PAPERMOLBENCH: BENCHMARKING AND UNDERSTANDING MOLECULAR STRUCTURE RECOGNITION IN SCIENTIFIC FIGURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific figures in chemistry and biology contain complex molecular structures and figure logic. Existing benchmarks mainly focus on single-molecule images and therefore do not measure whether visual language models (VLMs) can understand real scientific figures for research in these domains. We introduce PaperMolBench, a human-annotated benchmark for extracting target molecular structures and reactions from full scientific figures, comprising 355 test cases. We evaluate representative frameworks, including Tool-Cascaded Method, Agentic Pipeline, and End-to-End Generation, together with several widely used large models. On Molecule Extraction, the best configuration among evaluated methods reaches only 50.0% exact-match (EM) accuracy. To understand the gap, we build a capability diagnostic framework for the simple yet competitive end-to-end generation approach, covering localization, identification, and transcription. Using InternVL3.5-8B, we observe low accuracy even on synthetic molecular images, with further drops on crops from scientific figures, revealing a clear distribution gap. Further analysis indicates that effective molecular structure extraction requires addressing both limited data scale and distribution mismatch, and that linear text representations such as SMILES do not naturally capture molecular topology. These findings highlight key directions for future VLMs in chemistry.

1 INTRODUCTION

Scientific figures in chemistry and biology combine complex molecular structures with reaction arrows, compound identifiers, plots, and explanatory text (Rajan et al., 2023; Morin et al., 2023; Qian et al., 2023a). Recovering these structures supports molecular search and machine-readable analysis of the scientific literature (Filippov & Nicklaus, 2009; Staker et al., 2019).

Existing optical chemical structure recognition (OCSR) benchmarks mainly evaluate images of individual molecules (Clevert et al., 2021; Qian et al., 2023b; Morin et al., 2023; Fang et al., 2025). They therefore do not assess whether visual language models (VLMs) can understand complex molecular structures and figure logic in real scientific figures.

We introduce PaperMolBench, a human-annotated benchmark with 355 cases from 149 scientific figures, covering two tasks: Molecule Extraction and Reaction Extraction. *Molecule Extraction* asks for one query-specified molecular SMILES; *Reaction Extraction* asks for the reactant and product SMILES of a specified reaction. We score canonical isomeric-SMILES exact match (EM) and fingerprint similarity (FPS).

On Molecule Extraction, the Tool-Cascaded Method reaches 18.1% EM, End-to-End Generation reaches 44.1%, and the Agentic Pipeline reaches 46.1%. Three End-to-End Generation calls followed by model selection reach 50.0%. On Reaction Extraction, End-to-End Generation with Gemini 3.8 Flash reaches 33.8% joint EM.

To understand the remaining errors, we diagnose End-to-End Generation on 72 paired molecule cases. We compare full figures, target crops, and RDKit images using Gemini 3.8 Flash and the official InternVL3.5-8B and Qwen3-VL-8B models (Wang et al., 2025; Bai et al., 2025). Gemini's EM does

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

not improve with target crops or RDKit renderings, while both open models improve across these input conditions. Even on RDKit images, all three remain limited, with Gemini reaching 54.2% EM.

These findings motivate our study of molecular data scale and depiction distribution through continued pretraining (CPT) and paired supervised fine-tuning (SFT) of the open models. After SFT, models initialized from the final CPT checkpoints outperform those initialized from the official base models across both SFT variants and all three input conditions. Real-depiction SFT improves crop EM over RDKit-rendered SFT at most checkpoints in both model families, but gives no consistent full-figure advantage. This contrast suggests that real depictions help recognition when the target is provided as a crop, while improving full-figure extraction may require supervision with real scientific figures. Structural probes further show that local structural distinctions remain recoverable from both projected visual features and the language backbone's image-token states, even as exact extraction remains limited.

Our contributions are a full-figure benchmark spanning molecule and reaction extraction; direct baselines for both tasks and three system families for Molecule Extraction; and controlled analyses linking input conditions, molecular data scale, depiction distribution, and internal representations to molecular extraction performance.

2 Related Work

Molecular structure recognition. Classical OCSR systems combine image processing, symbol recognition, and graph assembly (Filippov & Nicklaus, 2009); later systems learn document segmentation and structure prediction (Staker et al., 2019). Img2Mol generates SMILES from molecular depictions (Clevert et al., 2021), and DECIMER provides segmentation and recognition tools for scientific publications (Rajan et al., 2023). MolScribe predicts molecular graphs with geometric and chemical constraints, while MolGrapher also uses a graph representation (Qian et al., 2023b; Morin et al., 2023). MolParser and MolRecBench-Wild study recognition on document-derived depictions (Fang et al., 2025; Yang et al., 2026). PaperMolBench retains the full figure and requires a natural-language query to select the target before recognition.

Document and chemical vision. DocVQA and TextVQA pair questions with document images and text-rich natural images, respectively (Mathew et al., 2021; Singh et al., 2019), while RxnScribe parses complete reaction diagrams into structured records (Qian et al., 2023a). Reaction Extraction selects one queried reaction from a full scientific figure and scores its reactant and product structures. ChemVLM, ChemDFM-X, and CheMM-R1 evaluate chemical visual recognition and reasoning (Li et al., 2025; Zhao et al., 2024; Huang et al., 2026). PaperMolBench adds original figure context and query-specified targets; Qwen3-VL and InternVL3.5 provide the open-model baselines (Bai et al., 2025; Wang et al., 2025). ReAct and Self-Refine motivate the iterative actions and revision used in our Agentic Pipeline (Yao et al., 2023; Madaan et al., 2023).

Real depictions and adaptation. Document molecules vary in line style, image quality, abbreviations, and surrounding annotations (Morin et al., 2023; Fang et al., 2025). Our paired crops and RDKit renderings measure recognition when the target crop is provided and when its depiction is standardized. Continued pretraining provides a route to domain adaptation (Gururangan et al., 2020). Using fixed downstream inputs, we study molecular data scale through CPT and depiction distribution through paired SFT. To examine the remaining generation errors, we probe the trained representations for structural information (Hewitt & Liang, 2019).

3 PaperMolBench

3.1 Tasks and benchmark organization

Given a scientific figure I and query q, PaperMolBench defines two tasks.

Molecule Extraction. The system predicts one requested molecular SMILES, $\hat{s} = f_\theta(I, q)$. Queries refer to identifiers, names, panels, spatial relations, or molecular roles; the reference $s^\star$ specifies connectivity, bond types, charges, and annotated stereochemistry.

