LEARNING TRANSFERABLE REACTION MECHANISMS FROM VISUAL CHEMICAL KNOWLEDGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Reaction mechanisms describe the step-by-step transformations underlying chemical reactions and are central to reaction analysis and synthesis. Learning-based models have achieved strong performance on established mechanism-prediction benchmarks, but transferring them to unseen chemistry remains challenging. Such transfer is difficult because familiar mechanisms must be applied to unfamiliar molecular structures, and some target mechanisms may be poorly covered by the training data. To address these challenges, we introduce **MechaVLM**, a visual framework that combines transferable chemical representations with external mechanistic knowledge. It learns reusable visual features through multiscale chemical grounding and cross-rendering contrastive learning. For open-book prediction, MechaVLM retrieves a fixed set of precedents from 70,384 literature mechanism figures and re-reads relevant visual evidence as the molecular state evolves, directly using the figures without symbolic mechanism parsing. An atom-indexed language decoder then recursively generates executable electron edits to construct the complete mechanism. We further introduce **MechBench**, a challenging literature-derived benchmark with 2,184 mechanisms and 9,146 elementary steps. Across cross-dataset and literature-derived benchmarks, MechaVLM establishes strong zero-shot mechanism prediction. Its closed-book model alone improves Step/Pathway Top-1 by 12.50/13.93 on FlowER→ReactMech, while external visual precedents unlock further gains on challenging OOD reactions. The learned representation also generalizes beyond mechanism prediction to atom mapping and reaction center prediction. Codes and benchmarks will be released.

1 INTRODUCTION

Understanding how chemical reactions proceed through elementary transformations is central to reaction analysis and synthesis. Recent learning-based models have achieved strong reaction mechanism prediction performance on large mechanistically annotated datasets (Joung et al., 2025; Das et al., 2026). However, transferring these models to a different reaction distribution beyond the training distribution remains challenging (Dang et al., 2026). Because obtaining mechanism annotations for every new domain is costly, repeatedly adapting a model to new chemistry is impractical. We therefore study *transferable reaction mechanism prediction*, where a model trained on existing mechanism data must generalize zero-shot to unseen chemical domains.

Such transfer can be difficult for two reasons. First, a familiar mechanism may need to be applied to molecular structures that differ substantially from those seen during training. This requires representations that retain reusable local chemical information. Second, the relevant mechanism may be poorly covered by the training data, so the model must rely on knowledge beyond its learned parameters. To address these, we propose using **visual modeling** as a natural way to support both forms of transfer: (1) First, molecular diagrams expose atoms, bonds, and local structures directly, providing a representation space in which reusable chemical patterns can be learned across different reactions. (2) Second, the same visual space provides direct access to external mechanistic knowledge. Chemical literature contains abundant expert-curated mechanism figures showing molecular structures, intermediates, and electron-flow arrows across diverse real-world reactions (Dang et al., 2026). Using such figures in a symbolic pipeline typically requires first extracting their molecular structures and mechanistic information, which can be costly, error-prone, and may discard useful visual context (Qian et al., 2023). In contrast, when both the input reaction and external precedents

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.

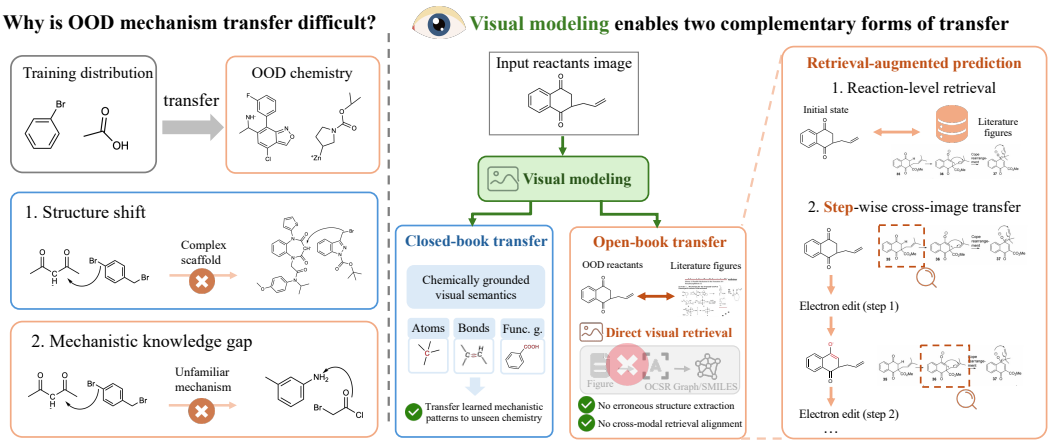


Figure 1: Transferable reaction mechanism prediction with MechaVLM. Visual modeling supports complementary closed-book and open-book transfer for OOD chemistry, while retrieved pathway-level precedents are dynamically transferred to step-wise electron-edit predictions.

are represented as images, retrieval can be performed directly in the same visual space. This avoids both prerequisite symbolic extraction and an additional cross-modal retrieval alignment stage, while preserving the original information in the literature figures.

Retrieving a relevant mechanism figure is not enough. A literature figure usually shows a complete multi-step pathway, while the model predicts one elementary step at a time. Only part of the retrieved figure may be useful for the current step, especially when the precedent and target reaction differ in structure. As the molecular state evolves, the useful part may also change. The key challenge is to transfer the right local mechanistic evidence to each prediction step.

