

ReactBench: A Benchmark for Topological Reasoning in MLLMs on Chemical Reaction Diagrams

Qiang Xu^{1,2*}, Shengyuan Bai^{1*}, Yu Wang³, He Cao¹, Leqing Chen¹,
Yuanyuan Liu¹, Bin Feng¹, Zijing Liu¹, Yu Li^{1†}

¹International Digital Economy Academy (IDEA)

²Arizona State University

³University of Wisconsin–Madison

xuqiang@asu.edu baishengyuan@idea.edu.cn

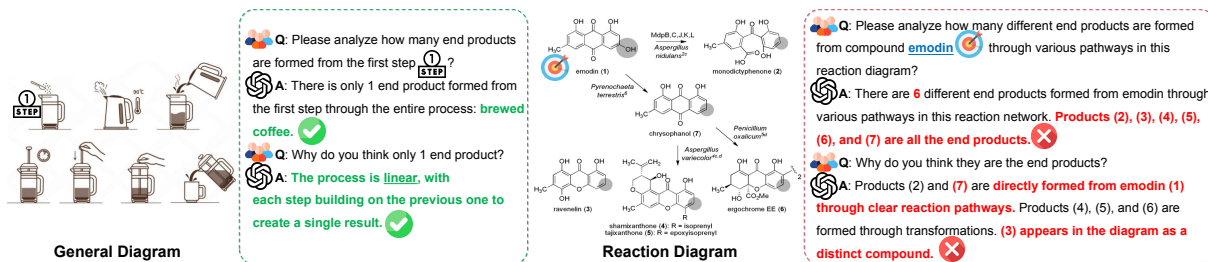


Figure 1: **Comparing General Diagram and Chemical Reaction Diagram on Element Localization Tasks.** GPT-4o performs well on general diagrams for identification tasks. However, for chemical reaction diagram, the structural complexity of molecular diagrams and reaction pathways leads to errors.

Abstract

Multimodal Large Language Models (MLLMs) excel at recognizing individual visual elements and reasoning over simple linear diagrams. However, when faced with complex topological structures involving branching paths, converging flows, and cyclic dependencies, their reasoning capabilities degrade sharply, even on tasks as basic as counting endpoints. Existing benchmarks fail to probe this gap, focusing on semantic comprehension rather than structural reasoning. We introduce **ReactBench**, a benchmark that reveals fundamental limitations in structural reasoning through chemical reaction diagrams. These real-world scientific diagrams offer an ideal testbed because they naturally span diverse structures from linear chains to cyclic graphs, while requiring both precise local recognition and coherent global reasoning. Our benchmark comprises 1,618 expert-annotated QA pairs across four hierarchical task dimensions. Extensive evaluation across 24 MLLMs reveals a significant performance gap exceeding 30% between anchor-based tasks and holistic structural reasoning tasks. Controlled ablations confirm this bottleneck lies in reasoning, not perception. These findings expose a fundamental deficit in structural understanding and establish directions for advancing visual reasoning. Our dataset is publicly available at: <https://huggingface.co/datasets/IDEA-AI4S/ReactBench>.

1 Introduction

Recent advances in MLLMs have demonstrated remarkable capabilities in visual-linguistic reasoning across diverse domains (OpenAI, 2024; Anthropic, 2024; Gemini Team et al., 2024; Bai et al., 2025c). These models excel at recognizing individual visual elements and reasoning over simple linear diagrams. However, real-world diagrams often encode information through complex topological structures involving branching paths, converging flows, and cyclic dependencies. The capacity of MLLMs to reason about such structural connections remains largely unexplored, as existing benchmarks (Mathew et al., 2021; Singh et al., 2019; Masry et al., 2022; Lu et al., 2022) primarily evaluate semantic comprehension rather than structural reasoning capabilities.

We illustrate this gap through a controlled comparison in Fig. 1. Both tasks require the same topological operation of identifying terminal nodes in a diagram. When presented with a simple linear topology such as a sequential coffee brewing process, GPT-4o correctly identifies the single endpoint. However, when the same task is posed on a branching chemical network, the model fails systematically, misidentifying six endpoints instead of the actual three. Notably, neither task requires domain expertise. The failure is attributable to an inability to reason about complex branching structures rather than a lack of chemical knowledge.

* Equal contribution. † Corresponding author.

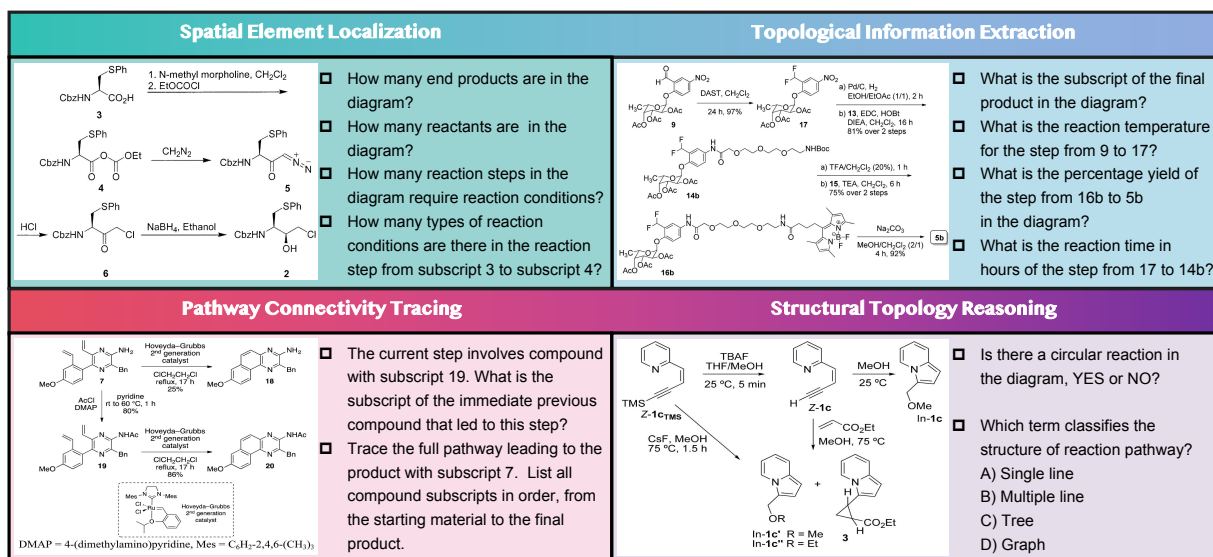


Figure 2: **Overview of ReactBench.** ReactBench systematically evaluates topological reasoning on chemical reaction diagrams across four complexity-stratified dimensions.

These observations reveal a fundamental limitation in current MLLMs. The ability to perceive local elements does not guarantee global structural understanding. As topological complexity increases, models systematically fail to integrate visual elements into coherent global structures.

Diagnosing this limitation requires a testbed that combines diverse topological structures with demands for both local recognition and global reasoning. Chemical reaction diagrams provide such a testbed: they span structures from linear chains to cyclic graphs, while requiring accurate perception of molecules, conditions, and reaction connectivity.

We introduce **ReactBench**, a benchmark designed to diagnose structural reasoning limitations in MLLMs through chemical reaction diagrams. ReactBench comprises 1,618 question-answer pairs carefully curated from real-world chemistry literature and patents, organized into four hierarchical task dimensions: spatial element localization, topological information extraction, pathway connectivity tracing, and structural topology reasoning. This design enables precise identification of where model capabilities degrade along the perception-to-reasoning spectrum. Our main contributions are:

- We present ReactBench, the first benchmark that systematically diagnoses structural reasoning limitations in MLLMs through hierarchical task decomposition, comprising 1,618 expertly annotated question-answer pairs across four skill dimensions.
- We identify a persistent gap between local

diagram understanding and global structural reasoning in current MLLMs. Although models perform well on anchor-based perception and local pathway tracing, they struggle to integrate local visual cues into coherent topological representations, showing that structural complexity remains a key challenge.

- Through controlled ablations across multiple model families, we show that the main bottleneck lies in reasoning rather than perception, and identify recurring structural failure modes that inform future visual reasoning systems.

2 Related Work

2.1 Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) extend text-only LLMs by integrating visual comprehension capabilities. Early works established fundamental vision-language alignment (e.g., CLIP (Radford et al., 2021)) and cross-modal fusion (e.g., BLIP-2 (Li et al., 2023)). Through large-scale pre-training and instruction tuning on diverse multimodal datasets, the field has rapidly advanced, achieving remarkable proficiency in general visual reasoning and Optical Character Recognition (OCR) understanding. Today, the landscape is driven by highly capable open-source models (e.g., Qwen3-VL (Bai et al., 2025b), Google Gemma 4 (Google DeepMind, 2026b), and Kimi-K2.6 (Moonshot AI, 2026)), alongside powerful proprietary models (e.g., GPT-5.5 (OpenAI, 2026), Claude-Sonnet-4.6 (Anthropic, 2026), and Gemini-

Category	Number
Question Dimensions	
Spatial Element Localization	835
Topological Information Extraction	414
Pathway Connectivity Tracing	167
Structural Topology Reasoning	202
Question types	
Numerical questions	1094
Multiple-choice questions	202
Free-form questions	322
Total	1618

Table 1: Key statistics of ChemReaction.

