LVLM inference of Pilot Study

The prompt templates for VQA are shown in Table 1 (reaction counting), Table 2 (molecule counting), Table 3 (cyclic reaction detection), and Table 4 (tree structure detection). These VQA tasks are primarily designed to evaluate the model's basic visual question answering capabilities regarding chemical diagrams.

Prompt: Reaction Counting

You are given an image containing one or more chemical reaction equations. Your task is to count how many distinct reaction equations are present in the image. Output a JSON object with a single key 'reaction_count' and an integer value indicating the number of reactions. For example: {"reaction_count": 2}.
Do not include any explanation or additional text.
<image>Now output your JSON format result:

Table 1. Prompt template for counting reaction

Prompt: Molecular Structure Counting

You are given an image containing one or more chemical reaction equations. Your task is to count how many distinct molecule structures (i.e., chemical structure diagrams) are present in the image. Output a JSON object with a single key 'structure_count' and an integer value indicating the number of molecular structures. For example: {"structure_count": 4}.
Do not include any explanation or additional text.
<image>Now output your JSON format result:

Table 2. Prompt template for counting molecular structures

Prompt: Cyclic Reaction Detection

You are given an image of a chemical reaction. Your task is to determine whether the image contains a cyclic (graph-style) chemical reaction diagram. Output a JSON object with a single key 'cyclic' and a boolean value indicating whether the image contains a cyclic (graph-style) chemical reaction diagram. If the image contains a cyclic reaction, output: {"cyclic": true}. Otherwise, output: {"cyclic": false}.
Do not include any explanation or additional text.
<image>Now output your JSON format result:

Table 3. Prompt template for detecting cyclic chemical reactions

The Prompt template for BROS pattern is shown in Table 5. The prompt template for BIVP is shown in Table 6.

2.3. Metric and Detailed Results of Pilot Study

HardMatch: A strict reaction-level matching strategy that requires all components of a reaction, including molecular structures and textual elements in reactants, products, and conditions, to be correctly matched. Each predicted component must have an IoU >0.5 with its ground truth counterpart. A reaction is counted as a true positive only if all associated components meet this criterion.

SoftMatch: A relaxed reaction-level matching strategy that ignores all text components and merges conditions into reactants during evaluation. A reaction is counted as a true positive only if all molecular structures in the reactants and products achieve IoU >0.5 with the ground truth. This strategy focuses solely on structural alignment.

Prompt: Tree Reaction Detection

You are given an image of a chemical reaction. Your task is to determine whether the image contains a tree (tree-style) chemical reaction diagram. Output a JSON object with a single key 'tree' and a boolean value indicating whether the image contains a tree (tree-style) chemical reaction diagram. If the image contains a tree reaction, output: {"tree": true}. Otherwise, output: {"tree": false}. Do not include any explanation or additional text.
<image> Now output your JSON format result:

Table 4. Prompt template for detecting tree-structured chemical reaction

Prompt: BROS pattern

You are given an image containing one or more chemical reaction equations. Each equation has three parts: reactants, conditions, and products. Each part may include multiple objects, and each object is either a structure, text, identifier, or supplement. Please extract all objects and their bounding boxes, and return them in the following strict JSON format: a list, where each element represents a reaction with keys 'reactants', 'conditions', and 'products'. Each key maps to a list of objects, and each object has:
- category: one of "structure", "text", "identifier", or "supplement"
- bbox: a list of four normalized values [x1, y1, x2, y2], representing the bounding box of the object relative to the image size. Each value must be between 0 and 1.
The coordinates must satisfy: x1 <x2, y1 <y2. (x1, y1) corresponds to the top-left corner, and (x2, y2) to the bottom-right corner of the object in the image.

```
[{
  "reactants": [{"category": "structure",
        "bbox": [0.1, 0.2, 0.3, 0.4]}],
  "conditions": [{"category": "text",
        "bbox": [0.32, 0.21, 0.4, 0.25]}],
  "products": [{"category": "structure",
        "bbox": [0.45, 0.2, 0.6, 0.4]}]
}]
```

Output only the JSON. Do not include any explanation or additional text.
<image>Now output your JSON format result:

Table 5. Prompt for BROS pattern

Prompt: BIVP pattern

You are an expert in chemical image structure analysis. You will be given an image in which all "molecular structures" have been boxed and numbered. Your task is to identify and reconstruct the chemical reaction(s) in the image based on these boxed structures. Please follow the rules below:
1. Each reaction must contain three fields: reactants, conditions, products (each is a List[Dict])
2. Each Dict must include:
- type: one of "mol", "txt", or "idt"
- if type is "mol", second field is `index` (box ID);
- if type is "txt" or "idt", second field is `content` (raw text)
3. Boxed molecular structures must be type "mol". Other elements must use "txt" or "idt"
4. Ignore decorations, arrows, and illustrations; only extract real elements
5. Clean up identifiers and retain only the core ID
6. If no reaction exists in the image, return an empty list []
7. Use arrow direction: tail = reactants, head = products

Example format:

```
[
  {
    "reactants": [
    {"type": "mol", "index": 1},
    {"type": "txt", "content": "NaCl"}
    ],
    "conditions": [
    {"type": "mol", "index": 3},
    {"type": "txt", "content": "H2O, 25°C"}
    ],
    "products": [
    {"type": "mol", "index": 2},
    {"type": "idt", "content": "1a"}
    ]
  }
]
```

Do not include any explanations, comments, or non-structured data.
<image> Please extract the list of reactions from the image. Return only the JSON reaction list without adding any other content.

Table 6. Prompt for BIVP pattern

The results of RxnDP task on RxnScribe-test-slct for General-purpose VLM in VQA task, BROS strategy, and RxnDP task in BIVP strategy are shown in Tables 9, 7, and 8, respectively.

Model	Hard Match			Soft Match		
	P	R	F1	P	R	F1
Gemini-2.5-Pro	0.7	0.6	0.6	38.2	33.0	35.4
GPT4o-2024-11-20	0.4	0.3	0.3	0.4	0.3	0.3
QwenVL-Max	0	0	0	5.0	4.0	4.4

Table 7. Zero-shot results of RxnDP task on RxnScribe-test-slct for General-purpose VLM in BROS strategy

Model	Hard Match			Soft Match		
	P	R	F1	P	R	F1
Gemini-2.5-Pro	17.2	18.8	17.9	77.8	84.5	81.0
GPT4o-2024-11-20	9.9	9.7	9.8	59.7	55.7	57.6
QwenVL-Max	14.5	13.4	13.9	69.6	63.6	66.5

Table 8. Zero-shot results of RxnDP task on RxnScribe-test-slct for General-purpose VLM in BIVP strategy

VQA Question	Gemini-2.5-Pro	GPT4o-2024-11-20	QwenVL-Max
How many reactions are in the image?	72.72	61.81	74.54
How many molecular structures are in the images?	66.36	25.45	56.36
Are there any cyclic reactions?	91.82	87.27	77.27
Are there any tree-structured reactions?	72.73	71.82	86.36
Average	75.91	61.59	73.63

Table 9. Zero-shot results of VQA task on RxnScribe-test-slct for General-purpose VLM