To address these challenges, we introduce **MechaVLM**, a visual framework for transferable reaction mechanism prediction. MechaVLM combines representation learning with retrieval-based mechanism transfer in a common visual space. It first learns chemically grounded representations through multiscale chemical supervision and cross-rendering contrastive learning, encouraging reusable local chemical structure to remain stable across reaction distributions. When external knowledge is available, MechaVLM retrieves a fixed set of visual precedents using the initial reactant image and introduces step-wise cross-image transfer to re-read their local evidence as the molecular state evolves. The resulting evidence is decoded into atom-indexed electron edits, which are applied recursively to generate the complete mechanism.

To evaluate transfer to challenging mechanisms, we introduce **MechBench**, a literature-derived benchmark containing 2,184 mechanisms and 9,146 elementary steps. Across cross-dataset and literature-derived benchmarks, MechaVLM consistently improves zero-shot mechanism prediction. In FlowER→ReactMech transfer, the closed-book model improves Step and Pathway Top-1 accuracy over the strongest baselines by 12.50 and 13.93. Visual precedents provide further gains in the open-book setting and outperform alternative retrieval-augmented baselines. The learned representation also transfers effectively to reaction-related tasks (atom mapping & reaction center prediction).

Our primary contributions are summarized as follows:

- We study transferable reaction mechanism prediction, focusing on zero-shot transfer across reaction distributions, and introduce **MechBench**, a literature-derived benchmark with 2,184 mechanisms and 9,146 elementary steps for evaluating transfer to unseen chemistry.
- We introduce **MechaVLM**, a visual framework that combines transferable chemical representation learning with retrieval-augmented mechanism prediction. It learns reusable local structure through multiscale grounding and cross-rendering contrastive learning, and uses step-wise cross-image transfer to adapt precedent evidence as the molecular state evolves.
- Extensive experiments show that MechaVLM achieves state-of-the-art zero-shot transfer across cross-dataset and literature-derived mechanism benchmarks. The learned visual representation further transfers to atom mapping and reaction center prediction, while access to external visual precedents provides substantial additional gains over closed-book results.

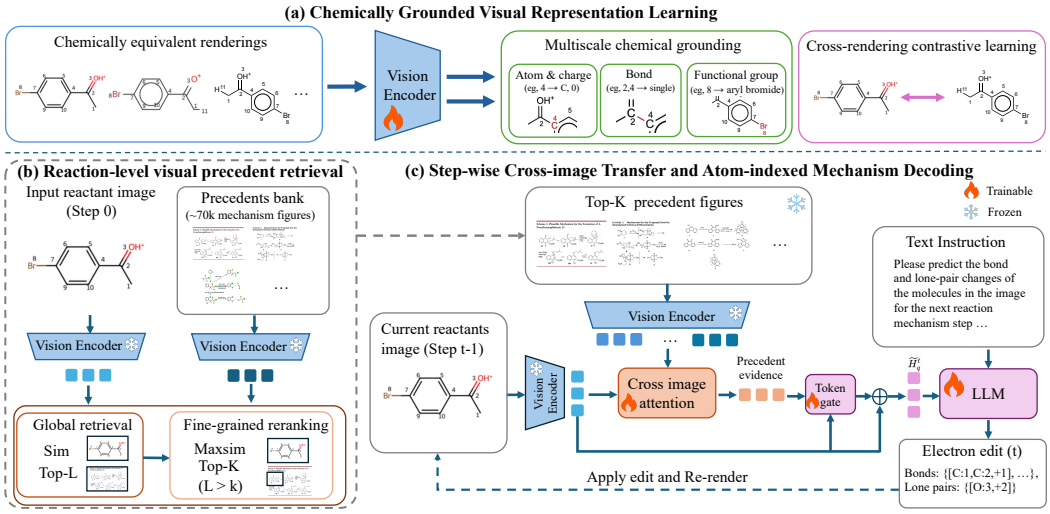


Figure 2: Overview of **MechaVLM**. (a) Chemically grounded and rendering-robust visual representation learning. (b) Reaction-level precedent retrieval via global screening and fine-grained reranking. (c) Step-wise cross-image transfer dynamically extracts relevant evidence from the retrieved precedents and grounds it to indexed query atoms for executable electron-edit decoding. The updated state is re-rendered for the next step.

2 Related Work

Reaction Mechanism Prediction. Reaction mechanism prediction models chemical reactions as sequences of elementary electron movements and molecular state transitions. Early approaches predicted electron-transfer paths for textbook organic reactions (Kayala & Baldi, 2012), followed by learning-based methods that model electron flow through graph edits or electron redistribution (Bradshaw et al., 2019; Bi et al., 2021). More recently, FlowER (Joung et al., 2025) models continuous electron redistribution with Bond–Electron matrices, while DeepMech (Das et al., 2026) further evaluates generalization across reaction classes and chemical components. Despite strong performance on established benchmarks, these methods degrade substantially when transferred to chemistry beyond the training distribution (Dang et al., 2026). We address this limitation from a visual perspective to learn transferable chemical representations.

Chemical Vision-Language Learning and Visual Retrieval. Our work is related to two lines of visual modeling. Recent chemical VLMs, including ChemVLM (Li et al., 2025), TinyChemVL (Zhao et al., 2026b), and ChemVLR (Zhao et al., 2026a), have demonstrated that molecular and reaction images can encode rich chemical information for understanding, reasoning, and prediction. However, these models are not designed specifically for transferring reaction mechanisms across different chemical distributions. A complementary line of work explores retrieval directly from visual content: ColPali (Faysse et al., 2025) and VisRAG (Yu et al., 2025), for example, preserve document information in image form rather than relying entirely on intermediate textual representations. In contrast, our MechaVLM performs image-to-image retrieval from reactant images to literature mechanism figures, and further transfers the retrieved evidence to step-wise mechanism prediction.