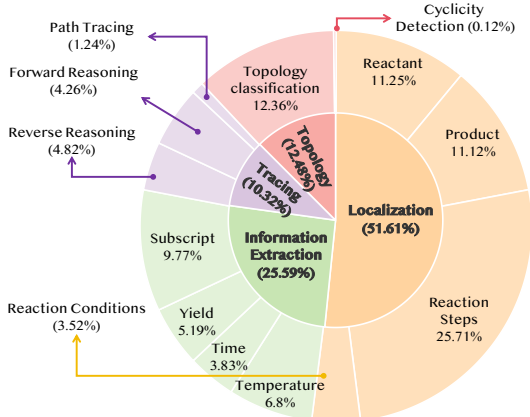


Figure 3: Composition of ChemReaction.

3.1-Pro (Google DeepMind, 2026a)). However, while these state-of-the-art MLLMs excel at recognizing individual visual elements and extracting localized information, their capacity to perform holistic structural reasoning (Dalke, 2019; Xu et al., 2025; Cao et al., 2026) over complex topological diagrams remains largely unexplored.

2.2 Visual Question Answering Benchmarks

Visual Question Answering (VQA) benchmarks have been instrumental in driving the development and evaluation of Multimodal Large Language Models. Existing datasets like DocVQA (Mathew et al., 2021) and ChartQA (Masry et al., 2022) primarily evaluate visual information extraction capabilities across relatively simple, standardized layouts (Li et al., 2025c, 2026b). On the other hand, science-oriented benchmarks such as ScienceQA (Lu et al., 2022) and AI2D (Kembhavi et al., 2016) test domain-specific scientific knowledge and basic diagrammatic understanding. Chemical-domain benchmarks such as MoleculeQA (Lu et al., 2024) and ChemCoT-Bench (Li et al., 2026a) further evaluate molecular comprehension and chemical reasoning, respec-

tively. However, these benchmarks do not systematically evaluate structural reasoning over complex topological diagrams. In these scenarios, visual inputs typically act as static semantic references rather than intricate relational networks (Bai et al., 2025a). Consequently, while current VQA benchmarks heavily focus on general semantic comprehension and basic grounding, they lack a systematic evaluation of complex topological reasoning. Our work addresses this critical gap by specifically isolating and evaluating multi-hop, graph-theoretic reasoning within dense structural topologies.

2.3 Chemical Reaction Diagram Understanding

Existing approaches to chemical reaction diagram understanding employ multi-stage pipelines that first perform Optical Chemical Structure Recognition (OCSR) to detect individual molecular structures, then convert detected elements into text-based representations such as the Simplified Molecular Input Line Entry System (SMILES) (Rajan et al., 2020; Qian et al., 2023b,a; Clevert et al., 2021). Representative methods include RxnScribe (Qian et al., 2023a) and Reaction-DataExtractor 2.0 (Wilary and Cole, 2023). Such pipeline-based frameworks are fundamentally incompatible with end-to-end MLLMs, which directly process visual inputs without intermediate symbolic conversion. Furthermore, text-based representations discard spatial positioning and pathway directionality, precluding differentiation between recognition errors and reasoning deficits (Filippov and Nicklaus, 2009; Oldenhof et al., 2020; Rajan et al., 2023). In contrast, ReactBench evaluates MLLMs’ topological reasoning capabilities through visual question-answering tasks, enabling direct assessment of structural understanding without pipeline-induced confounds.

3 The ReactBench Benchmark

We introduce ReactBench, a specialized benchmark designed to evaluate MLLMs’ ability to comprehend and reason about chemical reaction diagrams. ReactBench consists of a high-quality dataset ChemReaction with 1,618 rigorously curated QA pairs derived from real-world chemical reaction diagrams, and a comprehensive evaluation framework assessing four key dimensions of reaction diagram understanding: (1) **spatial element localization**, (2) **topological information extrac-**

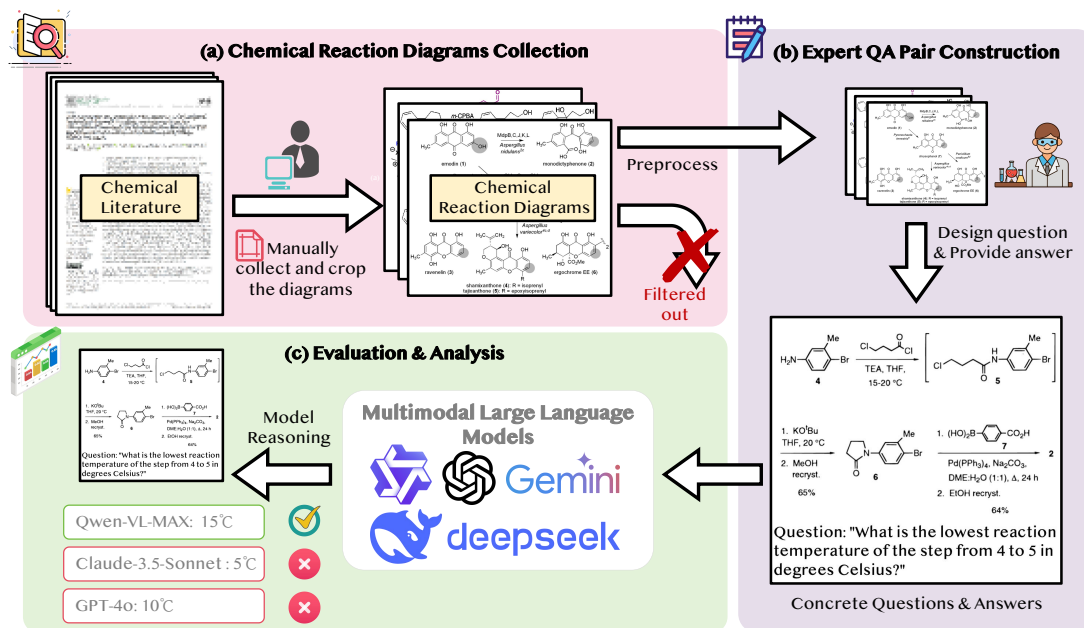


Figure 4: **ReactBench pipeline.** We collect and curate chemical reaction diagrams from literature and patents, construct expert-validated QA pairs, and evaluate MLLMs across predefined topological reasoning tasks.

tion, (3) pathway connectivity tracing, (4) structural topology reasoning. Key statistics of ChemReaction are summarized in Tab. 1, and its composition is illustrated in Fig. 3. The complete benchmark construction pipeline is depicted in Fig. 4. Below, we detail the dataset collection and annotation methodology for ChemReaction, and the evaluation scope of ReactBench.

3.1 Dataset Details

Data Collection and Curation. ReactBench is constructed from over 1,300 unique real-world chemical reaction diagrams curated from chemistry literature and patent databases, ensuring both academic rigor and practical relevance. The academic sources comprise diagrams extracted from articles published between 1996 and 2016 in major Chemical Society journals, including *The Journal of Organic Chemistry*, *Journal of the American Chemical Society*, and *Organic Letters*. To ensure diversity, each diagram is sourced from a distinct article, covering varied layouts, notation styles, and information densities. We further filter out diagrams with incomplete mechanisms, ambiguous visual structures, or duplicate content. The final dataset contains 1,618 question-answer pairs.

Taxonomy of Reaction Topologies. To capture diverse visual complexity and reasoning challenges, we stratify the curated diagrams into four topology categories: (a) *Single-line*: linear pathways with unidirectional flows; (b) *Multiple-lines*: parallel or branched mechanisms with competing intermedi-

ates; (c) *Tree*: hierarchical branching with divergent synthesis routes; (d) *Graph*: cyclic networks requiring non-linear reasoning.

3.2 Question Design

To evaluate MLLMs’ chemical reaction diagram understanding, we design questions across four dimensions, as shown in Fig. 2.

Spatial Element Localization. This task evaluates fine-grained spatial reasoning within structured diagrams. Models must identify node roles, such as sources, sinks, and intermediates, based on topological position and visual context.

Topological Information Extraction. This dimension assesses structured information extraction under topological context. Models must extract numerical and textual attributes, such as reaction conditions, yields, and subscripts, while preserving their correspondence to graph elements.

Pathway Connectivity Tracing. This dimension probes connectivity and path traversal in reaction networks. Models must trace pathways, predict sequential nodes, and reconstruct complete paths from topological connections alone, testing graph-theoretic reasoning rather than chemical prediction.

Structural Topology Reasoning. This dimension evaluates holistic understanding of reaction topology. Models must classify global structural patterns, such as linear, branched, cyclic, tree-like, or graph topologies, and detect properties such as cycles or terminal nodes.

Table 2: Detailed evaluation results on ReactBench across different models, showing the answer accuracy (Acc.) of each model in each task. **Average** denotes the arithmetic mean of a model’s scores over all four tasks. The **best** and **second-best** per model category is highlighted with bold and underlined respectively.