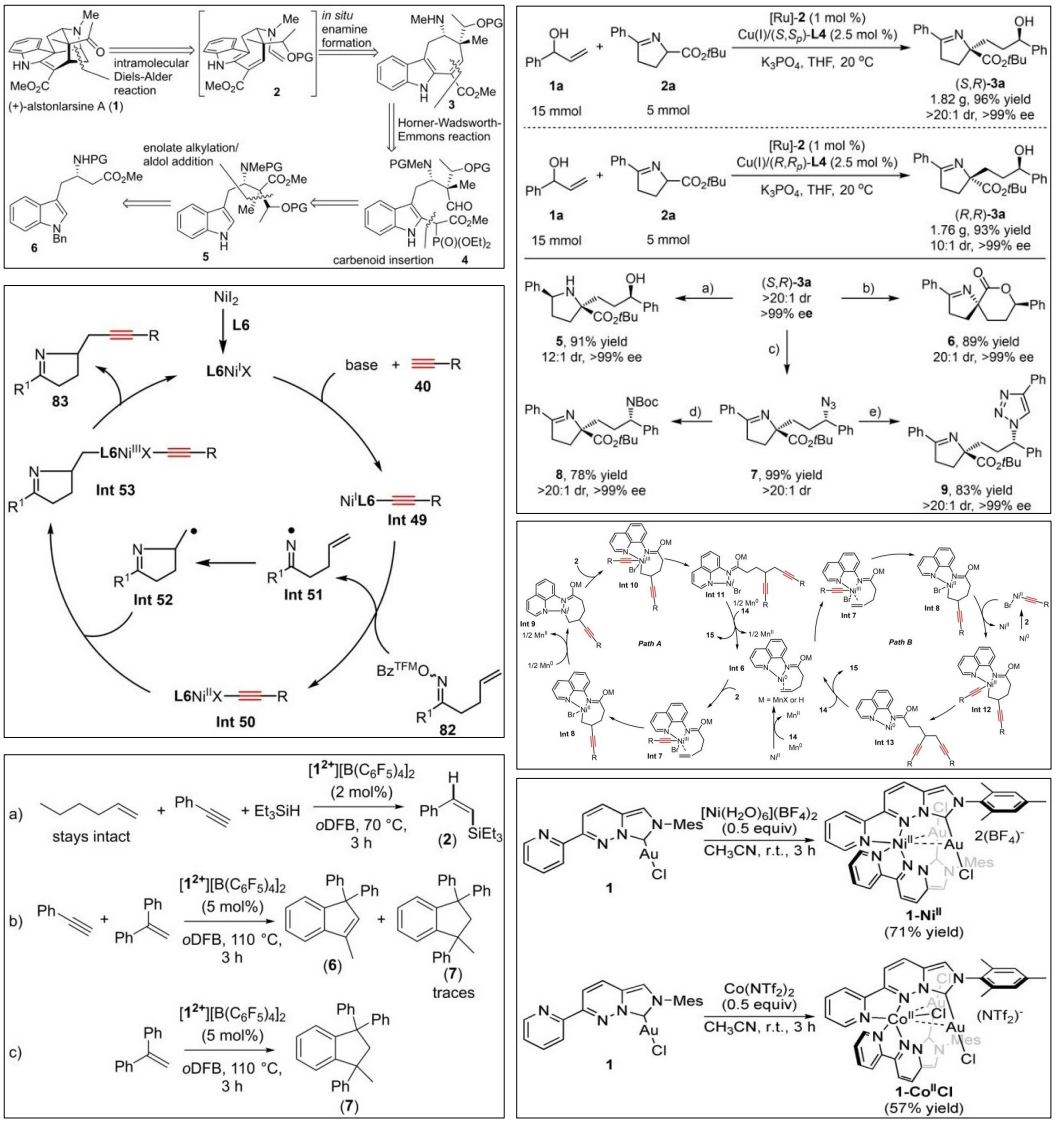

Figure 1. Visualization of four layouts of U-RxnDiagram-15k's training data.

3. Details of U-RxnDiagram-15k

Existing RxnDP datasets have notable limitations: the `RxnScribe` dataset, derived from real papers, includes only 1378 samples, while the `RxnIM` dataset, though large, is artificially synthesized, limiting diversity and model generalization (see § 4.2 of the Main paper). To overcome these issues, we developed the `U-RxnDiagram-15k` dataset, which is a large, high-quality, and diverse RxnDP dataset. The layout of training data of `U-RxnDiagram-15k` is shown in Figure 1.

Journal	Number of Papers
Organic Letters	661
The Journal of Organic Chemistry	518
Journal of the American Chemical Society	387
Angewandte Chemie International Edition	373
Chemical Communications	222
Chemical Science	164
ACS Catalysis	155
Chemistry – A European Journal	126
Nature Communications	94
Nature Chemistry	35
Science	26
Nature Synthesis	24
Nature	16
Chem	15
Nature Catalysis	14
CCS Chemistry	12
JACS Au	7
Science Advances	7
Accounts of Chemical Research	3
Chemical Reviews	1
Communications Chemistry	1
Chemical Society Reviews	1

Table 10. Journal distribution of the U-RxnDiagram-15k dataset

Year	Number of Papers
2021	36
2022	978
2023	1310
2024	513
2025	25

Table 11. Yearly distribution of papers in the U-RxnDiagram-15k dataset

U-RxnDiagram-15k comprises 2,862 organic chemistry papers (2021–2025) split temporally into training (pre-July 2024) and test sets. The detailed statistics of source are in Table 10, 11 and 12. Our future work will thus focus on expanding the dataset's temporal diversity.

