RobustRDP: Advancing Reaction Diagram Parsing via Synthetic-to-Real Data Scaling and Robustness-Oriented Training

Jianting Tang[1,2], Zezhong Wu[1,2], and Linli Xu[1,2(✉)]

[1] University of Science and Technology of China, Hefei, China
[2] State Key Laboratory of Cognitive Intelligence, Hefei, China
{jiantingtang,felix}@mail.ustc.edu.cn, linlixu@ustc.edu.cn

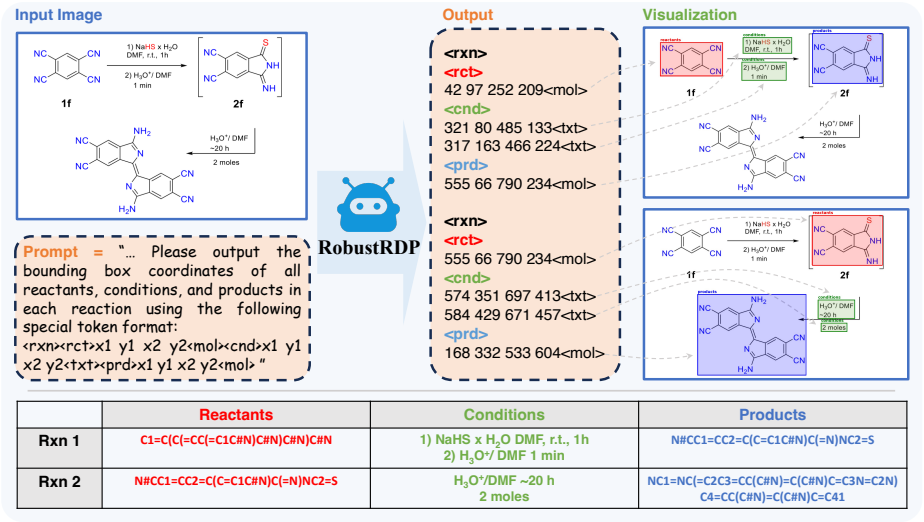
Abstract. Chemical reaction diagram parsing aims to automatically convert reactions from images into machine-readable formats. Current leading approaches employ end-to-end generative models to directly decode reaction diagrams into structured coordinate sequences, performing object localization and relationship modeling in a unified process. However, two critical bottlenecks remain: severe **data scarcity** due to the high cost of expert annotation, and **output instability** caused by the complex sequential dependencies between chemical reactions, where minor prefix errors can lead to cascading parsing failures. In this paper, we propose **RobustRDP**, a **Robust R**eaction **D**iagram **P**arser built on a multimodal large language model (MLLM). Its superior performance primarily stems from two advancements: **Synthetic-to-Real Data Scaling**: We develop a layout-driven synthesizer and an efficient annotation platform to create a large-scale training set, along with a more comprehensive evaluation benchmark. **Robustness-Oriented Training**: We propose a three-stage progressive training strategy (Pretraining, Multi-Task SFT, DPO). The SFT stage employs two auxiliary tasks including *region-guided reaction parsing* and *prefix-perturbed reaction parsing* to mitigate sequential dependencies between reactions and enhance the model's robustness against prior errors. The DPO stage further stabilizes outputs in failure-prone scenarios. Extensive experiments demonstrate the superior performance of RobustRDP, establishing a solid foundation for automated reaction diagram parsing. Code and model weights are available at https://github.com/jaydetang/RobustRDP.

Keywords: Reaction Diagram Parsing · Large Language Models

1 Introduction

The task of reaction diagram parsing focuses on converting reactions from images in chemical literature into structured, machine-readable formats. As diagrams are the primary form of conveying complex reaction pathways, the ability to parse them automatically is essential for building large-scale chemical

J. Tang and Z. Wu — Equal contribution.
L. Xu — Corresponding author.



	Reactants	Conditions	Products
Rxn 1	C1=C(C(=CC(=C1C#N)C#N)C#N)C#N	1) NaHS x H_2O DMF, r.t., 1h 2) H_3O^+/ DMF 1 min	N#CC1=CC2=C(C=C1C#N)C(=N)NC2=S
Rxn 2	N#CC1=CC2=C(C=C1C#N)C(=N)NC2=S	H_3O^+/DMF ~20 h 2 moles	NC1=NC(=C2C3=CC(C#N)=C(C#N)C=C3N=C2N)C4=CC(C#N)=C(C#N)C=C41

Fig. 1: RobustRDP generates object coordinates in a predefined format, allowing easy identification of reactants, conditions, and products for each reaction. Molecule and text object content can be then recognized using OCSR and OCR technologies.

databases [10, 14], advancing computer-aided synthesis planning [20, 28], and expediting automated drug discovery [26, 46]. Specifically, this task requires precisely identifying the reactants, conditions, and products associated with each reaction. However, the diverse layout configurations and complex reaction pathways exhibited in diagrams make accurate parsing highly challenging.

Traditional approaches primarily rely on cascaded pipelines and hand-crafted heuristics, such as ReactionDataExtractor [40, 41] and OChemR [24]. They first use object detection [32, 33] models to localize key elements in reaction diagrams, such as molecules, text, and arrows. Then, heuristic rules based on geometric relationships and arrow directions are applied to identify the chemical relationships, determining which objects are reactants, conditions, and products in each reaction. However, this approach suffers from cross-stage error accumulation and limited generalization, resulting in degraded performance. Current leading approaches reformulate reaction diagram parsing as a structured coordinate sequence generation task, as illustrated in Fig. 1. By generating a series of object coordinates in a predefined format, these approaches unify object localization and relationship modeling into an end-to-end process, yielding substantial performance gains. RxnScribe [29] pioneers this paradigm by adopting Pix2Seq [5] as the generative model backbone. RxnIM [6] subsequently leverages MLLMs [4, 39, 42] with a multi-stage SFT training strategy, achieving further improvements. RxnCaption-VL [35] overlays bounding boxes and indices produced by MolYOLO [35] onto images as visual prompts, enabling the MLLM to focus solely on generating object relationship sequences.