3 Method

We propose MechaVLM, a visual framework for transferable reaction mechanism prediction. MechaVLM combines two complementary sources of transfer. First, it learns chemically grounded visual representations through multiscale chemical supervision and cross-rendering contrastive learning, improving generalization across reaction distributions. Second, in the open-book setting, it retrieves mechanism figures directly as visual precedents and reuses a fixed set throughout the reaction, while step-wise cross-image transfer selects the relevant evidence as the molecular state evolves. Finally, an atom-indexed language decoder generates executable electron edits, which are applied and re-rendered recursively to produce the complete mechanism.

3.1 Problem Formulation: Transferable Reaction Mechanism Prediction

Reaction mechanism prediction. Given a reactant set R, let S^0 denote its initial molecular state. We represent a reaction mechanism as a sequence of elementary electron edits (Joung et al., 2025)

$$Y = \left(\Delta^1, \ldots, \Delta^T\right), \qquad S^t = S^{t-1} \oplus \Delta^t, \tag{1}$$

where each Δ^t denotes bond and lone-pair electron changes, and $\oplus$ applies the edit to reach the next molecular state. Starting from the initial reactant state S^0 with no ground-truth products provided, the predictor recursively generates elementary electron edits until termination.

Transfer setting. We focus on zero-shot transfer across reaction distributions. A model is trained on source reactions $\mathcal{D}_{\text{src}} \sim P_{\text{src}}(R, Y)$ and evaluated on target reactions $\mathcal{D}_{\text{tgt}} \sim P_{\text{tgt}}(R, Y)$, where $P_{\text{src}} \neq P_{\text{tgt}}$. No target-domain reactions are used for fine-tuning. The distribution shift may involve molecular scaffolds, reaction families, or broader chemical contexts. We evaluate two inference settings: (1) in the *closed-book* setting, prediction relies only on the knowledge encoded in the model parameters. (2) in the *open-book* setting, the model may additionally leverage a fixed external knowledge bank $\mathcal{B}$. This distinction studies two complementary forms of transfer: generalization from learned chemical representations and transfer from external mechanistic knowledge.

3.2 Chemically Grounded Visual Representation Learning

Transfer across reaction distributions requires visual representations that capture chemical structure rather than drawing-specific patterns. We therefore pretrain the visual encoder with two complementary objectives: *multiscale chemical grounding*, which teaches the model to recognize local chemical units, and *cross-rendering contrastive learning*, which makes the representation stable across different renderings of the same molecule. Given a rendered reactant image $I = \rho_\nu(R)$, the visual encoder $\mathcal{E}_\theta$ produces dense features $H = \mathcal{E}_\theta(I) = \{h_1, \ldots, h_L\}$.

Multiscale chemical grounding. We ground visual features in explicit chemical semantics at three levels: atoms, bonds, and functional groups. Specifically, the model predicts atom identity and formal charge, bond connectivity and type, and local functional-group context. The corresponding supervision is automatically derived from the reactant structure used to render each image. To identify the target chemical units, we render atom indices directly in the image. To specify the atoms involved in each auxiliary task, we overlay atom indices beside their corresponding atom in the image. These indices are randomly permuted across renderings. Let $\mathcal{T}_{\text{chem}} = \{\text{atom}, \text{bond}, \text{functional-group}\}$. For task $k \in \mathcal{T}_{\text{chem}}$, queried unit u, target y, and instruction π_k, we optimize

$$\mathcal{L}_{\text{chem}} = -\mathbb{E}_{(R,k,u,y),\,\nu}\left[\log P_\theta\left(y \mid \rho_\nu(R), u, \pi_k\right)\right], \qquad \nu \sim \mathcal{G}_{\text{vis}}. \tag{2}$$

By explicitly supervising these local chemical units, the encoder is encouraged to capture reusable chemical structure rather than relying on dataset-specific visual patterns.

Cross-rendering contrastive learning. The same chemical structure can appear in substantially different visual forms under different rendering conditions. We therefore generate multiple chemistry-preserving renderings of the same reactants and align their representations through contrastive learning. For each reactant structure R_i, we sample two renderings

$$I_i^{(a)} = \rho_{\nu_a}(R_i), \qquad I_i^{(b)} = \rho_{\nu_b}(R_i),$$

where $\nu_a, \nu_b \sim \mathcal{G}_{\text{vis}}$ vary molecular layout, drawing style, resolution, cropping, atom indexing, and additional graphical perturbations while preserving the underlying chemical structure. For each rendering, we obtain a normalized global representation of the vision encoder $\mathcal{E}_\theta$:

$$z(I) = \text{Norm}\left(\text{Pool}\left(\mathcal{E}_\theta(I)\right)\right). \tag{3}$$

We align representations of the same reactants using a symmetric in-batch contrastive objective,

$$\mathcal{L}_{\text{ctr}} = \frac{1}{2}\left[\text{InfoNCE}\left(Z^{(a)}, Z^{(b)}\right) + \text{InfoNCE}\left(Z^{(b)}, Z^{(a)}\right)\right], \tag{4}$$

with cosine similarity and temperature τ.