Publisher	Number of Papers
American Chemical Society (ACS)	1732
Wiley	499
Royal Society of Chemistry (RSC)	387
Springer Science and Business Media LLC	184
American Association for the Advancement of Science (AAAS)	33
Elsevier BV	15
Chinese Chemical Society	12

Table 12. Publisher distribution of papers in the U-RxnDiagram-15k dataset

For the `U-RxnDiagram-15k` dataset, each image contains annotated bounding boxes (bboxes) that identify specific regions of interest within chemical diagrams. We extract the bounding box coordinates and crop the corresponding image regions, then apply a specialized OCR prompt(see Table 13) to each cropped region using Gemini-2.5-Pro.

The visualization of chemical reaction processes typically follows specific logical layouts to accurately convey their intrinsic chemical relationships. There are four primary types of layouts, as shown in Figure 2: 1) the Single-line Layout, which is characterized by the linear display of a series of continuous chemical transformations along a single path, sequentially arranging starting materials, intermediates, and products to form an unbroken reaction chain; 2) the Multiple-line Layout, which structurally breaks down a long reaction route into several consecutive linear segments arranged in sections; 3) the Tree Layout, which features multiple independent reaction pathways diverging from a common starting material or intermediate to different products, forming a one-to-many radial structure; and 4) the Cyclic Layout, which uses a closed-loop diagram to describe the process of a species undergoing a series of transformations and ultimately regenerating to its initial state.

4. Details of the experiments

4.1. Results Categorized into Four Layout Types

We conducted a detailed analysis on four layout categories using four trained models from the main experiment table and one top-performing zero-shot VLM, Gemini-2.5-Pro. The results are shown in Figure 3. From the figures, we can observe that, when categorized by layout type, the difficulty of the RxnDP task follows the order: **Single-line** $<$ **Multi-line** $<$ **Tree** $<$ **Cyclic**. Performance across all categories is generally slightly lower on `U-RxnDiagram-15k-test`, which is due to our image crop being performed at the figure/table level, resulting in greater diversity and realism. In terms of individual model performance, our **RxnCaption-VL** demonstrates

Prompt: Gemini OCR

OCR_PROMPT = "You are a chemistry OCR expert. Analyze the given image snippet from a chemical diagram and output the most accurate text transcription.

Step 1: Determine content type

- If the image shows a **graphical structure** (e.g., bond lines, benzene rings, molecule drawings), return `[GRAPHICAL_STRUCTURE]`.
- If it is **textual content** (e.g., reagents, conditions, labels, simple structural formulas), transcribe it exactly as seen.

Step 2: OCR transcription rules

1. **Preserve chemical formatting:**
   - Keep all subscripts and superscripts in their Unicode form.
   - Example: H_2O, Fe^{3+}, SO_4^{2-}, S_n2, not `H^2O`, `S_n2`, or `Fe<sup>3+</sup>`.
2. **Maintain exact case:**
   - Output the text **exactly as it appears**, preserving all uppercase/lowercase letters.
   - Example: `Pd` $\neq$ `pd`, `tBu` $\neq$ `TBU`.
3. **Preserve all symbols and operators:**
   - Keep all characters such as +, −, ⊕, ⊖, °, /, ·, →, ⇌, =, (), [], {}, and charge signs exactly as shown.
   - **Do not reposition** any sign or charge, maintain their original location relative to the molecule or atom.
4. **Handle ions, radicals, and special symbols correctly:**
   - Keep radicals, dots, and charges (e.g., •, $^+$, $^-$) in their exact positions relative to the molecule.
   - If the symbol appears **directly above** the molecule or its position is unclear, place it at the **molecule's upper-right corner by default** (e.g., CH_3•, `Mes-Acr-`BF_4 *).
   - Do **not** generate or alter characters, e.g., if the image shows `CyS`•, do **not** output `CysS`•.
5. **Preserve spacing and punctuation:**
   - Do not insert or remove spaces unless clearly visible in the image.
   - Example: $NaBH_4$`/THF` stays as is; not $NaBH_4$ `/ THF`.
6. **Support simple structural notations:**
   - If a short structural fragment appears (e.g., `[Si]-H`, CH_2=CH_2, `Ph-OH`), transcribe it **as text** rather than `[GRAPHICAL_STRUCTURE]`.
7. **Multi-line handling:**
   - If text spans multiple lines, combine them into one line separated by spaces, keeping all internal formatting, subscripts, and symbols unchanged.
8. **No interpretation or correction:**
   - Do not normalize chemical names, fix spelling, or guess missing text.
9. **Exclude irrelevant graphical or contextual elements:**
   - Do **not** include arrows, labels, molecule indices (e.g., "1a", "2b"), or non-textual annotations as part of the OCR output.