Despite these advancements, two critical bottlenecks remain, as illustrated in Fig. 2. The first is severe data scarcity. Annotating object locations and the inter-object relationships in complex reaction diagrams is highly labor-intensive, thus most models still rely on the RxnScribe dataset [29], which contains only 1,240 training images, hardly sufficient for MLLMs training. The second is output instability. The structured coordinate sequence exhibits complex sequential dependencies between reactions, such that local errors can trigger a collapse of the entire sequence. This issue is further exacerbated by the vanilla next-token prediction training: the model may overly rely on object information from preceding reactions (shortcut learning), while being exposed only to ground-truth prefixes (oracle prefix). Consequently, minor prefix errors at inference can propagate and cause cascading failures, leading to unstable outputs.

To address these challenges, we propose RobustRDP, a robust reaction diagram parser built upon an MLLM backbone. To bridge the data gap, we develop a layout-driven synthesizer that mimics the compositional logic of reaction diagrams in chemical literature. By sampling molecules, text, and arrows and arranging them on blank canvases according to rules, the synthesizer generates diverse layouts with ground-truth labels, providing a large pre-training corpus. In addition, we develop an efficient annotation platform integrated with an object detection [34] model to facilitate semi-automated labeling, enabling the creation of a substantially larger dataset for training and evaluation. To mitigate the output instability, we introduce a three-stage robustness-oriented training strategy. During the SFT stage, *region-guided reaction parsing* requires RobustRDP to parse the standalone reaction within given regions, thus preventing it from using preceding reaction cues as shortcuts and compelling it to fully exploit visual information. *Prefix-perturbed reaction parsing* injects simulated errors into preceding sequences, training RobustRDP to recover from errors and maintain stability despite prior inaccuracies. During the DPO stage, we collect erroneous sequences generated by RobustRDP with high probability as rejected responses, further stabilizing outputs in failure-prone scenarios.

In summary, our main contributions are as follows:

- We propose a synthetic-to-real data scaling scheme involving a layout-driven synthesizer and an efficient annotation platform to produce extensive synthetic and real datasets for training and evaluation.
- We introduce a three-stage robustness-oriented training strategy, leveraging auxiliary SFT tasks to mitigate sequential dependencies and DPO to further stabilize outputs.
- Extensive experiments demonstrate the superior performance of RobustRDP, establishing a SOTA foundation for reaction diagram parsing.

2 Related Work

Reaction Diagram Parsing. A complete reaction diagram parsing workflow typically involves object localization, relationship modeling, and content recognition. With the maturity of Optical Chemical Structure Recognition (OCSR) [9,

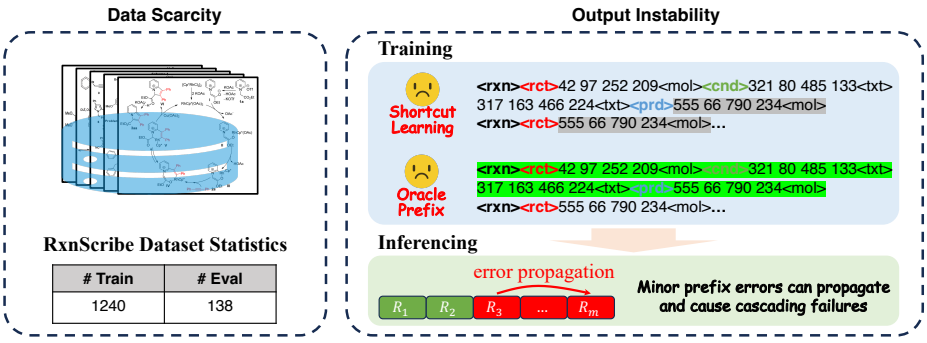


Fig. 2: Two critical bottlenecks in reaction diagram parsing. (Left) Predominant reliance on the RxnScribe dataset provides insufficient training samples for robust generalization. (Right) Complex sequential dependencies, exacerbated by shortcut learning and oracle prefix issues during training, undermine RobustRDP's inference robustness.

22,30,31,43] and Optical Character Recognition (OCR) [11,16,25] technologies, recognizing the content of individual molecules and text is well-established; consequently, the primary bottlenecks lie in the first two steps. Early approaches, such as ReactionDataExtractor [40,41] and OChemR [24] perform object localization and relationship modeling in a cascaded manner. They first employ object detection models, such as Faster R-CNN [33] or DETR [45], to identify the locations of molecules and text, alongside a dedicated arrow detection model to determine the start and end points of arrows. Relationships between objects are then inferred using heuristic rules—for example, objects near the arrow's start are considered reactants, those along the shaft as conditions, and those at the arrow's end as products. Currently, mainstream approaches formulate reaction diagram parsing as a coordinate sequence generation task, unifying object localization and relationship modeling into an end-to-end process. RxnScribe [29] pioneers this paradigm with a Pix2Seq [5] backbone and establishes the first standard training and evaluation datasets. RxnIM [6] upgrades the backbone from the small Pix2Seq model to an MLLM and adopts a multi-stage SFT training strategy, resulting in improved parsing accuracy. RxnCaption-VL [35] incorporates visual prompts by overlaying bounding boxes and indices from MolYOLO [35] onto the image, enabling the MLLM to focus solely on decoding relationship sequences and thus reducing the generation burden. Despite these improvements, robust parsing of diverse diagrams still leaves significant room for progress.

MLLMs. MLLMs [2, 12, 13, 36, 38] have recently achieved significant breakthroughs across a broad range of vision-language tasks [8,15,17]. Early MLLMs, such as Flamingo [1], BLIP-2 [19], and CogVLM [37], experimented with various cross-modal interaction architectures. This gradually converged on the widely adopted Encoder-Projector-LLM architecture, exemplified by models like LLaVA-1.5 [21], Qwen2.5-VL [3], and InternVL2.5 [7]. Within this framework, significant

research specifically focus on enhancing visual grounding capabilities. Models like Shikra [4], Kosmos-2 [27], Ferret-v2 [44], and Groma [23] have integrated coordinate-aware instruction tuning, demonstrating strong object localization and referring expression comprehension. However, these general-domain MLLMs often underperform on reaction diagrams due to a severe distribution shift from natural images. In addition, general MLLMs lack the capacity to perform object localization and relationship modeling simultaneously, which is essential for reaction diagram parsing. Therefore, careful domain adaptation is necessary for general MLLMs to effectively handle this task.