Reaction Extraction. The system predicts the requested reaction sides, $(\hat{r}, \hat{p}) = g_\theta(I, q)$. Periods separate multiple molecules on one side. Agents and conditions remain metadata; evaluation uses reactant and product structures.

The benchmark contains 355 reviewed cases from 149 figures in 129 papers: 204 Molecule Extraction cases from 83 figures and 151 Reaction Extraction cases from 84 figures, with 18 figures shared. Multiple queries can address different targets or reactions in one figure. Appendix A.1 describes annotation and adjudication. The 72-case diagnostic subset pairs each molecule query with its full figure, verified target crop, and RDKit rendering; crops retain an adjacent molecule identifier when it belongs to the target region. Table 1 gives the composition and annotations. Query-cue categories are nonexclusive.

Figure 1 summarizes the benchmark, system comparison, molecule diagnostics, and training study.

3.2 Evaluation Metrics

Our primary metric is exact match (EM) of canonical isomeric SMILES:

$$\mathrm{EM} = \frac{100}{N} \sum_{i=1}^{N} \mathbf{1}[C(\hat{s}_i) \neq \varnothing \ \wedge \ C(\hat{s}_i) = C(s_i^\star)] \,, \tag{1}$$

where C is RDKit canonicalization; invalid predictions score zero. Canonicalization removes graph-traversal differences while preserving molecular identity and stereochemistry. Unspecified substituents use wildcard atoms, `[*]`, and defined abbreviations are expanded. A fixed, reference-independent parser extracts the prescribed output and normalizes only known wrappers, without changing connectivity or stereochemistry. Appendix A specifies these rules.

Fingerprint similarity (FPS) averages Tanimoto similarity between chirality-aware radius-2, 2,048-bit Morgan fingerprints (Rogers & Hahn, 2010), reported on a 0–100 scale:

$$\mathrm{FPS} = \frac{100}{N} \sum_{i=1}^{N} \frac{|F(C(\hat{s}_i)) \cap F(C(s_i^\star))|}{|F(C(\hat{s}_i)) \cup F(C(s_i^\star))|}. \tag{2}$$

FPS captures partial agreement; EM measures complete extraction. Reaction Extraction applies both metrics separately to reactants and products. Joint EM requires both sides to match:

$$\mathrm{JointEM} = \frac{100}{N} \sum_{i=1}^{N} \mathbf{1}[C(\hat{r}_i) = C(r_i^\star) \ \wedge \ C(\hat{p}_i) = C(p_i^\star)] \,. \tag{3}$$

Prompts, image processing, and decoding settings appear in Appendix A.

Table 1: Benchmark composition. Diagnostic is a Molecule Extraction subset. Parenthesized percentages use cases except the multiple-cases row, which uses figures.

Statistic	Molecule ($N = 204$)	Reaction ($N = 151$)	Diagnostic ($N = 72$)
Cases	204	151	72
Source figures	83	84	35
Source papers	79	81	35
Figures with ≥ 2 cases	71 (85.5%)	67 (79.8%)	26 (74.3%)
Median cases per figure (range)	3 (1–3)	2 (1–2)	2 (1–3)
Location or panel cue	100 (49.0%)	5 (3.3%)	40 (55.6%)
Explicit *labeled* cue	32 (15.7%)	12 (7.9%)	10 (13.9%)
Cases with multiple reactants	–	4	–
Cases with multiple products	–	10	–
Cases with annotated agents	–	50	–
Cases with other conditions	–	8	–
Median target-box area (% of figure)	–	–	5.0

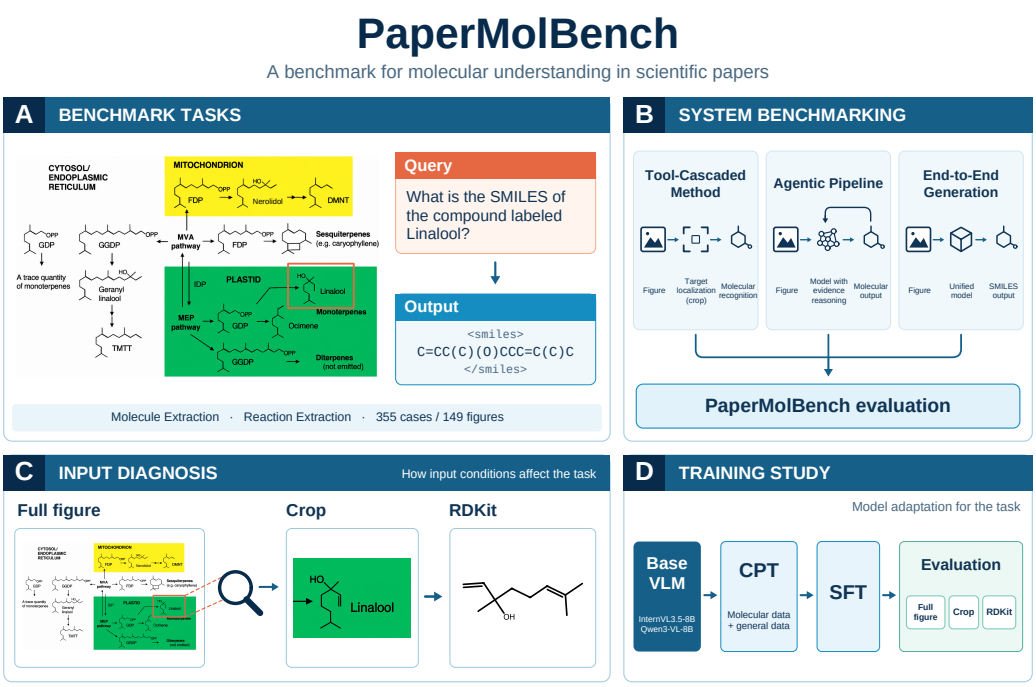


Figure 1: Overview of PaperMolBench. (A) Molecule and Reaction Extraction from queried scientific figures. (B) Molecule systems and direct model baselines. (C) Full-figure, crop, and RDKit molecule inputs. (D) CPT and SFT experiments on two open VLM families.

4 System Benchmarking

We evaluate representative frameworks, including the Tool-Cascaded Method, the Agentic Pipeline, and End-to-End Generation, together with widely used VLMs.

4.1 Molecule Extraction systems

We compare four configurations on all 204 molecule cases, using Gemini 3.8 Flash in each.

Tool-Cascaded Method. A molecular detector proposes candidate regions. The VLM selects the region answering the query, and DECIMER converts the selected crop into SMILES (Rajan et al., 2023).