Models	# Params	ReactBench				Average
		Localization	Extraction	Tracing	Reasoning	
Proprietary Models						
GPT-5.5	-	55.21	<u>88.41</u>	83.23	51.49	<u>69.59</u>
GPT-4o	-	44.07	79.71	82.04	49.50	63.83
Claude-Sonnet-4.6	-	<u>55.69</u>	75.36	<u>90.42</u>	46.53	67.00
Claude-Sonnet-3.5	-	50.78	81.88	92.22	49.50	68.60
Gemini-3.1-Pro	-	59.16	85.27	86.23	64.85	73.88
Gemini-1.5-Pro	-	45.75	75.60	71.86	<u>55.45</u>	62.17
Qwen-VL-MAX	-	49.22	90.10	88.62	32.18	65.03
Open Source Models						
InternVL2.5-MPO	8B	38.20	66.43	59.28	37.62	50.38
InternVL2.5	8B	27.43	65.22	45.51	26.73	41.22
InternVL2.5	26B	33.89	66.18	65.87	19.80	46.44
InternVL2.5	78B	41.08	79.47	67.66	26.73	53.74
Qwen3-VL-Instruct	8B	52.34	85.02	82.04	32.67	63.02
Qwen3-VL-Instruct	32B	<u>60.24</u>	86.96	91.02	57.43	73.91
Qwen3-VL-Thinking	8B	49.22	85.99	76.65	29.11	60.24
Qwen3-VL-Thinking	32B	63.59	88.89	<u>88.02</u>	43.07	<u>70.89</u>
Qwen2.5-VL	3B	24.43	<u>89.61</u>	49.70	34.65	49.60
Qwen2.5-VL	7B	37.96	85.75	66.47	39.11	57.32
Qwen2.5-VL	72B	46.47	89.86	83.83	<u>54.46</u>	68.66
DeepSeek-VL2	3B	45.39	50.72	16.17	33.66	36.49
DeepSeek-VL2	16B	28.62	78.50	8.38	28.71	36.05
LLaVA-NeXT-Mistral	7B	18.44	30.43	26.95	21.78	24.40
LLaVA-NeXT-Vicuna	13B	13.77	39.13	1.80	28.22	20.73
Phi-3.5-vision-instruct	4.2B	21.80	69.75	29.34	38.12	39.75
MiniCPM-o 2.6	8B	35.09	75.12	51.50	39.60	50.33

3.3 Evaluation Method

To simplify the evaluation process, all questions in ChemReaction are presented in either multiple-choice or open-ended formats with concise, easily verifiable ground truth answers. We utilize prompting and template matching to extract answers. Prompts guide the model in generating responses in both full and short-answer formats. After generations, the short answer is extracted to compare it with the ground truth. Detailed prompts used in our experiments can be found in Appendix D.2. This design eliminates the need for using LLM-as-a-judge (Chen et al., 2024a) and enables fully automated accuracy measurement through exact string matching.

4 Experiment

4.1 Experimental Setups

We comprehensively evaluate various MLLMs across multiple chemical reaction diagram understanding tasks. We test open-source models including Qwen3-VL-Instruct (8B, 32B) and Qwen3-VL-Thinking (8B, 32B) (Bai et al., 2025b), Qwen2.5-VL (3B, 7B, 72B) (Bai et al., 2025c), InternVL2.5 (8B-MPO, 8B, 26B, 78B) (Chen et al., 2024b), LLaVA-NeXT (Mistral-7B, Vicuna-13B) (Liu et al., 2024), MiniCPM-o 2.6 (8B) (Yao et al., 2024), DeepSeek-VL2 (3B, 16B) (Wu et al., 2024) and Phi-3.5-vision-instruct (4.2B) (Abdin et al., 2024). Furthermore, we evaluate proprietary models such as GPT-5.5 (OpenAI, 2026),

GPT-4o (OpenAI, 2024), Claude-Sonnet-4.6 (Anthropic, 2026), Claude-Sonnet-3.5 (Anthropic, 2024), Gemini-3.1-Pro (Google DeepMind, 2026a), Gemini-1.5-Pro (Gemini Team et al., 2024), and Qwen-VL-MAX (Qwen, 2024). Evaluation of these MLLMs ensures fair comparison considering model architecture differences and parameter scales.

4.2 Main Results

In this section, we evaluate both open-source and proprietary MLLMs on ReactBench. Tab. 2 reports model accuracy across four tasks. The results reveal a clear pattern: recent frontier MLLMs substantially improve the overall performance ceiling, yet a clear gap remains between anchor-based perception and holistic topological reasoning.

Recent frontier models improve the performance ceiling, but do not close the topology gap. The strongest models in our evaluation substantially outperform earlier MLLMs on ReactBench. Among all evaluated models, Qwen3-VL-Instruct-32B achieves the highest average accuracy of 73.91%, closely followed by Gemini-3.1-Pro with 73.88%. Qwen3-VL-Thinking-32B and GPT-5.5 also obtain strong average scores of 70.89% and 69.59%, respectively. These results indicate that recent MLLMs have made notable progress in chemical reaction diagram understanding. However, their gains are uneven across task types. The best scores on Extraction and Tracing reach 90.10% and 92.22%, respectively, while the best scores on Localization and Reasoning remain lower at 63.59% and 64.85%. This gap suggests that local visual recognition and pathway following are more mature than global topological abstraction.

Thinking does not consistently improve topological reasoning. A notable finding is that thinking-style models do not uniformly outperform their non-thinking counterparts. For proprietary models, GPT-5.5 improves over GPT-4o in average accuracy, increasing from 63.83% to 69.59%. Gemini-3.1-Pro also substantially improves over Gemini-1.5-Pro, increasing from 62.17% to 73.88%, and achieves the best Reasoning accuracy among all models at 64.85%. However, this trend is not universal. Claude-Sonnet-4.6 obtains a lower average score than Claude-Sonnet-3.5, and also slightly underperforms it on both Tracing and Reasoning.

The comparison between Qwen3-VL-Instruct and Qwen3-VL-Thinking provides a more controlled view of this effect. Although Qwen3-VL-

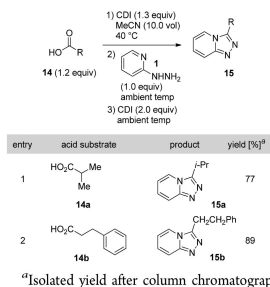
Thinking-32B achieves higher Localization and Extraction scores than Qwen3-VL-Instruct-32B, it performs substantially worse on Reasoning, dropping from 57.43% to 43.07%. A similar pattern appears at the 8B scale, where Qwen3-VL-Thinking-8B underperforms Qwen3-VL-Instruct-8B in average accuracy. These results suggest that explicit thinking-style inference may improve localized inspection or element-level grounding, but does not necessarily lead to better global topology abstraction. In some cases, the model may still rely on local visual evidence without successfully integrating it into a coherent global reaction structure.

Local pathway tracing does not imply global structural reasoning. Another important pattern is the large gap between Tracing and Reasoning within the same model. For example, Claude-Sonnet-4.6 achieves 90.42% on Tracing but only 46.53% on Reasoning. Qwen-VL-MAX reaches 88.62% on Tracing but drops to 32.18% on Reasoning. Similarly, Qwen3-VL-Thinking-32B obtains 88.02% on Tracing but only 43.07% on Reasoning. These results indicate that following local connectivity between adjacent nodes and inferring the global topology of the entire reaction network require different capabilities. A model may correctly trace local reaction paths while still failing to classify the overall reaction topology.

Open-source frontier models are competitive with proprietary models. The updated evaluation also shows that recent open-source models can match or even surpass proprietary systems on ReactBench. Qwen3-VL-Instruct-32B achieves the best overall average accuracy among all evaluated models, slightly outperforming Gemini-3.1-Pro. It also obtains 91.02% on Tracing and 57.43% on Reasoning, outperforming most proprietary models on these dimensions. This suggests that the performance gap between open-source and proprietary MLLMs has become much smaller for chemical reaction diagram understanding, particularly on tasks involving local pathway following and moderately structured visual reasoning.

Extraction is approaching saturation, while topology remains the main bottleneck. Topological information extraction shows high performance across many strong models. Qwen-VL-MAX achieves 90.10%, Qwen2.5-VL-72B achieves 89.86%, Qwen2.5-VL-3B achieves 89.61%, and Qwen3-VL-Thinking-32B achieves 88.89%. This indicates that local anchor recognition, including textual and numerical information extraction, is no

(a)



Q: How many total reaction steps are there in the diagram?

Gemini1.5 pro	LLaVA 1.6 13B	GPT-4o	Qwen-VL max	Claude3.5	GT
3	15	3	3	3	1

GPT-4o Reasoning:

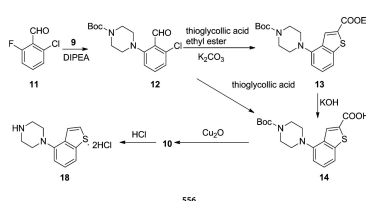
The reaction sequence in the diagram consists of three steps:

- ① Activation of the carboxylic acid (14) with CDI (Carbonyldiimidazole) in MeCN at 40°C – **This step likely forms an acyl imidazolide intermediate.**
- ② Reaction with the hydrazine derivative (1) at ambient temperature – **This step facilitates cyclization to form the fused pyrazolopyrimidine core.**
- ③ Additional CDI treatment at ambient temperature – **This likely completes the transformation into the final product (15).**

Thus, the total number of reaction steps in the diagram is **three**.



(b)



Q: What is the subscript of the final product in the diagram?