3 Methodology

3.1 Problem Formulation

The objective of reaction diagram parsing is to transform an input image I into a set of chemical reactions $\mathcal{R} = \{R_1, R_2, \ldots, R_n\}$. Each reaction is defined as a tuple $R_i = (\mathcal{S}_{rct}, \mathcal{S}_{cnd}, \mathcal{S}_{prd})$, consisting of three functional roles: reactants ($\mathcal{S}_{rct}$), conditions ($\mathcal{S}_{cnd}$), and products ($\mathcal{S}_{prd}$). Each $\mathcal{S}_{role}$ contains a variable number of objects o_j, represented as $o_j = (b_j, c_j)$, where $b_j = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ denotes the bounding box of the object, and $c_j \in \{\text{molecule}, \text{text}\}$ indicates its semantic category. To enable autoregressive generation with MLLMs, we serialize the structured reaction set $\mathcal{R}$ into a token sequence $\mathcal{Y}$ with explicit structural markers. As illustrated in Fig. 1, each reaction block begins with <rxn>, followed by role-specific sections separated by <rct>, <cnd> and <prd>. Within each section, an object is encoded as its bounding box coordinates followed by a category token (<mol> or <txt>).

3.2 Synthetic-to-Real Data Scaling

To address the severe data scarcity in reaction diagram parsing, we introduce a synthetic-to-real data scaling scheme. Large-scale synthetic images, despite certain distributional differences from real images, allow RobustRDP to acquire fundamental object localization and relationship modeling capabilities, achieving an initial adaptation. Extensive annotated real images extend the RxnScribe-train set, exposing RobustRDP to diverse layouts and stylistic nuances in genuine chemical literature, thereby ensuring robust generalization.

Layout-Driven Synthesizer. Due to the labor-intensive annotation of real reaction diagrams, we develop a layout-driven synthesizer to procedurally generates large-scale, high-quality synthetic data. The synthesis process follows a "sample-render-arrange" pipeline to mimic the compositional logic of real reaction diagrams. First, the synthesizer samples chemical elements, including molecules from the PubChem [18] dataset and text from a corpus of reaction conditions like reagents, catalyst and temperature. Second, the sampled elements are rendered into individual images with high visual variability, where molecules

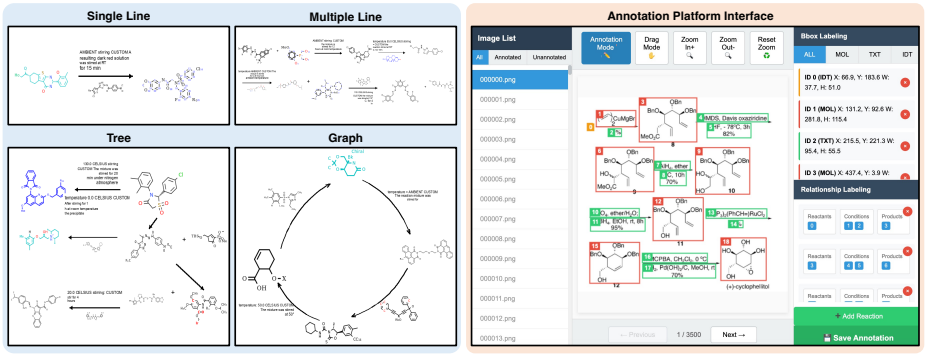


Fig. 3: (Left) Synthetic reaction diagrams with four representative layouts generated by our layout-driven synthesizer. (Right) Our efficient annotation platform, supporting semi-automatic annotation of object bounding boxes and chemical relationships, facilitating the construction of the RobustRDP-train and RobustRDP-test sets.

rendered with the Indigo toolkit are augmented through node shuffling and coordinate scaling, and text adopts diverse fonts and sizes to simulate various publication styles. Finally, the rendered components are placed on a blank canvas, connected by directional arrows and organized into four primary layouts in genuine chemical literature: Single-line for basic linear reactions; Multiple-line for wrapping pathways with shared intermediates across rows; Tree for convergent or divergent branches; and Graph for complex interconnected networks. By recording bounding box coordinates and functional roles during placement, the synthesizer automatically generates large-scale diagrams with ground-truth labels. Some representative synthetic examples are show in Fig. 3 (Left).

Efficient Annotation Platform. To scale annotated real diagrams, we develop an efficient annotation platform tailored for the reaction diagram parsing task. The platform supports labeling both object bounding boxes and object relationships. Given that object detection is a well-established field with many mature models, we leverage RxnScribe-train set and 500 manually labeled images in our early annotation phase to train a YOLO26 [34] model. This establishes a semi-automated labeling workflow, enabling subsequent annotation to focus solely on correcting occasional bounding box errors and labeling object relationships, reducing manual effort by approximately 40%. The annotation platform interface is shown in Fig. 3 (Right).

3.3 Robustness-Oriented Training

Chemical reaction diagram parsing requires precisely generating a series of object coordinates in a predefined format, which exhibits complex sequential dependencies between reactions. To mitigate output instability, we propose a three-stage robustness-oriented training strategy, as illustrated in Fig. 4. The Pretraining

stage enables RobustRDP initially adapt to this task, supporting more efficient subsequent learning on real data. The SFT stage adopting two auxiliary tasks to mitigate output instability arising from the shortcut learning and oracle prefix issues. The DPO stage further enhances robustness by explicitly suppressing RobustRDP's inherent failure modes.

Pretraining. While general MLLMs excel at natural image understanding, they are not well adapted to the abstract symbols and topological structures inherent in chemical reaction diagrams. To bridge this gap, we utilize the large-scale synthetic images generated by our layout-driven synthesizer for pretraining. This stage familiarizes RobustRDP with abstract reaction elements, equipping the MLLM with foundational parsing capabilities and enabling more effective training on real images in subsequent stages. Although synthetic data exhibits a distributional shift from diagrams in chemical literature, its extensive scale, diverse layout coverage, and precise ground-truth labels provide effective supervision. Through the large-scale synthetic pretraining, RobustRDP acquires the capabilities of both object localization and relationship modeling, while adapting to the predefined sequence format that incorporates special tokens as explicit structural markers.