End-to-End Generation. The VLM receives the full figure and query and directly generates the target SMILES.

Agentic Pipeline. Starting from the End-to-End Generation output, the system uses molecular-editing actions and rendered feedback to iteratively inspect and revise the predicted structure.

Test-time sampling and selection. We make three End-to-End Generation calls, canonicalize valid candidates, and remove duplicates. The model then selects among the remaining structures using the original figure and query, together with anonymously shuffled candidate SMILES and molecular renderings. It returns a candidate identifier; a single valid candidate is returned directly. Oracle@3 measures whether any of the three candidates matches the reference, quantifying the candidate pool's coverage before selection.

Table 2: Molecule Extraction on 204 cases. Tokens/query covers the complete pipeline; Oracle@3 is reference-assisted candidate coverage.

Model	System	EM (%)	FPS	Tokens/query
Gemini 3.8 Flash	Tool-Cascaded Method	18.14	46.2	1,202
	End-to-End Generation	44.12	70.3	3,853
	Agentic Pipeline	46.08	71.2	25,370
	End-to-End Generation (3 calls + selection)	**50.00**	74.0	14,363
	End-to-End Generation Oracle@3 (3 calls)	53.43	77.7	11,786

4.2 Molecule Extraction results

End-to-End Generation reaches 44.1% EM, compared with 18.1% for the Tool-Cascaded Method and 46.1% for the Agentic Pipeline (Table 2). Three calls plus selection reach 50.0% using 14,363 tokens/query, versus 25,370 for the Agentic Pipeline. The selector preserves 102 of the 109 questions covered by its candidate pool (93.6%). API configuration and accounting appear in Appendix B.

4.3 Reaction Extraction results

We evaluate End-to-End Generation with Gemini 3.8 Flash on all 151 reaction cases. The model predicts reactant and product SMILES from the full figure and query.

Table 3: Reaction Extraction on 151 cases. Reactant and product cells report FPS / EM (%).

Model	System	Reactant	Product	Joint EM (%)
Gemini 3.8 Flash	End-to-End Generation	76.7 / 49.01	76.2 / 46.36	33.77

Gemini 3.8 Flash reaches 49.01% EM for reactants and 46.36% for products, with both sides correct in 33.77% of cases (Table 3).

5 Capability Diagnostics and Training

End-to-End Generation with Gemini is simple and competitive, yet its best configuration reaches only 50.0% EM accuracy. To understand this gap, we focus on Molecule Extraction as a simpler, more atomic setting for analysis.

5.1 Capability Diagnostics

We compare Gemini and the official InternVL3.5-8B and Qwen3-VL-8B models on the same 72 cases (Figure 1C). Full figures require locating the queried molecule; providing the target crop removes this step, and RDKit images standardize its depiction. All three inputs require identifying the molecular structure and transcribing it into SMILES.

Table 4: Capability diagnostics on 72 matched cases, reported as FPS / EM (%).

Model	Full figure	Target crop	RDKit image
Gemini 3.8 Flash	79.3 / 59.7	76.3 / 51.4	81.9 / 54.2
InternVL3.5-8B	34.0 / 12.8	43.5 / 18.1	55.2 / 25.8
Qwen3-VL-8B	19.8 / 3.6	23.9 / 5.3	30.4 / 7.8



Figure 2: Paired-SFT trajectories on full figures, crops, and RDKit images. Rows report EM and FPS for InternVL3.5-8B and Qwen3-VL-8B. Dashed lines mark official bases; step 0 is SFT-only. Each point averages five generations.

Gemini's EM does not improve with target crops or RDKit renderings (Table 4). InternVL and Qwen improve from full figures to target crops and then to RDKit images. Their gains reflect the added difficulty of locating targets and recognizing real depictions.

All three models still make substantial errors on standardized single-molecule inputs: RDKit EM is 54.2% for Gemini, 25.8% for InternVL, and 7.8% for Qwen.

We therefore examine how molecular data scale and depiction distribution affect extraction.

5.2 Continued Pretraining

Starting from official InternVL3.5-8B and Qwen3-VL-8B weights (Wang et al., 2025; Bai et al., 2025), CPT mixes synthetic multi-molecule OCSR with general text, general multimodal data, and chemical data in roughly equal supervised-token budgets. Synthetic figures place multiple RDKit-rendered molecules in varied layouts and associate a query with one target.

We evaluate ten InternVL checkpoints through step 1852 and nine Qwen checkpoints through step 1694. Each is followed by an independent SFT run with a fresh optimizer; step 0 is an SFT-only baseline. The CPT sequence and training configuration remain fixed within each family. Figure 2 evaluates every point on full figures, crops, and RDKit images. Appendix C gives schedules, learning rates, and data budgets.

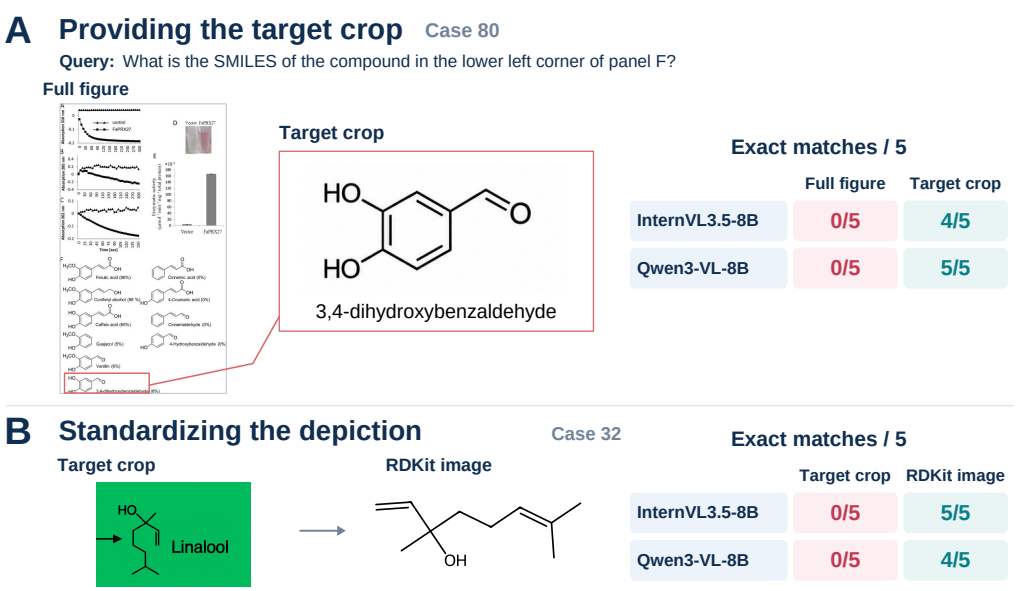


Figure 3: Two examples from the trained models: providing the target crop improves extraction in Case 80; standardizing its depiction improves extraction in Case 32. Counts show exact matches over five generations. Exact outputs appear in Appendix Table 7.