Gemini1.5 pro	LLaVA 1.6 13B	GPT-4o	Qwen-VL max	Claude3.5	GT
14	13	2	18	18	18

GPT-4o Reasoning:

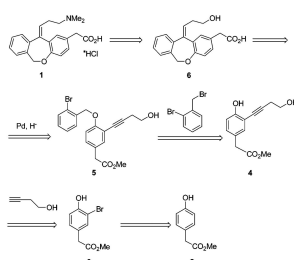
To determine the subscript in the final product, we trace the reaction scheme:

- ① Start with compound ****11****: (a benzaldehyde with fluorine and chlorine).
- ② React with DIPEA, adding a Boc-protected piperazine to form ****12****.
- ③ React with thioglycolic acid ethyl ester and K₂CO₃, yielding ****13****.
- ④ **Hydrolyze with KOH and thioglycolic acid, converting the ester to a carboxylic acid in ****14****.**
- ⑤ Treat with Cu₂O, **forming a thieno[3,2-d]pyrimidin-4-one structure in 18.**
- ⑥ React with HCl**, producing the final hydrochloride salt ****18·2HCl**.

Thus, **the final product has a **subscript of 2****.



(c)



Q: The current step involves compound with subscript 4. What is the subscript of the immediate next compound formed in this reaction?

Gemini1.5 pro	LLaVA 1.6 13B	GPT-4o	Qwen-VL max	Claude3.5	GT
2	6	5	5	5	3

GPT-4o Reasoning:

We observe **that 4 leads to the formation of compound 5, as indicated by the reaction arrow.**

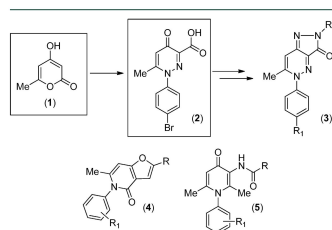
Compound 4 features an alkyne (-C≡C-) functional group and hydroxyl (-OH) groups. **The reaction conditions (Pd, H⁺) suggest a catalytic hydrogenation or coupling process.**

In compound 5, the alkyne functional group remains, and a new structural unit (a bromo-aryl ether moiety) appears.

This transformation suggests 4 → 5 is a key synthetic step.



(d)



The diagram illustrates a reaction. Which term classifies the structure of the reaction pathway? A) Single line B) Multiple line C) Tree D) Graph

Gemini1.5 pro	LLaVA 1.6 13B	GPT-4o	Qwen-VL max	Claude3.5	GT
C	A	B	C	C	A

GPT-4o Reasoning:

The reaction pathway in the diagram is best classified as a **multiple-line structure** because:

- ① Sequential Transformations: The diagram shows a stepwise progression from compound (1) → (2) → (3), indicating a sequential reaction mechanism.
- ② Branching Pathways: **After forming intermediate (3), the reaction further branches into multiple distinct final products (4) and (5). This signifies that the pathway is not a simple linear sequence but involves multiple outcomes.**

Thus, the correct answer is: **B) Multiple line**



Figure 5: Qualitative analysis of recurring failure patterns in ReactBench evaluation. Each case study exemplifies a characteristic reasoning deficit observed across multiple model architectures.

longer the primary bottleneck for frontier MLLMs. By contrast, holistic tasks such as endpoint localization and structural topology reasoning remain

substantially more challenging. Therefore, the key limitation of current MLLMs lies not merely in recognizing individual diagram elements, but in

Table 3: Performance comparison of Qwen2.5-VL using reaction-diagram images and linearized ground-truth JSON. **Loc.**: localization; **Ext.**: extraction; **Trac.**: tracing; **Reas.**: reasoning.

Model	Input	Loc.	Ext.	Trac.	Reas.
Qwen2.5-VL-3B	Image	24.43	89.61	49.70	34.65
	Text-Only	34.37	34.54	19.17	9.90
Qwen2.5-VL-7B	Image	37.96	85.75	66.47	39.11
	Text-Only	39.88	41.06	10.78	8.91

integrating these elements into a coherent global representation of the reaction topology.

5 Diagnosing the Reasoning Gap: Analysis and Ablations

The preceding results show that strong models can exceed 80% on extraction and tracing, yet still struggle with endpoint localization and topology reasoning. This gap raises a key question: *Does the bottleneck stem from visual perception or structural reasoning?* We investigate this through targeted diagnostic experiments.

5.1 Bottleneck: Perception or Reasoning

The preceding results reveal a clear performance gap. We now investigate whether this deficit originates from the visual modality itself, a failure in perception, or a limitation in reasoning.

Step 1: The Necessity of Visual Modality. We first investigate whether complex topological reasoning can be bypassed by providing perfectly structured text. We conduct an ablation study using ground-truth structured JSON (see Appendix D.1), linearized into a text-only prompt to completely replace the image input. This JSON representation encompasses all essential chemical information, including SMILES strings, molecular names, textual tags (e.g., A, B, C), bounding box coordinates (bbox), reaction roles, and reaction conditions. As shown in Tab. 3, while Localization scores slightly improve due to the explicit coordinates in the JSON, performance on Extraction, Tracing, and Reasoning tasks drops precipitously. Notably, for Qwen2.5-VL-7B, the Tracing score falls from 66.47% to a mere 10.78%. This contrast demonstrates that topological reasoning inherently requires the spatial and structural information encoded in visual diagrams, which is lost during the linearization process. However, this necessity of visual input does not imply the bottleneck lies in perception, a point we investigate next.

Step 2a: Does Structured Perception Help?

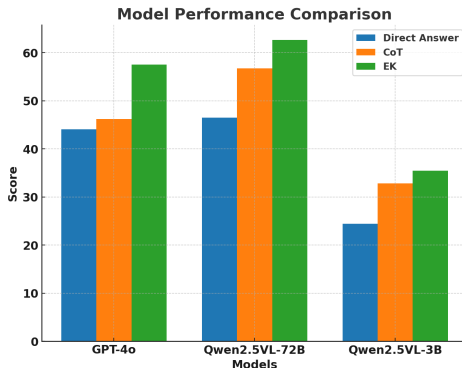


Figure 6: Localization performance comparison of three models with two enhancement techniques: Chain-of-Thought (CoT) and External Knowledge (EK).

To test whether perception is the bottleneck, we prompt models to first decompose the chemical diagram into a structured JSON representation before answering, following Chain-of-Thought methodologies (Wei et al., 2022; Sprague et al., 2024; Wang et al., 2025; Deng et al., 2024). As illustrated in Fig. 6, this structured perception step yields only modest improvements for GPT-4o, Qwen2.5-VL-3B, and Qwen2.5-VL-72B (gains of 2.16%, 8.38%, and 10.30%, respectively). While enforcing systematic visual analysis provides some benefit, the limited gains suggest the problem runs deeper than disorganized perception.

Step 2b: What If Perception Were Perfect? To eliminate the perception variable, we supply models with error-free structural information as External Knowledge (EK) alongside the image (see Appendix D.5). This bypasses the models’ visual processing, testing reasoning capabilities in a best-case scenario. Results in Fig. 6 are striking. While accuracy reaches 57.49% for GPT-4o and 62.63% for Qwen2.5-VL-72B, overall performance remains below 65%. This confirms the primary bottleneck is not visual perception, but a deep-seated limitation in core reasoning over complex topologies.

5.2 How Reasoning Failures Manifest

Having established that the bottleneck lies in reasoning rather than perception, we now examine how these failures manifest across the four task dimensions. The following cases highlight recurring failure patterns shared by different model families.

Localization: Failure in Hierarchical Abstraction. As shown in Fig. 5(a), MLLMs correctly identify local textual markers, such as 1), 2), and 3), but cannot perform the hierarchical abstraction needed to understand these are conditions for a single global transformation. This leads to systematic

Table 4: Impact of molecular placeholder masking on model performance.

Model	Task	Original images	Masked images	Δ
Qwen2.5-VL 3B	Localization	24.43	19.76	-4.67
	Extraction	89.61	78.50	-11.11
	Tracing	49.70	42.51	-7.19
	Reasoning	34.65	31.19	-3.46
Qwen2.5-VL 7B	Localization	37.96	35.81	-2.15
	Extraction	85.75	77.78	-7.97
	Tracing	66.47	56.89	-9.58
	Reasoning	39.11	47.52	+8.41
Qwen2.5-VL 72B	Localization	46.47	45.15	-1.32
	Extraction	89.86	80.68	-9.18
	Tracing	83.83	68.26	-15.57
	Reasoning	54.46	55.94	+1.48

over-segmentation when counting reaction steps, revealing an inability to move from local feature recognition to global process understanding.

Extraction: Loss of Structural Consistency.

Tracking entities across a diagram requires maintaining focus throughout multiple visual elements. In Fig. 5(b), GPT-4o correctly follows the main molecular transformation but is derailed by a local, irrelevant visual feature (*2HCl*) when identifying the final product subscript. This demonstrates a breakdown in structural consistency, where the model fails to filter distractors during extraction.

Tracing: Failure in Topological Path Tracing.

MLLMs often disregard explicit topological rules in favor of learned heuristics. In Fig. 5(c), models ignore directed arrows ($5 \rightarrow 4 \rightarrow 3$), predicting the next compound via superficial numerical sequences. This reveals a critical deficit in graph-theoretic reasoning and the inability to process connective logic.

Reasoning: Inability to Classify Global Topology.

The ultimate test of structural reasoning is classifying a diagram’s overall topology. In Fig. 5(d), models correctly perceive local features like a linear sequence and a branching point, but fail to integrate these into the correct global classification of “single line” structure. This directly confirms our central finding: local perception is intact, but global structural reasoning fails.