Multi-task SFT. In the Supervised Fine-tuning (SFT) stage, we utilize a combination of real images from our annotated RobustRDP-train set and the RxnScribe-train set. To further expand the dataset, we apply a combination of augmentation techniques following the practices in RxnScribe [29], including multi-image stitching (compositing multiple reaction diagrams into a single diagram), horizontal and vertical flipping, and rotation.

The objective of the primary task, Vanilla Reaction Parsing (VRP), is to optimize the model parameters θ to accurately predict all reactions in the reaction set $\mathcal{R}$, conditioned on the input image I. The training objective is formulated as a next-token prediction task:

$$\mathcal{L}_{VRP} = -\sum_{i=1}^{n} \log P(R_i \mid \mathcal{R}_{<i}, I; \theta) \tag{1}$$

where R_i denotes the i-th reaction and $\mathcal{R}_{<i} = \{R_1, \ldots, R_{i-1}\}$ represents the prefix including all preceding reactions. However, we identify two fundamental limitations in this standard SFT process that can exacerbate output instability:

(a) *Shortcut Learning.* Due to the causal attention mechanism, when learning to predict reaction R_i, the model can attend to additional information from the preceding reactions $\mathcal{R}_{<i}$, such as ground-truth object bounding boxes b_j and categories c_j. Additionally, the products of R_{i-1} often serve as the reactants of R_i. This allows the model to rely on preceding reaction cues as "shortcuts" to infer the current reaction, rather than fully exploiting the visual information in I. Consequently, the model tends to overfit the sequential dependency $P(R_i \mid \mathcal{R}_{<i})$, resulting in deficient local parsing accuracy for individual reactions.

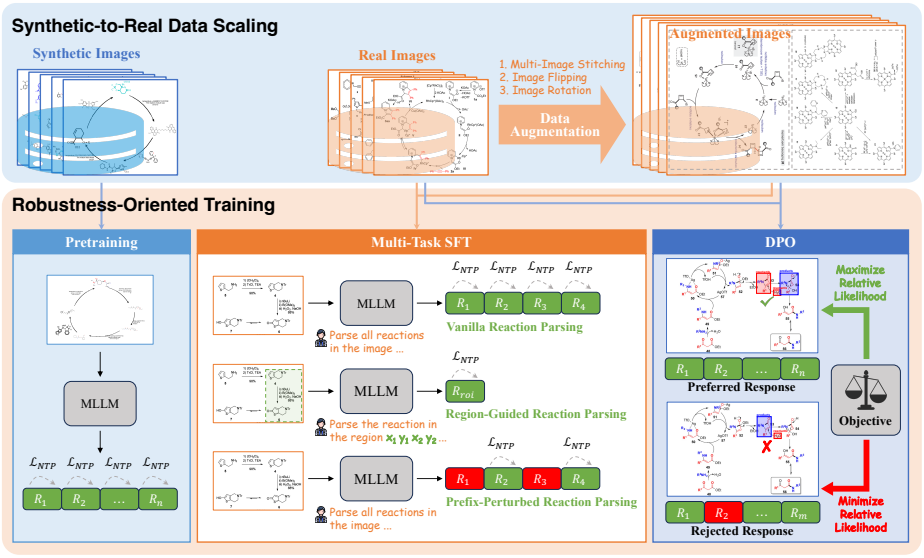


Fig. 4: (Up) Images utilized at each training stage. (Bottom) The three-stage progressive training strategy, comprising Pretraining for initial adaptation, Multi-Task SFT for robustness enhancement, and DPO for explicit failure mode suppression.

(b) *Oracle Prefix.* During training, the prefix sequence consists entirely of ground-truth reactions $\mathcal{R}_{<i}$, thereby forming an oracle prefix. However, during inference, the model relies on its own predictions $\hat{\mathcal{R}}_{<i}$, which may contain errors, forming an imperfect prefix. This train–inference mismatch makes the model highly sensitive to its own mistakes, preventing recovery from early-stage minor errors and easily leading to cascading failures. Collectively, the deficiency in local parsing accuracy, coupled with a high sensitivity to prefix errors, exacerbates output instability during inference. Therefore, to enhance model robustness, we propose two auxiliary SFT tasks:

(i) *Region-Guided Reaction Parsing (RGRP).* To mitigate shortcut learning, this task provides the image I along with an additional region coordinate B_{roi}, which specifies the boundaries of the region of interest encapsulating a single reaction R_{roi}. RobustRDP is required to only generate the reaction R_{roi} within that spatial constraint. Formally, the objective is defined as:

$$\mathcal{L}_{RGRP} = -\log P(R_{roi} \mid B_{roi}, I; \theta) \tag{2}$$

By providing B_{roi}, we disrupt RobustRDP's reliance on the preceding reactions $\mathcal{R}_{<i}$, thereby compelling it to enhance the standalone local parsing accuracy.

(ii) *Prefix-Perturbed Reaction Parsing (PPRP).* To mitigate the oracle prefix and the resulting train–inference mismatch problem, we simulate imperfect prefixes during training. Given the ground-truth sequence $\mathcal{R} = \{R_1, \ldots, R_n\}$, we randomly select a subset of indexes $\mathcal{P} \subset \{1, \ldots, n\}$ and apply rule-based perturbations to corresponding reactions, ultimately constructing a perturbed

sequence $\tilde{\mathcal{R}}$. These perturbations include randomly scaling object coordinates, omitting essential objects, and injecting redundant objects to simulate common inference errors. RobustRDP is then trained to predict only the unperturbed reactions conditioned on the perturbed prefix:

$$\mathcal{L}_{\text{PPRP}} = -\sum_{i \notin \mathcal{P}} \log P\big(R_i \mid \tilde{R}_{<i}, I; \theta\big) \tag{3}$$

By exposing RobustRDP to perturbed prefixes during training, this task improves its robustness to prefix errors. RobustRDP learns to recovery from early-stage minor errors and generate correct subsequent reactions during inference, thereby alleviating cascading failures.

During the SFT stage, the primary reaction diagram parsing task is jointly optimized with the two auxiliary tasks. This multi-task training strategy specifically enhances both local parsing accuracy for individual reactions and robustness to prefix errors, leading to more stable inference.