Examples:

- Image shows a benzene ring → **Output**: `[GRAPHICAL_STRUCTURE]`
- Image shows a molecule labeled "1a" → **Output**: `[GRAPHICAL_STRUCTURE]`
- Image shows text: `Pd/C \n`H_2 → **Output**: `Pd/C` H_2
- Image shows two lines: NaH and THF → **Output**: `NaH THF`
- Empty or unreadable image → **Output**: ''

Final Output: Return only the recognized text or `[GRAPHICAL_STRUCTURE]`. If recognition is uncertain or blank, return an empty string."

Output only the JSON. Do not include any explanation or additional text. <image> Now output your JSON format result:

Table 13. Prompt for Gemini OCR

more balanced results across all categories, especially on `RxnScribe-test`, whereas Gemini-2.5-Pro exhibits a more pronounced imbalance in performance across different categories.

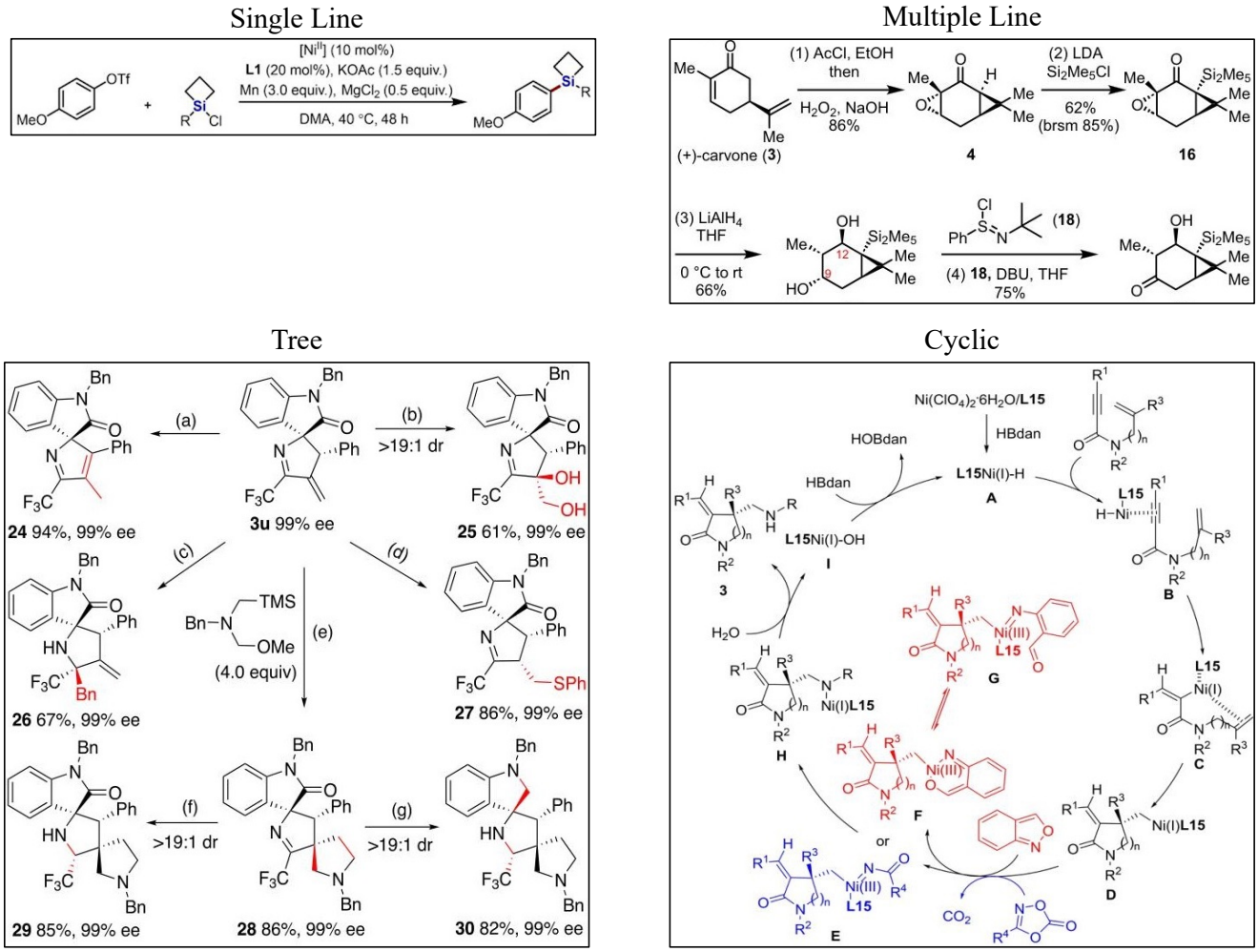

Figure 2. Four layouts of Reaction Diagram.

Detector Setup	Rxn Extractor	RxnScribe-test						U-RxnDiagram-15k-test					
		Soft Match			Hybrid Match			Soft Match			Hybrid Match		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
MolYOLO	Gemini-2.5-pro	67.9	86.5	76.1	44.7	56.1	49.8	64.2	69.2	66.2	38.9	42.1	40.4
	RxnCaption-VL	85.3	87.1	86.2	71.6	72.7	72.2	71.3	69.4	70.4	60.3	59.3	59.8
GT bbox + MolYOLO Recall	Gemini-2.5-pro	74.5	91.7	82.2	45.9	58.7	51.5	70.6	76.6	73.5	44.7	48.3	46.4
	RxnCaption-VL	87.2	88.1	87.6	73.3	73.5	73.4	75.9	76.7	76.3	63.8	63.9	63.8
MolYOLO	Ground Truth	100.0	99.4	99.7	100.0	99.4	99.7	100.0	91.0	95.3	100.0	91.0	95.3

Table 14. Detailed results (Table 6 in full paper) of Error Attribution Analysis of BIVP Strategy in ablation studies.

Besides, by comparing the pix2seq-based models (RxnScribe-official and RxnScribe_w/15k) with the VLM-trained models, we observe that traditional approaches maintain relatively stable performance on categories such as single-line and multi-line, but experience a significant drop when handling more complex images like tree and cyclic structures. This suggests that the VLM's strong visual-language understanding capability can greatly benefit reaction diagram parsing tasks involving intricate graph structures. Moreover, by comparing the performance of RxnScribe-official and RxnScribe_w/15k across the two test sets, we observe that our proposed `U-RxnDiagram-15k` not only yields substantial improvements on its corresponding test set, but also leads to noticeable gains on the