5.3 The Role of Molecular Structures in Topological Reasoning

Having shown that reasoning is the main bottleneck, we further examine whether molecular structures serve as useful visual anchors or distractors for topology reasoning (Li et al., 2025b). Following the masking procedure in Appendix D.6, we replace molecular structures with black borders

while preserving the original topological connectivity.

As shown in Tab. 4, masking molecular structures consistently reduces performance on Localization, Extraction, and Tracing across all model sizes. This suggests that models use molecular structures as visual anchors for organizing reaction connectivity and flow. In contrast, Reasoning shows a different pattern: the 3B model slightly declines (-3.46%), whereas the 7B and 72B models improve by 8.41% and 1.48%. This suggests that molecular details help local tasks but may distract larger models from abstract topology reasoning.

Overall, molecular structures play a dual role: they support local perception and pathway understanding, but can impede the hierarchical abstraction required for global structural reasoning.

6 Conclusion

We introduce ReactBench, a benchmark to systematically investigate MLLMs’ comprehension capabilities in chemical reaction diagrams. We evaluate MLLMs using 1,618 QA pairs based on real-world reaction images, covering tasks from basic recognition to advanced reasoning. Our findings reveal persistent limitations in MLLMs’ ability to perform topological reasoning over chemical reaction diagrams, particularly when tasks require integrating local visual cues into a holistic understanding of reaction structure. This highlights the need for improved multimodal reasoning in scientific domains and provides a critical evaluation framework.

Limitations

Our benchmark focuses on chemical reaction diagrams rather than abstract graphs or other diagram types. This choice is deliberate. Chemical reaction diagrams provide natural topological diversity from linear chains to cyclic networks while demanding both local perception and global reasoning. Abstract synthetic graphs lack the visual complexity that makes findings practically meaningful, and other diagram types such as flowcharts typically exhibit limited topological variety. Extending this framework to other scientific domains remains valuable future work. Furthermore, although the scale of 1,618 instances is modest compared to massive, synthetically generated datasets, the exceptionally high density of multi-hop reasoning steps and the rigorous, multi-round expert annotation protocol ensure a high signal-to-noise ratio, effectively prioritizing analytical quality over sheer quantity.

References

- Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- Anthropic. 2024. [Claude 3.5 sonnet](#).
- Anthropic. 2026. Introducing claude sonnet 4.6. <https://www.anthropic.com/news/claude-sonnet-4-6>. Accessed: 2026-05-23.
- Shengyuan Bai, Qibin Li, Zhe Wang, Nai Zhou, and Nianmin Yao. 2025a. Enhancing nlu in large language models using adversarial noisy instruction tuning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 23451–23459.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025b. Qwen3-v1 technical report. *arXiv preprint arXiv:2511.21631*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025c. [Qwen2.5-v1 technical report](#). *Preprint*, arXiv:2502.13923.
- He Cao, Siyu Liu, Fan Zhang, Zijing Liu, Hao Li, Bin Feng, Shengyuan Bai, Leqing Chen, Kai Xie, and Yu Li. 2026. Mozi: Governed autonomy for drug discovery llm agents. *arXiv preprint arXiv:2603.03655*.
- Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.
- Djork-Arné Clevert, Tuan Le, Robin Winter, and Floriane Montanari. 2021. Img2mol—accurate smiles recognition from molecular graphical depictions. *Chemical science*, 12(42):14174–14181.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Andrew Dalke. 2019. The chemfp project. *Journal of cheminformatics*, 11(1):76.
- Yuntian Deng, Yejin Choi, and Stuart Shieber. 2024. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*.
- Igor V Filippov and Marc C Nicklaus. 2009. Optical structure recognition software to recover chemical information: Osra, an open source solution.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. [Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context](#). *Preprint*, arXiv:2403.05530.
- Google DeepMind. 2026a. Gemini 3.1 pro model card. <https://deepmind.google/models/model-cards/gemini-3-1-pro/>. Accessed: 2026-05-23.
- Google DeepMind. 2026b. [Gemma 4](#). Accessed: 2026-05-23.
- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer.
- Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. 2025a. Preference leakage: A contamination problem in llm-as-a-judge. *arXiv preprint arXiv:2502.01534*.
- Hao Li, He Cao, Bin Feng, Yanjun Shao, Xiangru Tang, Zhiyuan Yan, Yonghong Tian, Li Yuan, and Yu Li. 2026a. Beyond chemical QA: Evaluating LLM’s chemical reasoning with modular chemical operations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. [Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). *Preprint*, arXiv:2301.12597.
- Qibin Li, Shengyuan Bai, Nai Zhou, and Nianmin Yao. 2025b. Causally graph-guided counterfactual analysis to biomedical named entity recognition. *Expert Systems with Applications*, page 130485.
- Qibin Li, Shengyuan Bai, Nai Zhou, and Nianmin Yao. 2025c. Enhancing entity and relation extraction with dynamic hard negative augmentation framework. *Engineering Applications of Artificial Intelligence*, 161:112211.

- Qibin Li, Shengyuan Bai, Nai Zhou, and Nianmin Yao. 2026b. Counterfactual-based cognitive alignment in-context learning for relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 23087–23095.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. *Llava-next: Improved reasoning, ocr, and world knowledge*.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.
- Xingyu Lu, He Cao, Zijing Liu, Shengyuan Bai, Leqing Chen, Yuan Yao, Hai-Tao Zheng, and Yu Li. 2024. Moleculeqa: A dataset to evaluate factual accuracy in molecular comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3769–3789.
- Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.
- Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.
- Moonshot AI. 2026. *Kimi k2.6*. Accessed: 2026-05-23.
- Martijn Oldenhof, Adam Arany, Yves Moreau, and Jaak Simm. 2020. Chemgrapher: optical graph recognition of chemical compounds by deep learning. *Journal of chemical information and modeling*, 60(10):4506–4517.
- OpenAI. 2024. *Gpt-4o system card*. Preprint, arXiv:2410.21276.
- OpenAI. 2026. Introducing gpt-5.5. <https://openai.com/index/introducing-gpt-5-5/>.
- Yujie Qian, Jiang Guo, Zhengkai Tu, Connor W Coley, and Regina Barzilay. 2023a. Rxnscribe: a sequence generation model for reaction diagram parsing. *Journal of chemical information and modeling*, 63(13):4030–4041.
- Yujie Qian, Jiang Guo, Zhengkai Tu, Zhening Li, Connor W Coley, and Regina Barzilay. 2023b. Molscribe: robust molecular structure recognition with image-to-graph generation. *Journal of chemical information and modeling*, 63(7):1925–1934.
- Qwen. 2024. *Qwen-vl-max*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. *Learning transferable visual models from natural language supervision*. *CoRR*, abs/2103.00020.
- Kohulan Rajan, Henning Otto Brinkhaus, M Isabel Agea, Achim Zielesny, and Christoph Steinbeck. 2023. Decimer. ai: an open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. *Nature communications*, 14(1):5045.
- Kohulan Rajan, Achim Zielesny, and Christoph Steinbeck. 2020. Decimer: towards deep learning for chemical image recognition. *Journal of Cheminformatics*, 12(1):65.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.
- Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*.
- Yu Wang, Xiaogeng Liu, Yu Li, Muhao Chen, and Chaowei Xiao. 2025. Adashield : Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting. In *Computer Vision – ECCV 2024*, pages 77–94, Cham.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Damian M Wilary and Jacqueline M Cole. 2023. Reactiondataextractor 2.0: a deep learning approach for data extraction from chemical reaction schemes. *Journal of Chemical Information and Modeling*, 63(19):6053–6067.
- Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. *Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding*. Preprint, arXiv:2412.10302.
- Qiang Xu, Shengyuan Bai, Leqing Chen, Zijing Liu, and Yu Li. 2025. Chemlabs on chemo: a multi-agent system for multimodal reasoning on icho 2025. *arXiv preprint arXiv:2511.16205*.
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

A Robustness of the Evaluation Protocol

To ensure our findings are not artifacts of an overly strict evaluation metric, we compare exact-match scoring against an LLM-as-a-judge approach. Strict exact-match risks penalizing semantically correct answers due to surface-form variations (e.g., “2” vs. “two”).

We explicitly mitigate such variations in ReactBench through constrained task design, including single-character multiple-choice questions and strict prompts for numeric outputs. To quantify potential underestimation, we relax these constraints and use Gemini-2.5-Flash (Comanici et al., 2025) as an external semantic judge.

As shown in Tab. 5, semantic evaluation changes the absolute scores substantially (e.g., Tracing improves from 49.70% to 70.06% for Qwen2.5-VL-3B). Despite some task-level variations, the overall model ranking remains consistent, and the significant capability gaps persist. This indicates that while the absolute scale shifts, the overall comparative assessment of model capabilities remains stable.

We therefore retain exact-match as the primary metric for ReactBench. It guarantees absolute reproducibility, eliminates the computational overhead and inherent biases of LLM judges (Li et al., 2025a), and provides a rigorous lower-bound standard for future benchmarking.