DPO. While SFT builds a strong foundation for reaction diagram parsing by maximizing the likelihood of ground-truth sequences, this purely imitation objective does not explicitly suppress RobustRDP's inherent failure modes, allowing it to generate certain erroneous predictions with high probability. To further strengthen RobustRDP's robustness, we incorporate Direct Preference Optimization (DPO) to explicitly suppress these predictions.

To construct the preference dataset $\mathcal{D}_{DPO} = \{(I, y_w, y_l)\}$, we pair a preferred response y_w with a rejected response y_l for a given image I. The preferred responses y_w correspond to the ground-truth reaction sequences. The rejected responses y_l are sampled from the model's outputs after the SFT stage on the training set, retaining those with evaluation metrics below a threshold. We optimize the policy π_θ to maximize the margin between the preferred and rejected responses. The DPO loss is formulated as follows:

$$\mathcal{L}_{DPO} = -\mathbb{E}_{(I,y_w,y_l)\sim\mathcal{D}_{DPO}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid I)}{\pi_{ref}(y_w \mid I)} - \beta \log \frac{\pi_\theta(y_l \mid I)}{\pi_{ref}(y_l \mid I)} \right) \right] \tag{4}$$

where π_{ref} denotes the frozen SFT model used as a reference to prevent excessive distribution shift, $\sigma(\cdot)$ denotes the sigmoid function and β is a hyperparameter controlling the sharpness of the preference margin.

Through this stage, RobustRDP further enhances its output robustness in failure-prone scenarios.

4 Experiments

4.1 Experimental Setup

Datasets. The pretraining dataset consists of 60,000 synthetic images generated by our layout-driven synthesizer, specifically including 10,000 single-line images, 10,000 multiple-line images, 20,000 tree images, and 20,000 graph images.

Table 1: Data statistics across the Pretraining, Multi-task SFT, and DPO stages.

	Pretraining	**Multi-task SFT**	**DPO**
Data Source	60,000 Synthetic Images	4,240 Real Images 63,600 Aug. Images	4,240 Real Images 63,600 Aug. Images
Total Samples	60,000	127,200 (VRP) 31,800 (RGRP) 31,800 (PPRP)	14,169

The SFT dataset is built upon a collection of 4,240 real images, which includes 3,000 newly annotated images and 1,240 existing images from the RxnScribe-train [29] set. By applying data augmentations to these real images, we generate an additional 63,600 augmented images. For the primary VRP task, the real and augmented images are mixed at a $1:1$ ratio, resulting in 127,200 training samples. To support the robustness-oriented auxiliary tasks, we further derive 31,800 training samples for RGRP task and 31,800 training samples for PPRP task from the real and augmented images. The DPO dataset consists of 14,169 preference pairs. Specifically, we first perform inference with the SFT model on the real and augmented images, retaining images with predictions yielding a hard match F1 score of less than 0.8. The model's prediction is treated as the rejected response, and the corresponding ground-truth reaction sequence serves as the preferred response. Furthermore, we annotate 500 real images with 2,634 reactions to create the RobustRDP-test benchmark, substantially larger than the RxnScribe-test [29] which contains 138 images with 392 reactions, providing a more comprehensive assessment. The data statistics are shown in Table 1.

Evaluation Metrics. We adopt the same evaluation protocol as RxnScribe [29] to ensure a consistent comparison. Given a ground-truth set $G = \{R_1, \ldots, R_n\}$ and a predicted set $P = \{\hat{R}_1, \ldots, \hat{R}_m\}$, each predicted reaction is compared against all ground-truth reactions. Object-level correspondences are established by matching bounding boxes using Intersection over Union (IoU), where a pair is considered matched if the IoU exceeds 0.5. To accommodate potential annotation ambiguities, there are two metrics of distinct criteria. **Hard Match:** This represents the most strict criterion, where a predicted reaction $\hat{R}$ is considered correct only if all its components—reactants, conditions, and products—exactly match the corresponding objects in the ground truth R. **Soft Match:** Accounting for common annotation ambiguities, such as molecules visually placed as conditions but functioning as reactants, this metric focuses solely on molecule objects and does not distinguish between reactants and reagents in conditions. For both matching criteria, we calculate the precision, recall, and F1 scores. The final performance is reported as micro-averaged metrics across the entire evaluation set.

Table 2: Performance comparison of various reaction diagram parsing models. The evaluation includes Hard Match and Soft Match metrics (Precision, Recall, and F_1 score) on the RxnScribe-test and RobustRDP-test sets.

Dataset	Model	Hard Match			Soft Match		
		Precision	Recall	F_1	Precision	Recall	F_1
RxnScribe -test	RxnDE	4.1	1.3	1.9	19.4	5.9	9.0
	OChemR	4.4	2.8	3.4	12.4	7.9	9.6
	RxnScribe	72.3	66.2	69.1	83.8	76.5	80.0
	RxnIM	74.7	69.7	72.1	86.9	82.8	84.8
	RxnCaption-VL	71.6	72.7	72.2	85.3	87.1	86.2
	RobustRDP	**81.0**	**82.4**	**81.7**	**90.2**	**91.3**	**90.8**
RobustRDP -test	RxnScribe	69.7	61.7	65.5	82.7	72.3	77.2
	RxnIM	57.5	50.2	53.6	76.2	65.8	70.6
	RxnCaption-VL	-	-	-	-	-	-
	RobustRDP	**82.7**	**82.4**	**82.5**	**91.1**	**90.8**	**91.0**

Implementation Details. RobustRDP adopts Qwen2.5-VL-3B-Instruct as the backbone. The model's vocabulary is extended with several special tokens: <rxn>, <rct>, <cnd>, <prd>, <txt>, and <mol>, which serve as explicit structural markers in reaction sequences. The training process follows a three-stage pipeline, all utilizing a cosine learning rate scheduler with a warmup ratio of 0.03. In the pretraining stage, the vision encoder and projector are frozen to optimize only the LLM, which is trained for 1 epoch on the pretraining dataset with a learning rate of 1.0×10^{-6} and a global batch size of 16. In the SFT stage, the primary parsing task is optimized together with the two auxiliary tasks, with full-parameter fine-tuning performed for 1 epoch at a learning rate of 1.0×10^{-5} and a global batch size of 4. In the DPO stage, only the LLM is optimized for 1 epoch with a learning rate of 3.0×10^{-7} and a global batch size of 64.