5.3 Paired Supervised Fine-Tuning

For SFT, we combine general text and multimodal data, chemical instruction data, and synthetic multi-molecule OCSR examples. We add 3,000 USPTO and 2,205 MolRecBench-Wild examples (Yang et al., 2026), using their original depictions in Real-depiction SFT and RDKit renderings of the same molecules in RDKit-rendered SFT. All other data and the target SMILES are shared between the two variants. For source $v \in \{\text{real}, \text{RDKit}\}$,

$$\mathcal{L}_{\text{SFT}}^{(v)}(\theta) = -\sum_{(I_i^{(v)}, q_i, s_i) \in \mathcal{D}^{(v)}} \sum_{t=1}^{|s_i|} \log p_\theta(s_{i,t} \mid I_i^{(v)}, q_i, s_{i,<t}). \tag{4}$$

Both variants use one SFT epoch.

5.4 Training Results

Figure 2 compares CPT trajectories against SFT-only models at step 0 and untouched official bases. All scores average five generations.

With Real-depiction SFT, Qwen EM rises from 7.2% to 8.9%, 9.7% to 15.6%, and 17.5% to 36.4% between step 0 and the final checkpoint on full figures, crops, and RDKit images. InternVL rises from 12.5% to 17.2%, 15.8% to 27.8%, and 24.4% to 35.0%. RDKit-rendered SFT also improves all six endpoint comparisons, and FPS agrees across all 12 family–variant–input comparisons.

Real-depiction SFT exceeds RDKit-rendered SFT in crop EM at 10/11 InternVL and 9/10 Qwen checkpoints, with one InternVL tie. This advantage does not consistently extend to full figures or RDKit images. Together with the gains from CPT, these results motivate scaling synthetic molecular data and adding supervision from real scientific figures. Complete checkpoint values appear in Appendix Tables 10 and 11.

Figure 3 illustrates the effects of providing the target crop and standardizing its depiction in the trained models. In Case 80, both models miss the target in the full figure but recover it from the crop. In Case 32, both fail on the real crop but recover the molecule from its RDKit rendering.

6 PROBING STRUCTURAL INFORMATION

Even on RDKit images, Gemini and the trained open models still make substantial errors. We therefore probe whether local structural information remains accessible in a trained Qwen3-VL-8B model. Separate probes compare pairs of images using visual representations or image representations from the language model (Figure 4).

The probe uses 5,000 parent molecules split into 3,500/750/750 train/validation/test sets. Alternative renderings form positive pairs; negatives introduce one bond, atom, substituent, regioisomeric, or stereochemical change. Table 5 reports test ROC-AUC; Appendix E gives category counts and readout details.

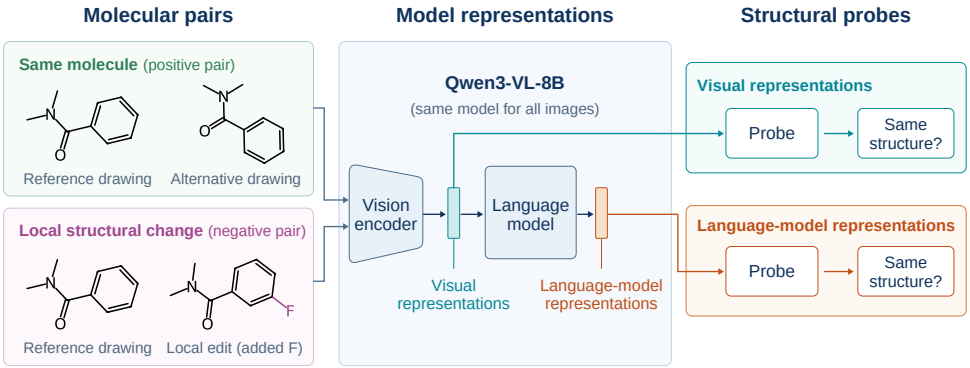


Figure 4: Structural information probes. Visual and language-model representations are probed separately to distinguish different depictions of the same molecule from local structural changes.

Table 5: Structural discrimination from model representations (test ROC-AUC).

Perturbation	Visual features	LLM image states
Bond-order change	0.944	0.984
Charge change	0.824	0.974
Chirality flip	0.891	0.921
E/Z flip	0.924	0.898
Functional-group swap	0.922	0.985
Heteroatom swap	0.935	0.978
One-carbon difference	0.904	0.961
Regioisomer	0.897	0.937
Substituent change	0.927	0.980
Macro-average	0.908	0.957

Mean AUC is 0.908 for visual features and 0.957 for language-backbone image states. The latter reaches 0.921 for chirality flips and 0.898 for E/Z flips; bond-order, heteroatom, and substituent changes are also strongly separable. These local distinctions remain readable in both representations. Exact SMILES generation must assemble them into a complete structure while preserving connectivity and stereochemistry (Weininger, 1988; O'Boyle, 2012). This gap motivates representations and decoding methods that capture molecular topology more directly.

7 CONCLUSION

PaperMolBench measures Molecule Extraction and Reaction Extraction from scientific figures through 355 cases. The best evaluated configuration reaches only 50.0% EM accuracy on Molecule

Extraction. Providing the target crop and standardizing its depiction improve the open models, while even Gemini leaves many standardized molecules unresolved.

Our training experiments show complementary benefits from greater data exposure and real-depiction supervision. CPT improves extraction, and Real-depiction SFT improves crop recognition, though this advantage does not consistently transfer to full figures. These findings motivate combining scalable synthetic data with supervision from real scientific figures. Probing further shows that local structural distinctions remain readable in Qwen representations. Turning this information into complete molecular outputs calls for representations and decoding methods that capture topology more directly.

Limitations. Computational constraints limit the amount of data we can use for training. Our evaluation captures variability across five generations, while variation across independent training runs remains unmeasured. The benchmark could also be expanded to cover more scientific figure styles and literature sources. Finally, our structural probe examines one Qwen model and local molecular changes; extending it to other models and whole-molecule structures would broaden our understanding of the structural information encoded in VLMs.