Together with multiscale chemical grounding, the representation objective is

$$\mathcal{L}_{\text{repr}} = \mathcal{L}_{\text{chem}} + \lambda_{\text{ctr}}\mathcal{L}_{\text{ctr}}. \tag{5}$$

3.3 Retrieval-Augmented Prediction with Visual Precedents

In the open-book setting, MechaVLM augments mechanism prediction with mechanism figures retrieved from an external precedent bank. MechaVLM retrieves a fixed set of precedents from the initial reactants and reuses them throughout the mechanism rollout. At each prediction step, it rereads these precedents against the current molecular state, allowing different regions to contribute as the mechanism evolves.

Reaction-level precedent retrieval. We construct an external precedent bank $\mathcal{B} = \{I_i^p\}_{i=1}^M$ containing M mechanism figures. Each figure is kept in its original visual form, including molecular structures, intermediates, electron-flow arrows, and textual annotations. Given the initial reactant image $I_q^{\text{init}} = \rho(S_q^0)$, we retrieve precedents using the frozen encoder $\mathcal{E}_{\theta_0}$ from Sec. 3.2. We first use the global representation to shortlist the top-L candidates:

$$s_i^{\text{g}} = \text{sim}\left(z(I_q^{\text{init}}), z(I_i^p)\right), \qquad \mathcal{C}_L = \text{TopL}_{I_i^p \in \mathcal{B}}\, s_i^{\text{g}}. \tag{6}$$

Global similarity may miss a useful precedent when the query reactants match only a small region of a complete mechanism figure. We therefore rerank the shortlisted candidates using late interaction (Faysse et al., 2025) between their dense visual features. Let $H_q^{\text{init}} = \{h_{q,m}^{\text{init}}\}_{m=1}^{N_q}$ and $H_i^p = \{h_{i,n}^p\}_{n=1}^{N_i}$ denote the N_q query and N_i precedent visual features, respectively. We compute

$$\begin{aligned} s_i^{\text{LI}} &= \frac{1}{N_q} \sum_{m=1}^{N_q} \max_{1 \le n \le N_i} \text{sim}\left(h_{q,m}^{\text{init}}, h_{i,n}^p\right), \\ \mathcal{R}_K &= \text{TopK}_{I_i^p \in \mathcal{C}_L}\, s_i^{\text{LI}}. \end{aligned} \tag{7}$$

where s_i^{LI} measures the local visual similarity between the query and precedent i. This allows different parts of the reactant image to match relevant regions within a complete mechanism figure. The resulting top-K precedent set $\mathcal{R}_K$ is kept fixed throughout the mechanism rollout.

Step-wise cross-image transfer. Although the retrieved precedents remain fixed, the useful information within them can change as the molecular state evolves. At prediction step t, we render and encode the current state:

$$I_q^t = \rho(S_q^{t-1}), \qquad H_q^t = \mathcal{E}_{\theta_0}(I_q^t). \tag{8}$$

We concatenate the dense features of the retrieved precedents and use cross-image attention to read precedent features relevant to the current molecular state:

$$H_{\mathcal{R}} = \text{Concat}_{I_i^p \in \mathcal{R}_K}\, H_i^p, \tag{9}$$

$$O_{\text{ext}}^t = \text{softmax}\left(\frac{(H_q^t W_Q)(H_{\mathcal{R}} W_K)^\top}{\sqrt{d}}\right)(H_{\mathcal{R}} W_V), \tag{10}$$

where W_Q, W_K, and W_V are learned projections and d is the attention dimension. Because H_q^t changes after each predicted edit, this attention is recomputed at every step. The model can therefore draw on different regions of the same precedent figures as the mechanism progresses.

Finally, as not every matched precedent region is useful for every query feature, we use a token-wise gate to control how much retrieved information is introduced into each query feature:

$$\begin{aligned} G^t &= \sigma\left(W_G\left[H_q^t; O_{\text{ext}}^t W_O\right] + b_G\right), \\ \widetilde{H}_q^t &= H_q^t + G^t \odot \left(O_{\text{ext}}^t W_O\right), \end{aligned} \tag{11}$$

where σ is the sigmoid function and $\odot$ denotes element-wise multiplication. The resulting representation $\widetilde{H}_q^t$ is used to predict the next elementary edit in Sec. 3.4.

3.4 Executable Mechanism Decoding

At each mechanism step, MechaVLM predicts an executable electron edit on the current molecular state. The same decoder is used in both inference settings: it operates on H_q^t in the closed-book setting and on the precedent-enhanced representation $\widetilde{H}_q^t$ from Sec. 3.3 in the open-book setting.

Atom-indexed electron-edit decoding. At step t, the current molecular state is rendered with atom indices, allowing the model to refer directly to specific atoms in the predicted transformation. The visual representation $\widetilde{H}_q^t$ is provided to the language decoder of the VLM, which autoregressively generates the structured electron edit:

$$\Delta_q^t \sim P_\phi\left(\cdot \mid \widetilde{H}_q^t\right), \tag{12}$$

where $\widetilde{H}_q^t = H_q^t$ in closed-book and denotes the precedent-enhanced representation in open-book.

Each elementary step consists of bond and lone-pair electron changes:

$$\begin{aligned} \Delta_q^t &= \left(\Delta_{q,\mathrm{bond}}^t, \Delta_{q,\mathrm{lp}}^t\right), \\ \Delta_{q,\mathrm{bond}}^t &= \{(a_i, a_j, \delta_{ij})\}, \qquad \Delta_{q,\mathrm{lp}}^t = \{(a_i, \delta_i^e)\}, \end{aligned} \tag{13}$$

where a_i denotes an indexed atom, δ_{ij} specifies a bond change between atoms a_i and a_j, and δ_i^e specifies a lone-pair electron change on atom a_i. The predicted edit can therefore be directly applied to the current molecular state.