Baselines. ReactionDataExtractor (RxnDE) [41] and OChemR [24] rely on cascaded pipelines and hand-crafted heuristic. The remaining baselines follow the sequence generation paradigm. RxnScribe [29] employs the Pix2Seq model as its backbone, RxnIM [6] leverages MLLMs, and RxnCaption-VL [35] utilizes MolYOLO to provide visual prompts and predicts only relationship sequences.

4.2 Main Results

Table 2 presents the overall comparison on the RxnScribe-test and RobustRDP-test benchmarks. Traditional approaches like RxnDE and OChemR exhibit significantly lower performance, largely due to their heavy reliance on handcrafted heuristics which fail to generalize to diverse real-world diagrams. RobustRDP

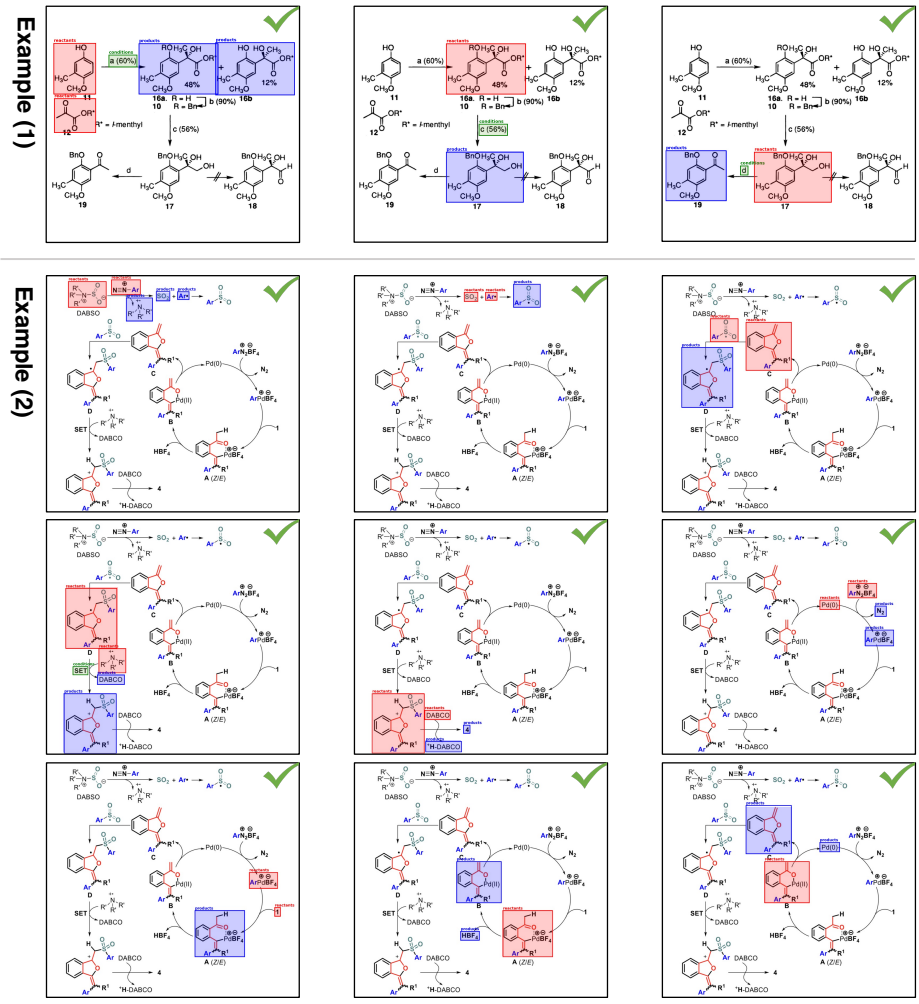


Fig. 5: Representative parsing results of RobustRDP on the RobustRDP-test set. RobustRDP demonstrates robust performance across reaction diagrams of diverse layouts and varying complexity.

consistently outperforms existing sequence-based models, achieving the best performance across all metrics. On the RxnScribe-test benchmark, RobustRDP achieves an F_1 score of 81.7% under the hard match criterion and 90.8% under soft match, demonstrating a substantial improvement over the strongest baseline, RxnCaption-VL, by 9.0% and 4.6% in F_1, respectively. This performance advantage is further maintained on the more challenging RobustRDP-test benchmark, where RobustRDP achieves a hard match F_1 score of 82.5% and a soft match F_1 score of 91.0%, yielding substantial gains of 17.0% and 13.8% in F_1 scores compared to RxnScribe. These results underscore RobustRDP's superior robust-

Table 3: Ablation study on the key method components of RobustRDP. Individual contributions are evaluated by incrementally removing each component. The experimental results are obtained from the RxnScribe-test set.

Method Components					Hard Match			Soft Match		
DPO	Pretrain	RGRP	PPRP	RobustRDP -train	Precision	Recall	F_1	Precision	Recall	F_1
✓	✓	✓	✓	✓	**81.0**	**82.4**	**81.7**	**90.2**	**91.3**	**90.8**
–	✓	✓	✓	✓	80.4	81.4	80.9	89.7	90.3	90.0
–	–	✓	✓	✓	78.2	79.5	78.8	87.9	89.0	88.5
–	–	–	✓	✓	76.7	78.6	77.6	86.9	88.8	87.8
–	–	✓	–	✓	75.2	76.3	75.7	86.2	87.5	86.8
–	–	–	–	✓	74.6	74.7	74.6	86.0	85.7	85.9
–	–	–	–	–	71.1	69.6	70.4	81.5	80.1	80.8

ness to handle the diverse layouts and varying complexity inherent in authentic reaction diagrams. Some parsing results are shown in Fig. 5.

4.3 Ablation Studies

To assess the contribution of each component in the RobustRDP framework, we conduct an incremental ablation study. The results, as shown in Table 3, highlight the impact of our training strategy and data scaling approach.

Effect of DPO. Comparing Row 1 and Row 2 in Table 3, removing the DPO stage causes a noticeable performance drop, with hard match F_1 decreasing from 81.7% to 80.9% (-0.8%) and soft match F_1 decreasing from 90.8% to 90.0% (-0.8%). This demonstrates that DPO effectively enhances RobustRDP's output robustness. By explicitly contrasting correct and incorrect outputs, DPO suppresses RobustRDP's inherent failure modes and mitigates the drawback of SFT learning solely from positive supervision by mimicking ground-truth sequences. Comparative visual examples are presented in Fig. 6 (a).