B Question-Only Blind Baseline

A natural concern for any diagram-based benchmark is whether models can exploit linguistic patterns or dataset priors to guess answers without actually reading the diagram. To rule out this shortcut, we conduct a *question-only* (blind) ablation in which the chemical reaction diagram is removed entirely and the model receives the textual question alone. We evaluate the strongest open-source family in our study, Qwen3-VL-Instruct (8B and 32B), under identical prompts and answer extraction rules as the main experiments.

Performance collapses without visual input. As shown in Tab. 6, removing the diagram degrades performance across all four dimensions. Extraction, which depends on reading numerical attributes directly off the diagram, drops to below 7% for both model scales, and Reasoning falls below 10%. Notably, these scores fall *below* the random-guessing level of the corresponding answer spaces, indicating that blind models do not guess uniformly but

Table 5: Comparison of evaluation protocols for Qwen2.5-VL. **EM**: Exact Match; **Judge**: LLM-as-a-Judge; **Loc.**: Localization; **Ext.**: Extraction; **Trac.**: Tracing; **Reas.**: Reasoning; **Avg.**: arithmetic mean over the four tasks.

Size	Protocol	Loc.	Ext.	Trac.	Reas.	Avg.
3B	EM	24.43	89.61	49.70	34.65	49.60
	Judge	36.40	83.33	70.06	25.25	53.76
7B	EM	37.96	85.75	66.47	39.11	57.32
	Judge	54.01	87.44	87.42	42.57	67.86

instead commit to systematically incorrect default answers (e.g., a canonical endpoint count). This confirms that ReactBench contains no exploitable language prior: the questions are not answerable from templated phrasing or from the marginal distribution of ground-truth answers.

Visual topology is required, not optional. Questions such as “*What is the subscript of the immediate next compound?*” cannot be resolved by heuristics, since the answer depends on multi-hop traversal over the specific topology of each diagram. Because our 1,300+ diagrams span linear chains, branched mechanisms, trees, and cyclic graphs, no single structural prior generalizes across instances. The residual accuracy observed on Tracing (32.93–36.53%) reflects the fact that a subset of tracing questions admits a small candidate set of adjacent subscripts, rather than any genuine reasoning ability.

Corroboration from thinking-style models. We repeat the same blind experiment on the reasoning-focused variants (Qwen3-VL-Thinking 8B/32B). Because these models perform explicit chain-of-thought before answering, they consistently detect that no diagram is available and decline to produce an answer rather than hallucinating one. This behavior further indicates that ReactBench questions are strictly grounded in the visual input and are not solvable from the text alone.

C Empirical Validation of OCSR Limitations

To illustrate the information loss inherent in current OCSR tools, we analyze the output of RxnScribe (Qian et al., 2023a), a state-of-the-art method. The following example displays a typical prediction for a reaction scheme:

Model	Input	Loc.	Ext.	Trac.	Reas.
Qwen3-VL-Instruct-8B	Image + Question	52.34	85.02	82.04	32.67
	Question Only	15.57	6.76	36.53	9.90
Qwen3-VL-Instruct-32B	Image + Question	60.24	86.96	91.02	57.43
	Question Only	10.30	5.56	32.93	8.91

Table 6: Question-only (blind) baseline on ReactBench. Removing the reaction diagram causes performance to collapse across all four dimensions. Loc.: localization; Ext.: extraction; Trac.: tracing; Reas.: reasoning.

Example output from RxnScribe

```

1 [
2   {
3     "reactants": [
4       {
5         "category": "[Mol]",
6         "smiles": "Cc1ncc(C[n+]2csc(CCO)
          c2C)c(N)n1.C1"
7       }
8     ],
9     "conditions": [
10      { "category": "[Txt]", "text": "
11        28.5% aq" },
12      { "category": "[Mol]", "smiles": "O=
13        S(=O)([O][Na])SCC1CCC01" },
14      { "category": "[Txt]", "text": "NaOH
15        " },
16      { "category": "[Txt]", "text": "rt,
17        10 min" }
18    ],
19    "products": [
20      {
21        "category": "[Mol]",
22        "smiles": "C/C(C(=C(\CCO)SSCC1CCC01
          )N(C=O)Cc1cnc(C)nc1N"
23      }
24    ]
25  }
26 ]

```

While this representation correctly recovers individual molecules and textual labels, it fails to encode the reaction topology targeted by our benchmark:

- **Connectivity:** It does not specify which arrow connects which reactant to which product.
- **Sequential Order:** It does not indicate the temporal order of individual steps in a multi-step scheme.
- **Global Structure:** It does not represent branching pathways or convergent steps within the global network.

Consequently, a model relying solely on OCSR-derived textual representations (SMILES or JSON) lacks the necessary structural context to perform complex topological reasoning.

D Detailed Experiment Setups

In this section, we provide more details about our experiment designs.

D.1 Illustration of Ground-Truth Structured JSON

In this section, we provide a concrete example of the ground-truth structured JSON used in the text-only ablation study discussed in the main paper. Fig. 7 shows the original chemical reaction diagram, while the subsequent JSON snippet demonstrates how its constituent elements are computationally structured.

Example of Ground-Truth JSON

```

1 [
2   {
3     "reactants": [
4       {
5         "category": "[Mol]",
6         "smiles": "C1C(CC1=CC=CC=C1)=O"
7       }
8     ],
9     "conditions": [
10      { "category": "[Txt]", "text": "
11        triethylamine" },
12      { "category": "[Txt]", "text": "2-
13        ethyl-1-hexanol" },
14      { "category": "[Txt]", "text": "
15        methylene chloride" },
16      { "category": "[Txt]", "text": "50C"
17        }
18    ],
19    "products": [
20      {
21        "category": "[Mol]",
22        "smiles": "[O]C(CC1=CC=CC=C1)=O"
23      }
24    ]
25  }
26 ]

```

D.2 Prompt for Answer Extraction

In our main experiment, we directly pose questions to the model and employ prompt engineering and template matching to extract answers. As illustrated in Fig. 9, prompts guide the model in

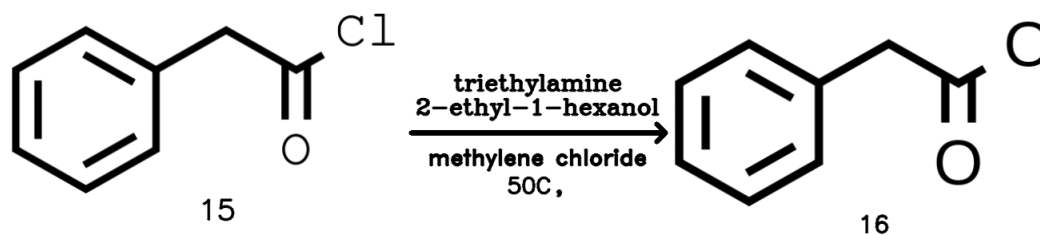


Figure 7: **Example of a chemical reaction diagram used for the text-only ablation study.** This visual representation is translated into the structured JSON format shown below, which preserves the chemical semantics (SMILES, text).

generating responses in both full and short answer formats, where the short answer format adheres to the requirements specified in Tab. 7 for different question types. After generation, the short answer is extracted for comparison with the ground truth, while the full answer is used to analyze the reasoning process of MLLMs.

Table 7: The prompt for different QA types in answer generation.

Answer type	Prompt
multiple choice	Just provide the corresponding choice option, such as 'A', 'B', 'C', or 'D'.
number	Provide ONLY the numerical answer without any units, symbols, or additional text.
text	Just give the subscript without any other text.

D.3 Fine-Grained Subtype Statistics

Fig. 3 reports the proportional composition of the fine-grained subtypes. Because rounded percentages may obscure the actual sample sizes of rare categories, we additionally report absolute counts alongside exact percentages in Tab. 8.

Imbalance among fine-grained subtypes. These statistics make the imbalance across subtypes explicit. Path Tracing contains 20 examples, while Cyclicity Detection contains only 2. We note that the Cyclicity Detection subset is too small to support a statistically reliable standalone evaluation or any broad claim about models’ cycle-detection capabilities; these examples were included to provide coverage of cyclic reaction structures rather than to establish an independent quantitative category. Path Tracing, although larger, is still substantially smaller than Forward and Reverse Reasoning, so we treat its results as exploratory and refrain from using small subtype-level differences to support

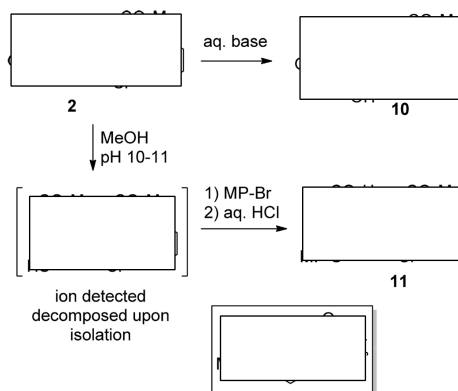


Figure 8: **Illustration of the visual context ablation.** Detailed molecule images are masked and replaced with rectangle placeholders to isolate the topological structure.

model rankings or general conclusions.

All quantitative findings reported in this paper are computed at the level of the four broader task dimensions, which contain 835, 414, 167, and 202 examples respectively. The subtype-level statistics are intended to clarify dataset composition and illustrate the coverage of different operations, not to imply that every subtype supports equally strong statistical conclusions.

D.4 Prompt of Chain-of-Thought

The prompt of the CoT prompting strategy is illustrated in Fig. 10, where the model is first instructed to convert the reaction diagram into a predefined JSON format, followed by the question being posed.