Iterative mechanism rollout. After predicting $\widehat{\Delta}_q^t$, we apply the edit and re-render the updated :

$$S_q^t = S_q^{t-1} \oplus \widehat{\Delta}_q^t, \qquad I_q^{t+1} = \rho(S_q^t). \tag{14}$$

The updated image is then used to predict the next elementary step. In the open-book setting, it also re-conditions the step-wise cross-image transfer in Sec. 3.3, while the retrieved precedent set $\mathcal{R}_K$ remains fixed. This process continues until a non-electron-movement action is predicted.

Training. After representation training, we freeze the vision encoder and fine-tune the language decoder on source-domain mechanism data. The open-book model additionally trains the cross-image transfer and fusion modules. At step t, the ground-truth state S_q^{t-1} is rendered as input, and the language decoder is trained with teacher forcing to generate the serialized electron edit Δ_q^t using the standard LLM loss. For open-book training, precedents are retrieved once from initial reactants and reused throughout the trajectory, while cross-image transfer is recomputed for each state.

4 Experiments

Datasets. We evaluate on four reaction-mechanism datasets. FlowER (Joung et al., 2025) contains 250,782/2,801/28,049 train/validation/test reactions (1.45M/15.7K/162.0K elementary steps). ReactMech (Das et al., 2026) contains 29,604 mechanism pathways and 104,964 elementary steps across 67 reaction classes. For zero-shot benchmarks, FukuyamaBench (Dang et al., 2026) contains 319 pathways with 1,997 elementary steps from Fukuyama Mechanism Book. MechBench is below.

MechBench is our proposed literature-derived benchmark constructed from reaction-mechanism figures in 417 publications from *JACS*, *JOC*, and *Organic Letters*. After manually screening articles and retaining only complete, self-contained mechanistic pathways, we curate 2,184 pathways comprising 9,146 elementary steps, with all reconstructed mechanisms verified against the original figures by three chemical experts. Data construction details are provided in the Appendix.

External visual precedent bank. We construct a fixed bank $\mathcal{B}$ of 70,384 reaction mechanism figures collected from the chemical literature. Figures are retained in their original visual form, including molecular structures, mechanistic arrows, and are indexed without conversion to symbolic molecular representations. The bank is used for open-book precedent retrieval during training and inference, without structured mechanism labels or query–precedent relevance annotations. We remove overlaps with all training and benchmark reactions to prevent retrieval leakage. Details are in Appendix.

Metrics. Following (Joung et al., 2025), we report *Top-k step accuracy* and *Top-k pathway accuracy*, with $k \in \{1, 5\}$. Step accuracy measures whether the ground-truth next intermediate is recovered among the top-k predictions for an elementary step. Pathway accuracy requires the complete sequence of mechanistic intermediates to be recovered within the corresponding beam.

Implementation. MechaVLM is initialized from Qwen2.5-VL-3B (Bai et al., 2025). In the first stage, we fully fine-tune the vision encoder with $\mathcal{L}_{\mathrm{repr}}$ for 2 epochs while keeping the LLM frozen, using AdamW with a learning rate of 1×10^{-4} and weight decay 0.1. We then freeze the vision encoder and use it to encode the external precedent bank once, building a nearest-neighbor index over the resulting global representations. In the second stage, we perform mechanism SFT by fully fine-tuning the LLM with a lr of 3×10^{-6}. Both stages use a batch size of 32 on 8 NVIDIA H800.

Table 1: Cross-domain reaction mechanism prediction (%). Models are trained on the source dataset indicated in the first column. White cells report in-distribution (ID) performance, while gray cells denote zero-shot out-of-distribution (OOD) transfer without target-domain adaptation.

		FlowER				**ReactMech**			
		Step		Pathway		Step		Pathway	
Source	**Method**	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
FlowER	NERF (Bi et al., 2021)	67.11	72.60	65.51	75.58	33.32	43.12	14.93	36.32
	Graph2SMILES (Tu & Coley, 2022)	89.09	98.66	92.51	97.76	42.71	55.00	31.84	61.19
	Graph2SMILES+H (Tu & Coley, 2022)	87.39	97.67	89.22	96.19	40.42	52.37	29.67	59.74
	MT (Schwaller et al., 2019)	88.75	98.92	88.31	97.59	44.52	52.45	39.30	58.21
	FlowER (Joung et al., 2025)	88.48	98.60	88.97	97.35	56.84	63.18	34.58	51.47
	ChemVLM-8B (Li et al., 2025)	86.73	96.52	86.94	95.72	31.42	39.08	12.26	16.84
	TinyChemVL (Zhao et al., 2026b)	88.21	98.34	91.06	97.89	37.84	45.26	16.05	22.71
	ChemVLR-8B (Zhao et al., 2026a)	88.54	98.61	91.87	98.32	41.56	48.73	19.83	27.16
	MechaVLM (ours)	90.02	99.20	93.89	99.14	**69.34**	**78.42**	**53.23**	**68.71**
ReactMech	NERF (Bi et al., 2021)	8.62	12.61	0.04	0.12	60.58	60.84	47.76	47.76
	Graph2SMILES (Tu & Coley, 2022)	3.03	4.82	0.60	0.89	97.46	98.84	84.83	89.55
	Graph2SMILES+H (Tu & Coley, 2022)	2.66	4.32	0.54	0.74	96.52	97.64	82.37	87.59
	MT (Schwaller et al., 2019)	5.88	9.39	0.30	0.63	92.73	94.90	74.38	81.34
	FlowER (Joung et al., 2025)	18.86	33.01	2.41	7.27	96.99	99.02	84.09	92.74
	DeepMech (Das et al., 2026)	24.93	32.13	5.60	11.56	98.98	99.58	95.94	96.98
	MechaVLM (ours)	**28.37**	**46.25**	**6.76**	**14.29**	98.38	99.47	96.32	98.21