Reproducibility Statement

Appendices A–E specify evaluation inputs, decoding, scoring, training mixtures, hyperparameters, and probe configurations. Complete checkpoint-level results are provided in Tables 10 and 11.

Ethics Statement

The task concerns extracting structures already depicted in scientific documents. Source attribution and redistribution permissions should accompany any release of figure assets. Predicted structures require verification before use in scientific databases or experimental decisions, particularly when stereochemistry or variable groups are involved.

AI Use Statement

AI assistants were used for literature retrieval, research ideation, implementation support, and drafting and editing manuscript text. Model-generated predictions are scored against fixed reference labels. Scientific claims and numerical results remain the responsibility of the authors.

References

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, et al. LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training. arXiv preprint arXiv:2509.23661, 2025. URL https://arxiv.org/abs/2509.23661.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, et al. Qwen3-VL technical report. arXiv preprint arXiv:2511.21631, 2025. URL https://arxiv.org/abs/2511.21631.

Djork-Arné Clevert, Tuan Le, Robin Winter, and Floriane Montanari. Img2Mol – accurate SMILES recognition from molecular graphical depictions. *Chemical Science*, 12:14174–14181, 2021. doi: 10.1039/D1SC01839F. URL https://pubs.rsc.org/en/content/articlelanding/2021/sc/d1sc01839f.

Xi Fang, Jiankun Wang, Xiaochen Cai, Shangqian Chen, Shuwen Yang, Haoyi Tao, Nan Wang, Lin Yao, Linfeng Zhang, and Guolin Ke. 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. URL https://openaccess.thecvf.com/content/ICCV2025/html/Fang_MolParser_End-to-end_Visual_Recognition_of_Molecule_Structures_in_the_Wild_ICCV_2025_paper.html.

Igor V. Filippov and Marc C. Nicklaus. Optical structure recognition software to recover chemical information: OSRA, an open source solution. *Journal of Chemical Information and Modeling*, 49(3): 740–743, 2009. doi: 10.1021/ci800067r. URL https://doi.org/10.1021/ci800067r.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740/.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Liting Huang, Zhihao Zhang, and Shoujin Wang. CheMM-R1: Enhancing chemical structure recognition and elucidation with reasoning multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26902–26921. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1341. URL https://aclanthology.org/2026.findings-acl.1341/.

Steven M. Kearnes, Michael R. Maser, Michael Wleklinski, Anton Kast, Abigail G. Doyle, Spencer D. Dreher, Joel M. Hawkins, Klavs F. Jensen, and Connor W. Coley. The open reaction database. *Journal of the American Chemical Society*, 143(45):18820–18826, 2021. doi: 10.1021/jacs.1c09820. URL https://pubs.acs.org/doi/10.1021/jacs.1c09820.

Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Wei Li, Mao Su, Shufei Zhang, Wanli Ouyang, Yuqiang Li, and Dongzhan Zhou. ChemVLM: Exploring the power of multimodal large language models in chemistry area. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(1):415–423, 2025. doi: 10.1609/aaai.v39i1.32020. URL https://ojs.aaai.org/index.php/AAAI/article/view/32020.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2200–2209, 2021. URL https://openaccess.thecvf.com/content/WACV2021/html/Mathew_DocVQA_A_Dataset_for_VQA_on_Document_Images_WACV_2021_paper.html.

Lucas Morin, Martin Danelljan, Maria Isabel Agea, Ahmed Nassar, Valery Weber, Ingmar Meijer, Peter Staar, and Fisher Yu. MolGrapher: Graph-based visual recognition of chemical structures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19552–19561, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Morin_MolGrapher_Graph-based_Visual_Recognition_of_Chemical_Structures_ICCV_2023_paper.html.

NVIDIA. NVIDIA Nemotron Nano 2: An accurate and efficient hybrid Mamba-Transformer reasoning model. arXiv preprint arXiv:2508.14444, 2025. URL https://arxiv.org/abs/2508.14444.

Noel M. O'Boyle. Towards a universal SMILES representation – a standard method to generate canonical SMILES based on the InChI. *Journal of Cheminformatics*, 4:22, 2012. doi: 10.1186/1758-2946-4-22. URL https://link.springer.com/article/10.1186/1758-2946-4-22.

Yujie Qian, Jiang Guo, Zhengkai Tu, Connor W. Coley, and Regina Barzilay. RxnScribe: A sequence generation model for reaction diagram parsing. *Journal of Chemical Information and Modeling*, 63(13):4030–4041, 2023a. doi: 10.1021/acs.jcim.3c00439. URL https://pubs.acs.org/doi/10.1021/acs.jcim.3c00439.

Yujie Qian, Jiang Guo, Zhengkai Tu, Zhening Li, Connor W. Coley, and Regina Barzilay. MolScribe: Robust molecular structure recognition with image-to-graph generation. *Journal of Chemical Information and Modeling*, 63(7):1925–1934, 2023b. doi: 10.1021/acs.jcim.2c01480. URL https://pubs.acs.org/doi/10.1021/acs.jcim.2c01480.

Kohulan Rajan, Henning Otto Brinkhaus, M. Isabel Agea, Achim Zielesny, and Christoph Steinbeck. DECIMER.ai: An open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. *Nature Communications*, 14:5045, 2023. doi: 10.1038/s41467-023-40782-0. URL https://www.nature.com/articles/s41467-023-40782-0.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t. URL https://pubs.acs.org/doi/10.1021/ci100050t.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Singh_Towards_VQA_Models_That_Can_Read_CVPR_2019_paper.html.

Joshua Staker, Kyle Marshall, Robert Abel, and Carolyn M. McQuaw. Molecular structure extraction from documents using deep learning. *Journal of Chemical Information and Modeling*, 59(3):1017–1029, 2019. doi: 10.1021/acs.jcim.8b00669. URL https://doi.org/10.1021/acs.jcim.8b00669.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265, 2025. URL https://arxiv.org/abs/2508.18265v2.

David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005. URL https://doi.org/10.1021/ci00057a005.

Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. FineVision: Open data is all you need. In *Advances in Neural Information Processing Systems*, 2026. URL https://neurips.cc/virtual/2026/poster/139537. To appear.