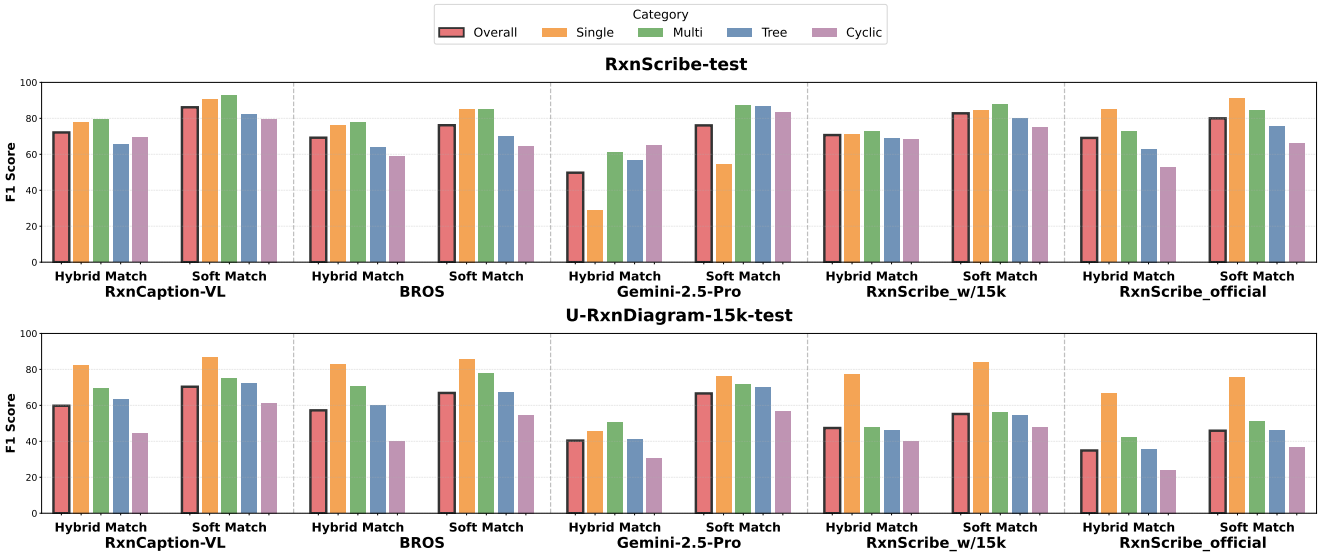

Figure 3. Results categorized into four layout types on `RxnScribe-test` and `U-RxnDiagram-15k-test`.

LVLM	Molecular Detector	RxnScribe-test						U-RxnDiagram-15k-test					
		Soft Match			Hybrid Match			Soft Match			Hybrid Match		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Gemini-2.5-pro	**YoDe**	49.4	60.7	54.5	30.1	41.0	34.7	58.9	59.9	59.4	35.0	37.1	36.0
	MolDetect	64.3	82.6	72.3	42.5	54.9	47.9	59.2	62.7	60.9	36.7	38.9	37.8
	MolYOLO	67.9	86.5	76.1	44.7	56.1	49.8	64.2	69.2	66.6	38.9	42.1	40.4
RxnCaption-VL	**YoDe**	63.4	59.7	61.5	54.3	52.4	53.3	62.2	58.3	60.2	49.9	48.5	49.2
	MolDetect	84.1	84.7	84.4	70.9	70.7	70.8	65.3	62.8	64.0	54.0	52.5	53.2
	MolYOLO	85.3	87.1	86.2	71.6	72.7	72.2	71.3	69.4	70.4	60.3	59.3	59.8

Table 15. Detailed results (Table 5 in full paper) of Influence of Molecular Detector in ablation studies.

`RxnScribe-test`. This demonstrates the strong generalizability of our constructed training dataset.

4.2. Detailed Results of Ablation Studies

For ablation experiments **Error Analysis of BIVP Strategy** and **Influence of Molecular Detector** in the main text, we present the detailed results in Table 14 and Table 15, which include the Precision, Recall, and F1 scores for each dataset and each evaluation metric.

4.3. Error Analysis of 2-Stage Pipeline

Table 16. Stage-wise Error Analysis.

Error Source	RxnScribe		RxnCaption-15k	
	Metric	Drop (↓)	Metric	Drop (↓)
Upper Bound	100.0	-	100.0	-
Detection Error (Ideal Extractor)	99.7	**0.3**	95.3	**4.7**
Assignment Error (Soft-F1)	86.2	**13.5**	70.4	**24.9**
Textual Error (Hybrid-F1)	72.2	**14.0**	59.8	**10.6**

Our qualitative analysis reveals that failures primarily stem from complex topologies, particularly in dense cyclic reaction flows (as we showed in Sec 4.3.3), and unconventional visual symbols, such as lightbulb or lightning icons representing reaction conditions instead of text. Additionally, we observed minor limitations in MolYOLO regarding pseudo-3D structures and rare detection misses.

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023. 1

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 1

[3] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12914–12923, 2024. 1

[4] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021. 1

[5] Yufan Chen, Ching Ting Leung, Jianwei Sun, Yong Huang, Linyan Li, Hao Chen, and Hanyu Gao. Towards large-scale chemical reaction image parsing via a multimodal large language model. *arXiv preprint arXiv:2503.08156*, 2025. 1

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 1

[7] Zhizhao Duan, Hao Cheng, Duo Xu, Xi Wu, Xiangxie Zhang, Xi Ye, and Zhen Xie. Cityllava: Efficient fine-tuning for vlms in city scenario, 2024. 1

[8] Junyuan Gao, Jiahe Song, Jiang Wu, Runchuan Zhu, Guanlin Shen, Shasha Wang, Xingjian Wei, Haote Yang, Songyang Zhang, Weijia Li, Bin Wang, Dahua Lin, Lijun Wu, and Conghui He. Pm4bench: A parallel multilingual multi-modal multi-task benchmark for large vision language model, 2025. 1