Table 2: Zero-shot prediction on FukuyamaBench and MechBench (%). Models are trained on the source dataset and evaluated without target-domain adaptation. *: derived from (Dang et al., 2026)

		FukuyamaBench				**MechBench**			
		Step		Pathway		Step		Pathway	
Source	**Method**	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
ReactMech	*Closed-book*								
	NERF (Bi et al., 2021)	6.56	8.86	0.31	0.31	0.72	1.20	0.14	0.18
	Graph2SMILES (Tu & Coley, 2022)	2.35	4.01	0.31	0.31	0.14	0.36	0.00	0.05
	Molecular Transformer (Schwaller et al., 2019)	3.41	7.51	0.00	0.63	0.39	1.07	0.09	0.14
	FlowER (Joung et al., 2025)	0.35	2.80	0.63	0.63	4.62	6.40	0.32	0.32
	DeepMech (Das et al., 2026)	9.01	15.27	1.57	1.57	2.19	3.86	0.14	0.32
	MechaVLM (ours)	**13.72**	**19.28**	**3.13**	**3.13**	**6.17**	**7.06**	**1.01**	**1.37**
	Open-book								
	ChemVLR-8B + CLIP retrieval (Radford et al., 2021)	10.32	14.97	1.57	1.88	3.52	4.89	0.46	0.64
	ChemVLR-8B + OCSR retrieval (Fan et al., 2024)	10.92	15.47	1.88	2.19	3.79	5.18	0.60	0.73
	MechaVLM (ours)	**16.83**	**22.93**	**4.08**	**5.33**	**8.48**	**9.96**	**1.88**	**2.38**
FlowER	*Closed-book*								
	NERF (Bi et al., 2021)	14.07	18.88	1.25	1.25	3.21	4.93	0.23	0.27
	Graph2SMILES (Tu & Coley, 2022)	16.22	21.53	1.25	1.25	1.83	4.76	0.37	0.64
	Molecular Transformer (Schwaller et al., 2019)	14.27	19.38	0.94	2.51	2.54	2.93	0.18	0.23
	FlowER (Joung et al., 2025)	13.92	24.84	1.57	2.51	6.18	8.09	0.14	0.23
	Qwen3-235B-A22B* (Yang et al., 2025)	10.07	–	4.08	6.90	–	–	–	–
	Qwen3-30B-A3B* (Yang et al., 2025)	2.35	–	3.45	4.39	–	–	–	–
	ChemDFM-R-14B* (Zhao et al., 2025)	0.70	0.70	3.13	3.76	–	–	–	–
	ChemVLM-8B (Li et al., 2025)	8.01	11.57	0.00	0.00	4.12	5.64	0.18	0.41
	TinyChemVL (Zhao et al., 2026b)	11.52	15.87	0.31	0.63	5.31	6.92	0.46	0.82
	ChemVLR-8B (Zhao et al., 2026a)	13.52	18.18	0.31	0.94	6.36	7.89	0.73	1.24
	MechaVLM (ours)	**22.28**	**28.24**	**5.96**	**7.52**	**9.59**	**12.46**	**1.56**	**1.92**
	Open-book								
	ChemVLR-8B + CLIP retrieval (Radford et al., 2021)	14.72	19.33	0.63	1.25	6.88	8.43	0.82	1.37
	ChemVLR-8B + OCSR retrieval (Fan et al., 2024)	14.92	20.08	0.94	1.25	7.24	8.60	1.01	1.47
	MechaVLM (ours)	**27.14**	**33.40**	**8.15**	**9.72**	**12.72**	**15.41**	**2.52**	**3.16**

4.1 Zero-Shot Transfer to Unseen Chemistry

We evaluate whether mechanism predictors trained on one reaction distribution transfer to unseen chemistry without target-domain adaptation, under both closed- and open-book settings.

Strong in-domain performance does not imply transferability. Tab. 1 reveals a clear separation between ID accuracy and cross-domain robustness. With FlowER training, MechaVLM improves the strongest baseline by only 0.93/1.38 in Step/Pathway Top-1 on the FlowER test set, whereas the margins increase to 12.50/13.93 after transfer to ReactMech. The same trend holds in reverse: despite lower ID Step accuracy on ReactMech, MechaVLM achieves the strongest zero-shot transfer to FlowER. Meanwhile, chemistry VLMs that are competitive in-domain deteriorate much more sharply under distribution shift. These results suggest that the advantage of MechaVLM primarily comes from better transfer of mechanistic knowledge rather than improved source-domain fitting.

Table 3: Transfer to atom mapping (%).