D.5 Prompt of External Knowledge

Here, we present the specific prompt and external knowledge format used in the main text. The prompt itself is shown in Fig. 11, and Fig. 12 gives a concrete example of the accompanying JSON.

Task Dimension	Subtype	Count	%
Spatial Element Localization	Reactant	182	11.25
	Product	180	11.12
	Reaction Steps	416	25.71
	Reaction Conditions	57	3.52
	<i>Total</i>	835	51.61
Topological Information Extraction	Temperature	110	6.80
	Time	62	3.83
	Yield	84	5.19
	Subscript	158	9.77
	<i>Total</i>	414	25.59
Pathway Connectivity Tracing	Forward Reasoning	69	4.26
	Reverse Reasoning	78	4.82
	Path Tracing	20	1.24
	<i>Total</i>	167	10.32
Structural Topology Reasoning	Topology Classification	200	12.36
	Cyclicity Detection	2	0.12
	<i>Total</i>	202	12.48
Overall		1,618	100.00

Table 8: Fine-grained subtype statistics of ChemReaction. Percentages are computed over all 1,618 QA pairs.

D.6 Illustration of Molecular Structure Masking

To provide a clear visual example of the visual context ablation described in the main paper, Fig. 8 illustrates how the chemical diagrams are specifically modified for this experiment. By replacing detailed molecular structures with black rectangular placeholders, we force the models to rely solely on the remaining topological elements (e.g., arrows and text) without the aid of familiar chemical visual cues.

E Dataset Details

E.1 Annotation Protocol and Annotator Qualifications

We established a rigorous annotation pipeline involving domain experts. All annotators are graduate-level chemists with extensive experience in reading and interpreting organic reaction schemes.

- **Guidelines:** Annotators followed detailed guidelines designed to standardize the extraction of structured information. This included specific instructions on (i) identifying and labeling reactants, products, reagents, and reaction conditions; and (ii) formulating diverse question-answer pairs that target recognition (e.g., entity identification) as well as higher-level reasoning (e.g., mechanistic relations,

condition analysis).

- **Training Phase:** Before large-scale annotation, all annotators underwent a training phase using pilot examples. Their initial annotations were reviewed and corrected by a senior expert to ensure alignment with the guidelines.
- **Data Format:** Each reaction diagram is annotated with structured ground truth in JSON format, capturing the semantic roles of all chemical entities within the diagram.

E.2 Quality Control and Inter-Annotator Agreement

To ensure high data quality, we implemented a multi-stage verification process. Each sample was independently reviewed by three distinct annotators. Disagreements regarding diagram interpretation or answer correctness were resolved through group discussion, with a senior expert making the final decision in ambiguous cases. Through this rigorous review process, the empirical consistency of the final annotations exceeded 95%.

F Examples in ReactBench

In this section, we present selected samples from different tasks in our dataset, as illustrated in Fig. 13, 14, 15, and 16. These examples highlight the diversity of our designed questions, which encompass a wide range of problem types.

Direct Answer System Prompt

Multimodal Chemical Reaction QA Task

Inputs:

1. Chemical reaction scheme image
2. Question

Answer the question by following these rules:

1. Provide **ONLY** the numerical answer (e.g., "80") without any units, symbols, or additional text.
2. Provide a detailed explanation of the step-by-step reasoning.

Final Output Format (as a JSON object):

```
{
  "answer": "<numerical_value>",
  "explanation": "<step-by-step reasoning>"
}
```

Figure 9: The direct answer prompt in our ReactBench.

Chain of Thought System Prompt

Chemical Reaction QA Task:

"Step 1 - Structural Parsing: Analyze the reaction image and convert it to this JSON format:"

```
[{
  "reactants": [
    {
      "category": "[Mol]",
      "category_id": 1,
      "bbox": [x_min, y_min, x_max, y_max],
      "compound_id": "..."}
    //Additional reactants...
  ],
  "conditions": [
    {
      "category": "[Txt]",
      "category_id": 2,
      "bbox": [x_min, y_min, x_max, y_max],
      "text": ["reagent", "(equiv)"]
    }
    / Additional conditions...
  ],
  "products": [
    // Product entries...
  ]
  // "Additional reactions..."
}]
```

"Step 2 - Question Answering:"

Using the parsed structure from Step 1, answer the question by following these rules:

1. Provide **ONLY** the numerical answer (e.g., "80") without any units, symbols, or additional text.
2. Provide a detailed explanation of the step-by-step reasoning.

"Final Output Format (as a JSON object): "

```
{
  "answer": "<numerical_value>",
  "explanation": "<step-by-step reasoning>"
}
```

Figure 10: The prompt of Chain-of-Thought in our ReactBench.

External Knowledge System Prompt

Multimodal Chemical Reaction QA Task

Inputs:

1. Chemical reaction scheme image
2. Supplemental JSON data

Task Requirements:

1. Cross-validate information between image and JSON data
2. Answer format requirements:
 - Return ONLY numerical value (e.g. "80")
 - No units, symbols or additional text"
3. Explicitly explain how both modalities contribute to the answer

Output Format (strict JSON):

```
{
  "answer": "<numerical_value>",
  "explanation": "<integration_steps>"
}
```

Figure 11: The prompt of external knowledge in our ReactBench.

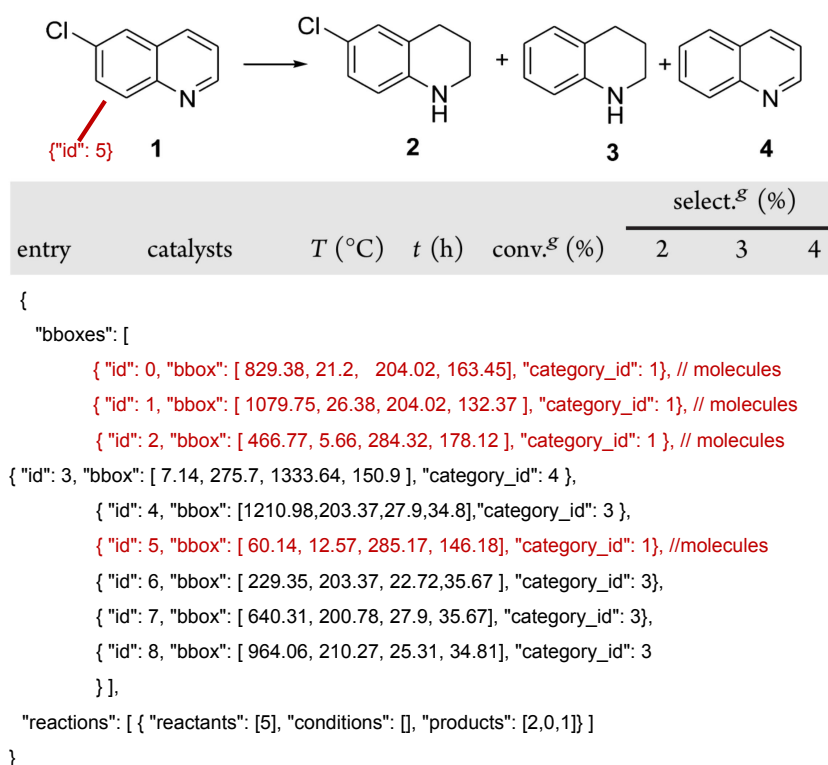


Figure 12: **External Knowledge Example in our ReactBench.** This JSON represents molecular structures and reaction relationships extracted from the image. The bboxes section defines detected elements with their respective IDs, bounding box coordinates, and the category_id, where 1 indicates molecules, 2 denotes text, 3 corresponds to compound identifiers (numerical labels without molecular structures), and 4 represents auxiliary information. The reactions section associates reactants and products by referencing their IDs, illustrating the transformations occurring between molecules. Overall, it systematically organizes visual and chemical data for structured interpretation.