Haote Yang, Hui Wang, Chen Zhu, Jingchao Wang, Linye Li, Hongbin Lai, Huijie Ao, Yongxuan Lv, Jiang Wu, Jiaxing Sun, Lua Chen, Yuanyuan Cao, Ruijie Zhang, Shengxin Lu, Lijun Wu, Bin Wang, and Conghui He. MolRecBench-Wild: A real-world benchmark for optical chemical structure recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 1924–1933, 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Yang_MolRecBench-Wild_A_Real-World_Benchmark_for_Optical_Chemical_Structure_Recognition_CVPRF_2026_paper.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Botao Yu, Frazier N. Baker, Ziqi Chen, Xia Ning, and Huan Sun. LlaSMol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=lY6XTF9tPv.

Zihan Zhao, Bo Chen, Jingpiao Li, Lu Chen, Liyang Wen, Pengyu Wang, Zichen Zhu, Danyang Zhang, Yansi Li, Zhongyang Dai, Xin Chen, and Kai Yu. ChemDFM-X: Towards large multimodal model for chemistry. *Science China Information Sciences*, 67(12):220109, 2024. doi: 10.1007/s11432-024-4243-0. URL https://link.springer.com/article/10.1007/s11432-024-4243-0.

A Evaluation Details

A.1 Dataset annotation and reference versions

Annotation and automated checks. Two annotators recorded answers separately for comparison. For Molecule Extraction, they provided the SMILES of the queried target. For Reaction Extraction, they recorded reactant and product SMILES, together with agents and conditions when present. We checked SMILES parseability, canonicalized valid strings, and compared paired labels using fingerprint similarity. Reaction agents and conditions were also compared as text. These checks supported the subsequent human adjudication.

Human adjudication. A third annotator adjudicated the paired labels using the source figures and queries. Review covered disagreements and cases where both initial annotations were incorrect, including target selection, ambiguous structures, and the assignment of reaction participants, agents, and conditions. For example, review corrected a reaction answer that included participants outside the scope of the query and classified UV-B as a condition rather than an agent. Adjudication sheets retained initial annotations, final labels, and written reasons for corrections.

Dataset composition and reference versions. The complete benchmark has 355 cases from 149 distinct source figures in 129 papers. Molecule Extraction has 204 cases from 83 figures in 79 papers, Reaction Extraction has 151 cases from 84 figures in 81 papers, and 18 figures occur in both tasks. The input-intervention set contains 72 Molecule Extraction cases from 35 figures drawn from 35 papers. One molecule reference does not parse with the evaluator; it remains in the denominator and receives zero for every system. Seven references shared by the 204-case and 72-case evaluations were updated before the paired analysis. The reaction references were reviewed separately; one displayed product label, `O[R]`, is scored as the prompt-specified wildcard form `O[*]`.

In the diagnostic subset, target bounding boxes occupy a median 5.0% of the source-figure area; the 25th and 75th percentiles are 2.9% and 7.1%, respectively.

A wildcard denotes an unspecified substituent at the depicted attachment point. It does not encode the full semantics of a Markush family or equality constraints between multiple variable labels.

A.2 Prompts and image processing

All target-specified diagnostic inputs use the following instruction. For full figures, the first sentence is replaced with the original benchmark query.

> What is the SMILES of the molecule in the image?
> Note: Encode unspecified substituents labeled R, R1, R2, etc. as [*] at their displayed attachment points. If the structure contains abbreviated groups (e.g., Glc, CoA, Ac, OMe, Boc, Ph, etc.), expand them into their full atomic SMILES representation.
> Output ONLY the SMILES string wrapped in XML tags:
> `<smiles>SMILES_STRING</smiles>`

The comparison uses full figures, target crops, and RDKit renderings. Cropping changes the surrounding context, apparent target scale, and query wording; RDKit redraws the same reference molecule with standardized layout and style.

Reaction Extraction uses the original target query followed by this output instruction:

> Note: Encode unspecified substituents labeled R, R1, R2, etc. as [*] at their displayed attachment points. If a structure contains abbreviated groups (e.g., Glc, CoA, Ac, OMe, Boc, Ph, etc.), expand them into their full atomic SMILES representation. If multiple molecules occur on one side of the reaction, join their SMILES with a period (.). Do not include agents or reaction conditions.
> Output ONLY the reactant and product SMILES wrapped in XML tags:
> `<reactant_smiles>REACTANT_SMILES</reactant_smiles>`
> `<product_smiles>PRODUCT_SMILES</product_smiles>`

Open-model diagnostics use the same five official-base generations as the training-curve baselines, with identical images, prompts, references, and scoring. Both omit the request seed. Gemini capability

Table 6: Decoding profiles. Maximum output length is 512 tokens and top-$p = 1$ throughout. Open-model diagnostics and paired training use five unseeded generations.

Evaluation	Temperature	Generations	Tensor parallelism
Capability diagnostics (InternVL / Qwen)	0.7	5	1
Capability diagnostics (Gemini)	0	1	API
Fixed-mixture InternVL study	0	1	4
Paired-SFT study	0.7	5	1
Reaction Extraction (Gemini)	0	1	API

diagnostics use one generation at temperature 0 with request seed 0. Gemini reaction extraction uses one request per case with the request seed omitted. Open-model diagnostics and the paired study use vLLM 0.11.0, Transformers 4.57.6, and PyTorch 2.8.0 for inference. They permit 12 InternVL dynamic patches and use the official model templates. Qwen's minimum and maximum pixel areas are 1,024 and 16,777,216. The image processor still performs its family-specific resizing and alignment after input construction.

A.3 Scoring and repeated generations

Prediction extraction is independent of the reference. Known wrappers are removed before RDKit parsing. The parser includes a conservative recovery rule for one unfinished known closing tag at the end of an otherwise extractable output, in addition to the unclosed-wrapper rule. These rules only normalize the container around the output; they do not alter molecular connectivity or stereochemistry. Canonical isomeric SMILES then determine EM. Fingerprint similarity uses chirality-aware radius-2, 2,048-bit Morgan fingerprints and is multiplied by 100 for display; invalid predictions or unparseable references score zero. Reaction predictions are parsed from the two prescribed tags and scored on each side; joint EM requires both canonical sides to match.

The Gemini reaction evaluation includes all 151 cases, with one provider refusal counted as a failed prediction.

For each paired-SFT point, the three image conditions each contain $72 \times 5 = 360$ predictions. Ten InternVL and nine Qwen CPT checkpoints, plus each family's SFT-only baseline, use both SFT image variants. Mean@5 averages the five dataset-level scores. Oracle@3, used separately in the API system comparison, uses reference labels to measure candidate-pool coverage.