Effect of Pretraining. As shown in Row 3 in Table 3, removing the pretraining stage further reduces hard match F_1 to 78.8% (-2.1%) and soft match F_1 to 88.5% (-1.5%). This result demonstrates that large-scale synthetic pretraining is essential for initializing the model with fundamental object localization and relationship modeling capabilities, while enabling adaptation to the specialized sequence format with newly introduced structural marker tokens. Such initialization facilitates more efficient and effective learning when fine-tuned on real-world data in subsequent stages.

Effect of RGRP and PPRP. Rows 3 $\sim$ 6 in Table 3 show that removing RGRP and PPRP leads to a substantial decline in hard match F_1 from 78.8%

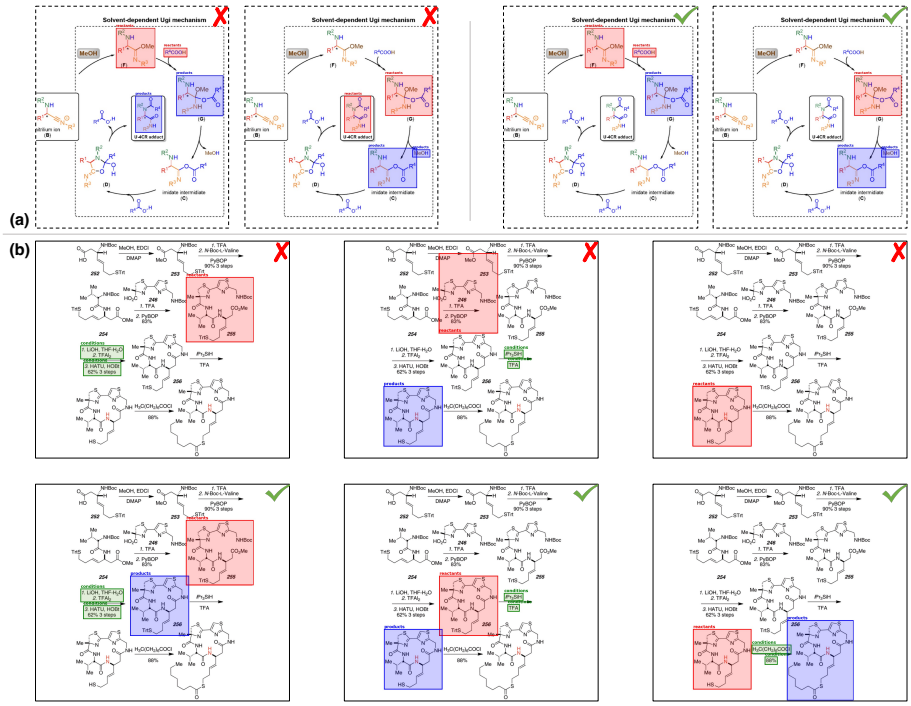


Fig. 6: (a) *Left.* Results of RobustRDP (w/o DPO, Row 2 in Table 3). *Right.* Results of RobustRDP (Row 1 in Table 3). (b) *Up.* Results of RobustRDP (w/o RGRP and PPRP, Row 6 in Table 3). *Bottom.* Results of RobustRDP (Row 3 in Table 3).

to 74.6% and in soft match F_1 from 88.5% to 85.9%, demonstrating the effectiveness of the proposed auxiliary SFT tasks in mitigating output instability. When removing RGRP and PPRP, we oversampled the VRP training data to ensure that all four experiments used the same number of training samples for fairness. Without RGRP and PPRP, RobustRDP faces shortcut learning and oracle prefix issues during training, resulting in deficient local parsing accuracy at inference, and thus caused minor errors in sequence can easily propagate, leading to cascading failures. Consequently, the evaluation metrics exhibit a substantial decline. Comparative visual examples are presented in Fig. 6 (b).

Effect of RobustRDP-train Data. Row 7 in Table 3 evaluates RobustRDP trained exclusively on the existing RxnScribe-train set, excluding our newly annotated images. The substantial decline in hard match F_1 (-4.2%) and soft match F_1 (-5.1%) highlights the importance of scaling high-quality real annotations, which introduce broader layout diversity and stylistic variations, enabling RobustRDP to better generalize to authentic reaction diagrams.

5 Conclusion

In this paper, we propose RobustRDP, a robust reaction diagram parser that achieves SOTA performance by addressing the critical bottlenecks of data scarcity and output instability. Specifically, we introduce a synthetic-to-real data scaling schema that significantly expands the training resources, and a three-stage robustness-oriented training framework that enhance robustness through mitigating sequential dependencies. In future work, we will further explore integrating rich chemical knowledge, such as structural transformation patterns between molecules, to enhance the model's ability to infer object chemical relationships in diagrams with ambiguous layouts.

Acknowledgements

This research was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (XDA0490000) and the National Natural Science Foundation of China (62276245).

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
2. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Zhu, D., et al.: Llava-onevision-1.5: Fully open framework for democratized multimodal training. arXiv preprint arXiv:2509.23661 (2025)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023)
5. Chen, T., Saxena, S., Li, L., Fleet, D.J., Hinton, G.: Pix2seq: A language modeling framework for object detection. arXiv preprint arXiv:2109.10852 (2021)
6. Chen, Y., Leung, C.T., Sun, J., Huang, Y., Li, L., Chen, H., Gao, H.: Towards large-scale chemical reaction image parsing via a multimodal large language model. Chemical Science **16**(45), 21464–21474 (2025)
7. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
8. Cheng, K., Song, W., Fan, J., Ma, Z., Sun, Q., Xu, F., Yan, C., Chen, N., Zhang, J., Chen, J.: Caparena: Benchmarking and analyzing detailed image captioning in the llm era. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 14077–14094 (2025)