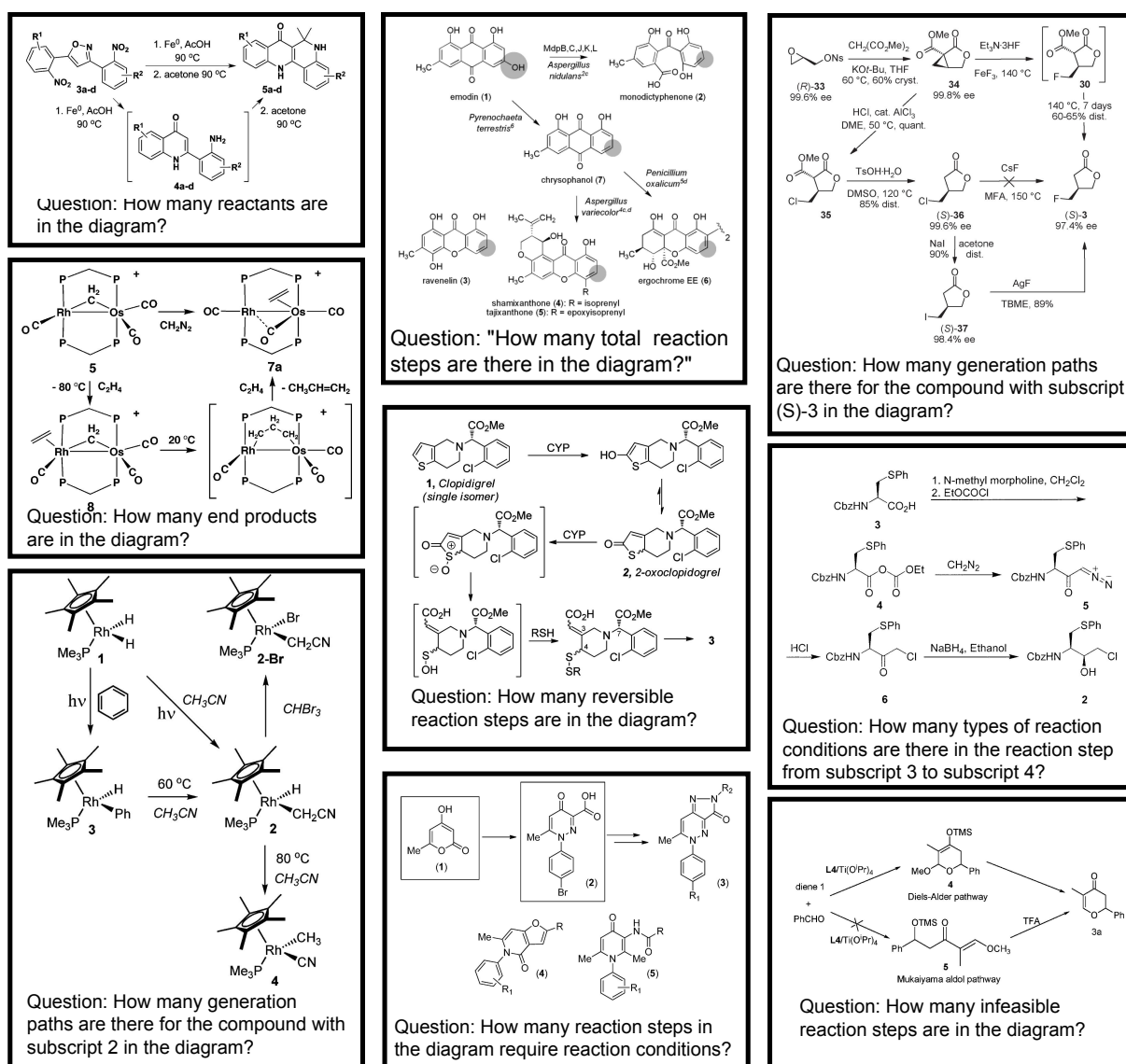


Figure 13: Examples of Element Localization Task in our ReactBench.

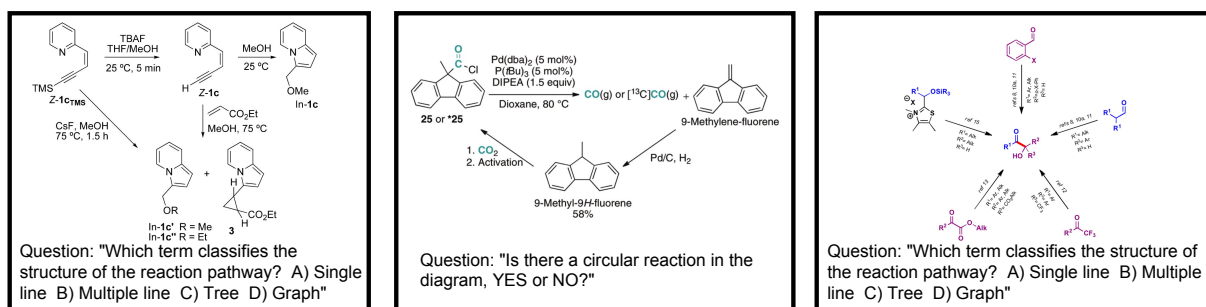


Figure 14: Examples of Topology Reasoning Task in our ReactBench.

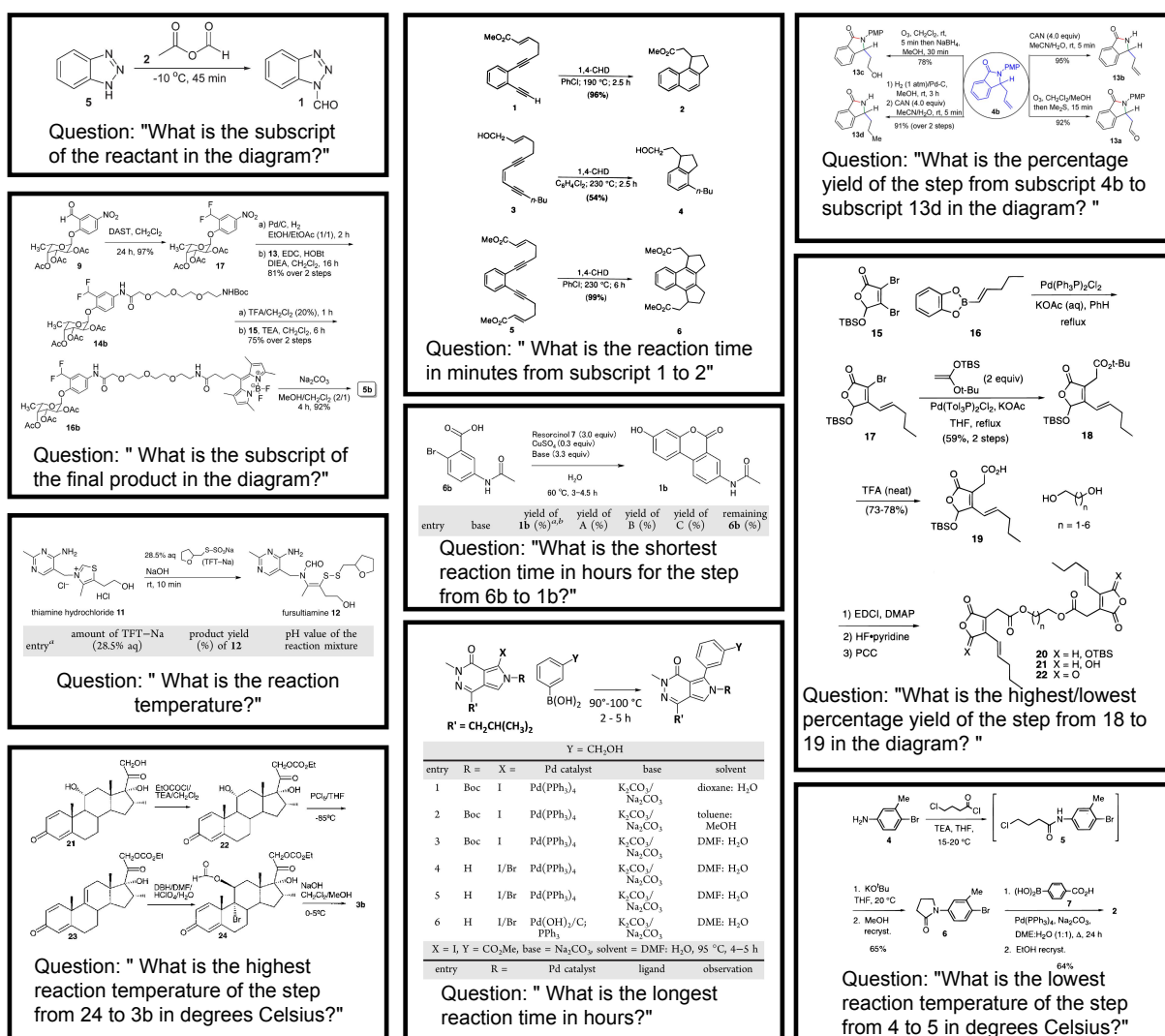


Figure 15: Examples of Information Extraction Task in our ReactBench.

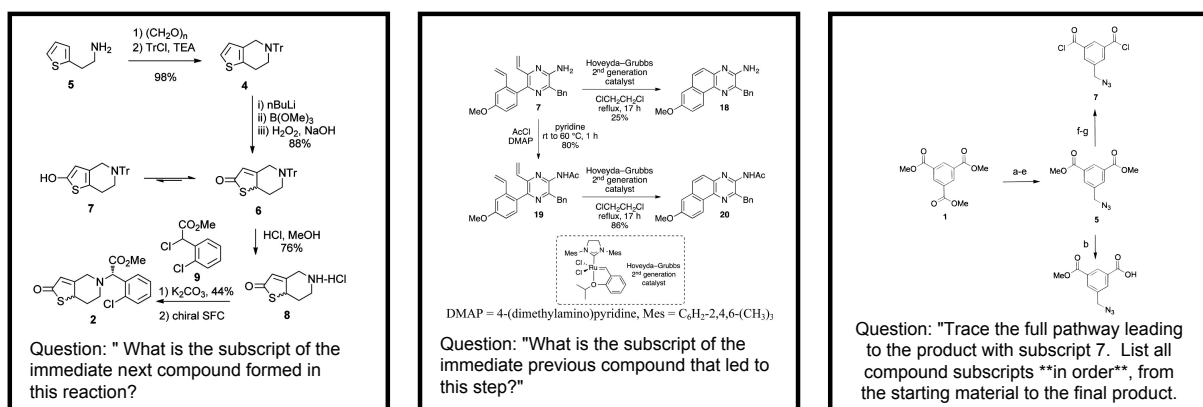


Figure 16: Examples of Connectivity Tracing Task in our ReactBench.