A.4 Qualitative example predictions

Figure 3 uses the final Real-depiction SFT checkpoints: InternVL CPT step 1852 and Qwen CPT step 1694. Each count summarizes five generations. Table 7 reports the first sampled output for each displayed input. The references are Case 80: `O=Cc1ccc(O)c(O)c1` and Case 32: `C=CC(C)(O)CCC=C(C)C`. Figure images are enhanced for readability; all reported predictions use the original benchmark inputs.

Table 7: Exact extracted SMILES from the first unseeded generation for the two selected cases. Correct means an exact match after canonicalization; equivalent SMILES strings need not be textually identical.

Case	Model	Input	EM	Extracted SMILES
80	InternVL	Figure	wrong	`COC1=CC(/C=C/C(=O)O)=CC=C1O`
80	InternVL	Crop	correct	`O=CC1=CC(O)=C(O)C=C1`
80	Qwen	Figure	wrong	`COc1c(O)cccc1C=O`
80	Qwen	Crop	correct	`O=CC1=CC(O)=C(O)C=C1`
32	InternVL	Crop	wrong	`CC(C)CCC(C)(O)C[*]`
32	InternVL	RDKit	correct	`C=CC(C)(O)CCC=C(C)C`
32	Qwen	Crop	wrong	`CC(C)CCC(C)(CC)O`
32	Qwen	RDKit	correct	`C=CC(C)(O)CCC=C(C)C`

B System Implementations and API Accounting

The Tool-Cascaded Method uses MolDet candidate regions, a Gemini selection call, and DECIMER recognition. Its 30 detection-stage failures remain in the 204-case evaluation. The Agentic Pipeline uses molecular-editing operations and rendered feedback to refine the shared End-to-End Generation prediction.

The API endpoint identified its model as `gemini-3.8-flash`, the channel identifier reported by the intermediary. The Molecule Extraction systems used temperature 0 and the pipeline's reasoning configuration; usage metadata included reasoning tokens. The displayed cost sums input and completion tokens, with reasoning already included in completion usage. Logical costs include the initial End-to-End Generation call used by the Agentic Pipeline and all three generations plus selection for the three-call End-to-End Generation configuration. These values describe token usage rather than hardware FLOPs.

End-to-End Generation with three calls and selection uses the original image, query, candidate SMILES, and RDKit renderings. The model sees anonymous shuffled candidate identifiers, without generation order, frequency, or ground truth. The 93.6% coverage-retention rate includes cases with only one valid distinct candidate, for which a separate selection decision is unnecessary. Oracle@3 FPS takes the highest candidate similarity per case; it is reference-assisted like its EM coverage score.

C Training Data and Hyperparameters

C.1 Continued pretraining and checkpoint adaptation

CPT starts from official InternVL3.5-8B and Qwen3-VL-8B-Instruct weights. Its pre-filter mixture targets four approximately equal supervised-token channels: general text (Nemotron pretraining corpora; NVIDIA, 2025), general multimodal data (FineVisionMax and LLaVA-OneVision-1.5 mid-training data; Wiedmann et al., 2026; An et al., 2025), chemical data (including ORD (Kearnes et al., 2021), SMolInstruct (Yu et al., 2024), reaction descriptions, and chemistry-paper material), and synthetic large-figure OCSR. Each synthetic image contains multiple rendered molecules, with a target specified by the associated text. The candidate source pool includes 600,000 such images. Length filtering and packing determine the consumed training sequences.

Table 8: Training settings. Both stages update all model components. Modality-specific learning rates are applied to the corresponding optimizer groups.

Parameter	CPT	Paired SFT
Language-model learning rate	2×10^{-5}	1.6×10^{-5}
Vision-encoder learning rate	2.5×10^{-6}	4×10^{-6}
Alignment-module learning rate	7.5×10^{-5}	8×10^{-5}
Schedule	80-step warmup, then constant	10% warmup, then cosine
Global batch (packed sequences)	168	28
Maximum sequence length	3,072	3,072
Weight decay	0.01	0.1
Adam β_1, β_2	0.9, 0.95	0.9, 0.95
Gradient clipping	1.0	1.0
Precision	bfloat16	bfloat16
Random seed	42	42

Training uses full-parameter adaptation, FlashAttention, gradient checkpointing, packed sequences, and ZeRO-2. The paired SFT implementation uses ms-swift 4.5.3 and Transformers 5.16.1. Seven GPUs with microbatch one and gradient accumulation four give a global SFT batch of 28. CPT applies loss to the training sequence, whereas SFT uses the official chat template and assistant-response loss. The training visual budget is up to three dynamic patches for InternVL and 768 visual tokens for Qwen. Evaluation uses the settings in Table 6.

Each SFT run starts independently from its selected CPT checkpoint. The step-0 SFT-only baselines instead start from the official base weights with the same family-specific data and optimization

settings. CPT and SFT use separate optimizer states, so evaluating a checkpoint does not alter the CPT trajectory.

C.2 Paired SFT mixture and actual sequence budgets

The base mixture is a fixed approximately half-size subset of the non-USPTO instruction mixture. It contains all SMolInstruct task types represented in that subset (Yu et al., 2024), synthetic multi-molecule OCSR, general instruction and image–text data, and chemistry-paper instruction data. The additional 5,205 pairs comprise 2,205 MolRecBench-Wild examples whose converted labels parse and render, plus 3,000 USPTO examples. Real-depiction and RDKit-rendered variants share target labels and intended source rows.

Table 9: Prepared paired SFT datasets. Token counts include encoded input, visual, and output tokens. Packing length is 3,072.

Family	Added images	Rows	Tokens (M)	Packs	Steps
InternVL3.5-8B	Real-depiction	61,314	20.485	6,676	239
InternVL3.5-8B	RDKit-rendered	61,314	20.954	6,829	244
Qwen3-VL-8B	Real-depiction	61,305	17.464	5,692	204
Qwen3-VL-8B	RDKit-rendered	61,305	16.474	5,370	192

InternVL retains 61,314 encoded rows in each variant. Qwen excludes the same nine overlength FineVision examples from both variants, retaining 61,305 rows. Native distributed sampling consumes 6,671/6,825 packed sequences for InternVL Real-depiction/RDKit-rendered training and 5,691/5,369 for Qwen. The small dropped tails leave 5,205 paired examples consumed in both InternVL variants and in Qwen Real-depiction training, and 5,204 in Qwen RDKit-rendered training. Thus the matching unit is the intended image–label pair, with the encoded-token and sampling budgets explicitly recorded.

C.3 Complete paired-SFT checkpoint results

Table 10: InternVL3.5-8B across all evaluated CPT checkpoints. Each cell reports mean EM (%) / mean FPS across five unseeded generations. Real and RDKit denote Real-depiction and RDKit-rendered SFT.