9. Fang, X., Wang, J., Cai, X., Chen, S., Yang, S., Tao, H., Wang, N., Yao, L., Zhang, L., Ke, G.: Molparser: End-to-end visual recognition of molecule structures in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24528–24538 (2025)
10. Fletcher, D.A., McMeeking, R.F., Parkin, D.: The united kingdom chemical database service. Journal of chemical information and computer sciences **36**(4), 746–749 (1996)
11. Guan, S., Lin, M., Xu, C., Liu, X., Zhao, J., Fan, J., Xu, Q., Greene, D.: Prep-ocr: a complete pipeline for document image restoration and enhanced ocr accuracy. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 15413–15425 (2025)
12. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1. 5-vl technical report. arXiv preprint arXiv:2505.07062 (2025)
13. Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., et al.: Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv e-prints pp. arXiv–2507 (2025)
14. Irwin, J.J., Tang, K.G., Young, J., Dandarchuluun, C., Wong, B.R., Khurelbaatar, M., Moroz, Y.S., Mayfield, J., Sayle, R.A.: Zinc20—a free ultralarge-scale chemical database for ligand discovery. Journal of chemical information and modeling **60**(12), 6065–6073 (2020)
15. Ishmam, F., Tashdeed, I., Saadat, T.A., Ashmafee, H., Kamal, A.R.M., Hossain, A.: Visual robustness benchmark for visual question answering (vqa). In: Proceedings of the Winter Conference on Applications of Computer Vision. pp. 6623–6633 (2025)
16. Kanerva, J., Ledins, C., Käpyaho, S., Ginter, F.: Ocr error post-correction with llms in historical documents: no free lunches. In: Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025). pp. 38–47 (2025)
17. Kang, S., Kim, J., Kim, J., Hwang, S.J.: Your large vision-language model only needs a few attention heads for visual grounding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9339–9350 (2025)
18. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., et al.: Pubchem 2023 update. Nucleic acids research **51**(D1), D1373–D1380 (2023)
19. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
20. Lin, Y., Zhang, R., Wang, D., Cernak, T.: Computer-aided key step generation in alkaloid total synthesis. Science **379**(6631), 453–457 (2023)
21. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
22. Liu, P., Ren, Y., Tao, J., Ren, Z.: Git-mol: A multi-modal large language model for molecular science with graph, image, and text. Computers in biology and medicine **171**, 108073 (2024)
23. Ma, C., Jiang, Y., Wu, J., Yuan, Z., Qi, X.: Groma: Localized visual tokenization for grounding multimodal large language models. In: European Conference on Computer Vision. pp. 417–435. Springer (2024)
24. Martori, M., Probst, D.: Machine Learning approach for chemical reactions digitalisation. (6), `https://github.com/markmartorilopez/,version={0.0},year={2022}`

25. Memon, J., Sami, M., Khan, R.A., Uddin, M.: Handwritten optical character recognition (ocr): A comprehensive systematic literature review (slr). IEEE access **8**, 142642–142668 (2020)
26. Ou-Yang, S.s., Lu, J.y., Kong, X.q., Liang, Z.j., Luo, C., Jiang, H.: Computational drug discovery. Acta Pharmacologica Sinica **33**(9), 1131–1140 (2012)
27. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
28. Plehiers, P.P., Coley, C.W., Gao, H., Vermeire, F.H., Dobbelaere, M.R., Stevens, C.V., Van Geem, K.M., Green, W.H.: Artificial intelligence for computer-aided synthesis in flow: analysis and selection of reaction components. Frontiers in Chemical Engineering **2**, 5 (2020)
29. Qian, Y., Guo, J., Tu, Z., Coley, C.W., Barzilay, R.: Rxnscribe: a sequence generation model for reaction diagram parsing. Journal of chemical information and modeling **63**(13), 4030–4041 (2023)
30. Qian, Y., Guo, J., Tu, Z., Li, Z., Coley, C.W., Barzilay, R.: Molscribe: robust molecular structure recognition with image-to-graph generation. Journal of chemical information and modeling **63**(7), 1925–1934 (2023)
31. Rajan, K., Brinkhaus, H.O., Agea, M.I., Zielesny, A., Steinbeck, C.: Decimer. ai: an open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. Nature communications **14**(1), 5045 (2023)
32. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
33. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
34. Sapkota, R., Cheppally, R.H., Sharda, A., Karkee, M.: Yolo26: key architectural enhancements and performance benchmarking for real-time object detection. arXiv preprint arXiv:2509.25164 (2025)
35. Song, J., Wang, C., Jiang, B., Wang, Y., Zheng, H., Wei, X., Liu, C., Nie, R., Gao, J., Sun, J., et al.: Rxncaption: Reformulating reaction diagram parsing as visual prompt guided captioning. arXiv preprint arXiv:2511.02384 (2025)
36. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., et al.: Kimi-vl technical report. arXiv preprint arXiv:2504.07491 (2025)
37. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., XiXuan, S., et al.: Cogvlm: Visual expert for pretrained language models. Advances in Neural Information Processing Systems **37**, 121475–121499 (2024)
38. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
39. Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al.: Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. Advances in Neural Information Processing Systems **36**, 61501–61513 (2023)
40. Wilary, D.M., Cole, J.M.: Reactiondataextractor: a tool for automated extraction of information from chemical reaction schemes. Journal of chemical information and modeling **61**(10), 4962–4974 (2021)

41. Wilary, D.M., Cole, J.M.: 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)
42. Wu, J., Zhong, M., Xing, S., Lai, Z., Liu, Z., Chen, Z., Wang, W., Zhu, X., Lu, L., Lu, T., et al.: Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. Advances in Neural Information Processing Systems **37**, 69925–69975 (2024)
43. Xu, Z., Li, J., Yang, Z., Li, S., Li, H.: Swinocsr: end-to-end optical chemical structure recognition using a swin transformer. Journal of cheminformatics **14**(1), 41 (2022)
44. Zhang, H., You, H., Dufter, P., Zhang, B., Chen, C., Chen, H.Y., Fu, T.J., Wang, W.Y., Chang, S.F., Gan, Z., et al.: Ferret-v2: An improved baseline for referring and grounding with large language models. arXiv preprint arXiv:2404.07973 (2024)
45. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)
46. Zumla, A., Chan, J.F., Azhar, E.I., Hui, D.S., Yuen, K.Y.: Coronaviruses—drug discovery and therapeutic options. Nature reviews Drug discovery **15**(5), 327–347 (2016)