Method	USPTO -50K (2K)	Schneider	Jaworski
Specialized atom-mapping methods			
Indigo (EPAM Systems, 2026)	30.35	38.98	15.93
G.Mapper (Nugmanov et al., 2022)	79.40	92.20	80.40
RxnMapper (Schwaller et al., 2021)	81.05	93.12	83.15
LocalMapper (Chen et al., 2024)	–	90.08	86.98
Vision-language models			
Qwen2.5-VL-3B (Base VLM)	72.23	84.36	79.63
ChemVLM-8B (Li et al., 2025)	72.94	84.91	80.21
TinyChemVL (Zhao et al., 2026b)	74.38	85.76	81.47
ChemVLR-8B (Zhao et al., 2026a)	75.86	86.63	82.35
MechaVLM(ours)	**83.62**	**93.14**	**89.72**

Table 4: Transfer to reaction center prediction.

Method	Class Unknown Top-1	Class Unknown Top-5	Class Known Top-1	Class Known Top-5
Specialized reaction-center methods				
RetroXpert (Yan et al., 2020)	64.9	–	86.0	–
G2Gs (Shi et al., 2020)	75.8	85.6	90.2	95.0
GDiffRetro (Sun et al., 2025)	86.2	98.8	–	–
Vision-language models				
Qwen2.5-VL-3B (Base VLM)	89.6	93.1	92.7	94.3
ChemVLM-8B (Li et al., 2025)	90.8	94.2	93.8	95.4
TinyChemVL (Zhao et al., 2026b)	92.1	95.7	95.0	96.6
ChemVLR-8B (Zhao et al., 2026a)	94.0	97.4	96.1	98.0
MechaVLM(ours)	**97.2**	**99.3**	**98.4**	**99.6**

Table 5: Cumulative ablation of MechaVLM (Top-1 %). All variants are trained on FlowER. White cells report in-distribution performance, while gray cells denote zero-shot OOD evaluation.

Variant	FlowER (ID) Step	FlowER (ID) Path.	ReactMech Step	ReactMech Path.	FukuyamaBench Step	FukuyamaBench Path.	MechBench Step	MechBench Path.
Input modality								
SMILES input	87.85	90.92	43.38	23.88	10.67	0.63	4.31	0.18
Graph input	88.35	91.36	45.87	25.91	12.32	0.94	5.08	0.32
Closed-book: transferable representation learning								
Mechanism SFT only (image input)	88.41	91.96	52.32	32.47	15.87	1.57	6.82	0.60
+ Multiscale chemical grounding	89.63	93.42	62.74	45.58	19.58	3.76	8.58	1.10
+ Cross-rendering contrastive learning	90.02	93.89	69.34	53.23	22.28	5.96	9.59	1.56
Open-book: visual precedent utilization								
+ Retrieved precedent context	89.91	93.78	70.61	54.72	23.59	6.27	10.08	1.60
+ Initial-state cross-image transfer	90.08	**93.96**	72.54	56.83	24.89	6.90	10.54	1.69
+ Step-wise cross-image transfer	89.97	93.85	75.09	60.18	26.39	7.52	11.96	2.15
+ Token-wise gated fusion (MechaVLM)	**90.11**	93.88	**76.21**	**61.52**	**27.14**	**8.15**	**12.72**	**2.52**

Held-out benchmarks expose a more severe generalization gap. We further evaluate on held-out FukuyamaBench and MechBench. As in Tab. 2, performance decreases substantially for all methods, with the drop being particularly pronounced for complete pathway prediction. This indicates that transferring individual elementary transformations is already challenging, while errors further accumulate during multi-step mechanism rollout. Despite this harder setting, MechaVLM consistently achieves the strongest closed-book transfer. Fig. 3 (left) shows two examples.

External precedents provide complementary knowledge under distribution shift. Open-book prediction further improves both held-out benchmarks. Fig. 3 (right) shows an example. With FlowER training, Step/Pathway Top-1 increases from 22.28/5.96 to 27.14/8.15 on FukuyamaBench and from 9.59/1.56 to 12.72/2.52 on MechBench, with similar gains from ReactMech. MechaVLM also outperforms chemistry VLMs with visual or OCSR-based retrieval, indicating that external precedents are most effective when their mechanistic evidence is transferred to the query (Sec. 4.3).

4.2 Transferable Chemical Representations Beyond Mechanism Prediction

We further explore whether the knowledge learned by MechaVLM transfers beyond mechanism prediction to broader reaction-related tasks: atom mapping and reaction center prediction. Molecular inputs are rendered into 2D images (see Appendix) for training and test in VLM-based models.

Atom mapping. Following (Chen et al., 2024), we adapt VLM-based models on USPTO-FULL and evaluate on a held-out 2K subset of USPTO-50K, Schneider, and Jaworski. As in Tab. 3, MechaVLM outperforms the base VLM without MechaVLM training, existing chemistry VLMs, and matches or surpasses specialized atom-mapping methods. These may be attributed to a twofold advantage: the pre-training phase improves the model's visual comprehension of 2D molecular geometries, and simultaneously transfers a powerful mechanistic prior for complex atom mapping.

Reaction center prediction. Tab. 4 further validates this dual benefit for identifying reaction sites on USPTO-50K. The test set is divided with no overlap with training set of MechaVLM. Pre-training significantly improves MechaVLM's Top-1 accuracy, outperforming strong graph-based baselines like G2Gs. This may be because our pre-training explicitly conditions the model to pinpoint active sites prior to predicting electron flows, intrinsically optimizing the model's localized visual attention.