	Full figure		Target crop		RDKit image	
CPT step	Real	RDKit	Real	RDKit	Real	RDKit
0	12.5 / 37.4	11.7 / 36.6	15.8 / 38.3	15.6 / 42.1	24.4 / 49.2	30.0 / 59.7
200	11.4 / 36.9	15.6 / 40.5	22.2 / 47.8	21.7 / 49.4	29.7 / 58.5	27.8 / 58.4
400	12.5 / 38.6	16.4 / 41.2	23.6 / 48.9	22.8 / 49.3	28.9 / 57.5	32.8 / 62.3
600	14.7 / 39.8	16.1 / 40.8	25.8 / 52.6	24.7 / 49.3	34.4 / 63.4	32.8 / 64.2
800	17.5 / 42.6	15.8 / 39.8	28.3 / 52.1	22.2 / 48.3	32.5 / 63.8	31.1 / 61.9
1000	16.7 / 40.6	16.9 / 42.0	25.6 / 50.7	21.9 / 47.7	36.9 / 66.1	36.9 / 66.9
1200	16.1 / 39.4	19.4 / 43.0	25.0 / 48.2	23.1 / 50.0	30.0 / 60.0	35.3 / 64.6
1400	18.9 / 43.2	18.1 / 41.7	25.8 / 51.0	24.2 / 49.8	35.6 / 66.0	32.8 / 64.9
1600	15.8 / 41.1	18.6 / 41.4	23.3 / 50.6	23.3 / 49.0	33.3 / 66.5	37.5 / 68.3
1800	17.5 / 40.7	15.6 / 40.9	26.9 / 52.2	24.7 / 48.7	35.0 / 65.5	35.3 / 66.1
1852	17.2 / 42.1	16.4 / 39.7	27.8 / 53.0	23.9 / 49.9	35.0 / 64.8	38.9 / 66.3

Table 11: Qwen3-VL-8B across all evaluated CPT checkpoints, using the same reporting convention as Table 10.

	Full figure		Target crop		RDKit image	
CPT step	Real	RDKit	Real	RDKit	Real	RDKit
0	7.2 / 29.7	7.8 / 28.8	9.7 / 34.0	8.1 / 31.0	17.5 / 43.7	16.9 / 40.4
200	6.4 / 29.7	8.1 / 30.8	12.2 / 37.2	9.7 / 33.1	18.6 / 47.8	25.8 / 51.6
400	6.7 / 31.0	7.5 / 31.7	7.8 / 37.4	11.1 / 35.5	23.3 / 53.2	26.4 / 55.3
600	8.9 / 33.6	6.9 / 31.4	11.9 / 40.8	9.2 / 33.6	26.4 / 57.9	26.4 / 55.8
800	12.2 / 38.0	8.6 / 32.2	15.6 / 42.8	11.7 / 36.8	30.6 / 61.4	34.7 / 62.0
1000	8.6 / 34.2	7.5 / 30.8	15.0 / 43.2	14.7 / 39.1	33.1 / 61.8	28.6 / 60.4
1200	6.7 / 34.7	11.1 / 35.4	16.4 / 45.1	13.3 / 40.2	31.1 / 61.8	36.4 / 65.5
1400	12.8 / 37.0	10.6 / 35.6	17.8 / 45.1	11.7 / 38.7	36.7 / 64.9	36.7 / 64.2
1600	11.4 / 37.3	8.1 / 31.7	14.4 / 42.1	11.4 / 36.0	38.6 / 67.0	34.7 / 65.2
1694	8.9 / 33.7	9.7 / 33.7	15.6 / 42.6	8.3 / 37.2	36.4 / 64.6	35.6 / 65.9

D Additional Fixed-Mixture SFT Results

This InternVL trajectory uses a larger fixed SFT mixture and single-generation evaluation at temperature 0. It is shown separately from the paired-SFT Mean@5 study. The mixture contains 126,395 examples: 22,624 synthetic large-figure examples, 13,575 USPTO examples, and 90,196 examples from the other sources. Its SFT global batch is 56 and the run lasts 263 steps. The auxiliary set contains 100 synthetic multi-molecule queries.

The synthetic-set score increases from 46% at step 200 to 71% at step 1852. The full-figure and crop trajectories show larger checkpoint-to-checkpoint variation. These auxiliary results complement the repeated-generation comparison of the two paired SFT variants.

Table 12: InternVL CPT plus fixed-mixture SFT results, EM (%). All rows, including the official baseline, use single-generation evaluation. The auxiliary synthetic set contains 100 queries.

CPT step	Figure	Crop	RDKit	Synthetic-100
Official	16.7	20.8	29.2	–
200	18.1	25.0	40.3	46
400	15.3	29.2	41.7	54
600	16.7	25.0	40.3	57
800	15.3	25.0	45.8	64
1000	20.8	27.8	45.8	66
1200	20.8	29.2	43.1	67
1400	22.2	30.6	45.8	69
1600	19.4	29.2	47.2	71
1800	25.0	30.6	52.8	70
1852	18.1	29.2	47.2	71

E STRUCTURAL INFORMATION PROBE DETAILS

The probe experiment uses a task-adapted Qwen3-VL-8B checkpoint with a frozen backbone. We mean-pool the merged visual features for the first readout and the final-layer image-position hidden states for the second. The latter is extracted under an OCSR prompt without the target answer. A nonlinear contrastive projection has hidden dimension 2,048 and output dimension 4,096. It is trained for 30 epochs with learning rate 10^{-3}, weight decay 10^{-4}, contrastive temperature 0.07, and seed 42.

The 5,000 parent molecules are split before constructing their rendering pairs, giving 3,500 training, 750 validation, and 750 test parents. Same-molecule views provide positive comparisons; each perturbation type provides structurally changed negatives. ROC-AUC measures separation of the positive and perturbed pairs. Each type has equal numbers of positive and negative pairs. Macro-AUC averages the nine type-specific AUCs with equal weight.

Table 13: Probe test counts and AUC. n is the number of negative pairs; each row has the same number of positive pairs.

Perturbation	n	Visual features	LLM image states
Bond-order change	997	0.944	0.984
Charge change	201	0.824	0.974
Chirality flip	531	0.891	0.921
E/Z flip	79	0.924	0.898
Functional-group swap	29	0.922	0.985
Heteroatom swap	778	0.935	0.978
One-carbon difference	1,453	0.904	0.961
Regioisomer	921	0.897	0.937
Substituent change	2,611	0.927	0.980