[9] Jiang Guo, A Santiago Ibanez-Lopez, Hanyu Gao, Victor Quach, Connor W Coley, Klavs F Jensen, and Regina Barzilay. Automated chemical reaction extraction from scientific literature. *Journal of chemical information and modeling*, 62 (9):2035–2045, 2021. 1

[10] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1

[11] Steven M Kearnes, Michael R Maser, Michael Wleklinski, Anton Kast, Abigail G Doyle, Spencer D Dreher, Joel M Hawkins, Klavs F Jensen, and Connor W Coley. The open reaction database. *Journal of the American Chemical Society*, 143(45):18820–18826, 2021. 1

[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. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 1

[13] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day, 2023. 1

[14] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), 2024. 1

[15] Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model, 2024. 1

[16] Chao Pang, Xingxing Weng, Jiang Wu, Jiayu Li, Yi Liu, Jiaxing Sun, Weijia Li, Shuai Wang, Litong Feng, Gui-Song Xia, and Conghui He. Vhm: Versatile and honest vision language model for remote sensing image analysis, 2024. 1

[17] Yujie Qian, Jiang Guo, Zhengkai Tu, Connor W Coley, and Regina Barzilay. Rxnscribe: a sequence generation model for reaction diagram parsing. *Journal of chemical information and modeling*, 63(13):4030–4041, 2023. 1

[18] Nadine Schneider, Daniel M Lowe, Roger A Sayle, Michael A Tarselli, and Gregory A Landrum. Big data from pharmaceutical patents: a computational analysis of medicinal chemists' bread and butter. *Journal of medicinal chemistry*, 59(9):4385–4402, 2016. 1

[19] Nadine Schneider, Nikolaus Stiefl, and Gregory A Landrum. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of chemical information and modeling*, 56(12):2336–2346, 2016. 1

[20] Jing Shao, Siyu Chen, Yangguang Li, Kun Wang, Zhenfei Yin, Yinan He, Jianing Teng, Qinghong Sun, Mengya Gao, Jihao Liu, Gengshi Huang, Guanglu Song, Yichao Wu, Yuming Huang, Fenggang Liu, Huan Peng, Shuo Qin, Chengyu Wang, Yujie Wang, Conghui He, Ding Liang, Yu Liu, Fengwei Yu, Junjie Yan, Dahua Lin, Xiaogang Wang, and Yu Qiao. Intern: A new learning paradigm towards general vision, 2022. 1

[21] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[22] Sarveswara Rao Vangala, Sowmya Ramaswamy Krishnan, Navneet Bung, Dhandapani Nandagopal, Gomathi Ramasamy, Satyam Kumar, Sridharan Sankaran, Rajgopal Srinivasan, and Arijit Roy. Suitability of large language models for extraction of high-quality chemical reaction dataset from patent literature. *Journal of Cheminformatics*, 16(1):131, 2024. 1

[23] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun,

Yuang Peng, Chunrui Han, and Xiangyu Zhang. General ocr theory: Towards ocr-2.0 via a unified end-to-end model, 2024. 1

[24] Damian M Wilary and Jacqueline M Cole. Reactiondataextractor 2.0: a deep learning approach for data extraction from chemical reaction schemes. *Journal of Chemical Information and Modeling*, 63(19):6053–6067, 2023. 1

[25] Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, et al. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*, 2024. 1

[26] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023. 1

[27] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2024. 1

[28] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark, 2024. 1

[29] Ming Zhong, Siru Ouyang, Minhao Jiang, Vivian Hu, Yizhu Jiao, Xuan Wang, and Jiawei Han. Reactie: Enhancing chemical reaction extraction with weak supervision. *arXiv preprint arXiv:2307.01448*, 2023. 1

[30] Ming Zhong, Siru Ouyang, Yizhu Jiao, Priyanka Kargupta, Leo Luo, Yanzhen Shen, Bobby Zhou, Xianrui Zhong, Xuan Liu, Hongxiang Li, et al. Reaction miner: An integrated system for chemical reaction extraction from textual data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 389–402, 2023. 1

[31] Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. Image-of-thought prompting for visual reasoning refinement in multimodal large language models. *arXiv preprint arXiv:2405.13872*, 2024. 1