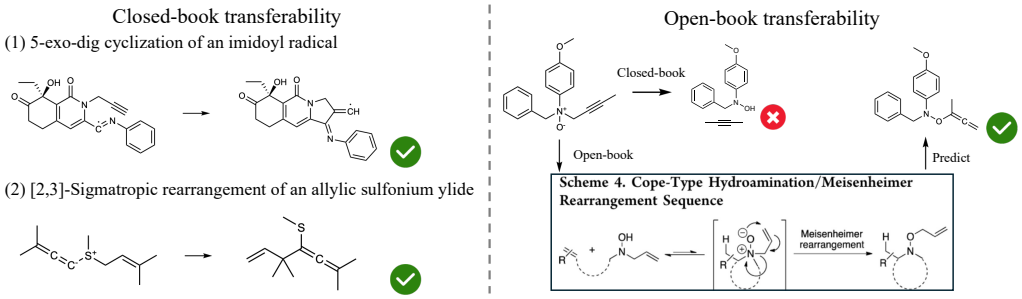


Figure 3: Examples of closed-book and open-book prediction of MechaVLM for OOD reactions.

Table 6: Ablation of external precedent retrieval strategies (%). All models are trained on FlowER.

	ReactMech		FukuyamaBench		MechBench	
Retriever	Step	Path.	Step	Path.	Step	Path.
Random	68.52	52.41	21.83	5.64	9.38	1.47
CLIP image similarity (Radford et al., 2021)	70.74	54.63	23.44	6.27	10.17	1.60
OCSR structure similarity (Fan et al., 2024)	71.08	55.12	23.74	5.96	10.37	1.65
Global similarity	72.93	56.76	24.89	6.90	11.31	2.15
Late interaction	74.41	58.92	25.99	7.52	12.13	2.15
Global → Late Interaction (ours)	**76.21**	**61.52**	**27.14**	**8.15**	**12.72**	**2.52**

4.3 Ablation and Analysis

We ablate the main components of MechaVLM with FlowER as the training source. We report Top-1 Step/Pathway accuracy on FlowER and three OOD benchmarks. More analyses are in Appendix.

Visual inputs improve cross-domain transfer under matched training. As in Tab. 5, with the same mechanism supervision, SMILES, graph, and image inputs perform similarly on FlowER but diverge substantially under OOD evaluation. Image input achieves the strongest transfer across ReactMech, FukuyamaBench, and MechBench, indicating that visual representations provide a better basis for cross-domain generalization even before additional representation learning or retrieval.

Representation learning drives closed-book transfer. As in Tab. 5, adding multiscale chemical grounding and cross-rendering contrastive learning improves FlowER Step/Pathway accuracy by only 1.61/1.93 points, but yields 17.02/20.76-point gains after transfer to ReactMech, with consistent improvements on both held-out benchmarks. The larger gains under distribution shift indicate that the learned features remain useful beyond the source reaction distribution. This suggests that these objectives mainly improve the transferability of the learned representation rather than source fitting.

Step-wise transfer makes better use of visual precedents. Using the same retrieved precedents, direct context addition gives only modest gains, while cross-image transfer performs better by matching the current state to relevant precedent regions. Recomputing this transfer at each mechanism step improves further, and token-wise gated fusion gives the best results. This supports retrieving precedents once but updating the used evidence as the mechanism evolves.

Coarse-to-fine retrieval improves precedent selection. Tab. 6 shows that our learned visual representations outperform both CLIP and OCSR-based retrieval. Late interaction further improves over global similarity by capturing local matches between the query reactants and regions of the full mechanism figure. Combining global retrieval with late-interaction reranking performs best, showing that overall reaction similarity and local structural matching provide complementary signals.

5 Conclusion

We introduced **MechaVLM**, a visual framework for zero-shot reaction mechanism transfer across chemical domains. MechaVLM combines transferable visual representations with external visual precedents, achieving state-of-the-art zero-shot performance across cross-dataset and literature-derived benchmarks. Its learned representation also transfers to atom mapping and reaction center prediction, while external precedents provide further gains in the open-book setting. A remaining limitation is the dependence on the coverage and quality of the external precedent bank, motivating future work on broader and more reliable mechanistic knowledge sources.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experimental design, assist with the interpretation and presentation of experimental results, and support parts of the data annotation process. In particular, generative AI was used to assist the extraction and transcription of reaction-mechanism information from literature figures, after which all annotations were manually reviewed and verified by the authors or domain experts. We did not use generative AI tools to generate synthetic datasets, implement the proposed methods, formulate or prove mathematical claims, or perform experiments. We also used generative AI tools to assist with manuscript drafting, language polishing, structural organization, and readability. All AI-assisted suggestions and annotations were critically reviewed, and all experimental results and scientific claims were verified against the underlying data and outputs. We take responsibility for the final content of this work, including text, claims, annotations, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses existing chemical datasets and published scientific literature and does not involve human subjects, personal data, or user studies. MechBench and the external precedent bank are derived from published literature; any release will respect applicable copyright and licensing restrictions and preserve source provenance. We also remove overlaps across training, retrieval, and evaluation data to reduce leakage. Reaction mechanism prediction may benefit chemical understanding, synthesis planning, and scientific discovery, but may also have dual-use applications involving hazardous chemistry. This work is intended for research and education, and we encourage responsible use under relevant legal, safety, and ethical standards.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architecture, training procedure, datasets, evaluation protocols, and implementation settings in the main text and Appendix. The Appendix further documents the construction and processing of MechBench and the external precedent bank, together with additional experimental details and ablations. Code and benchmark resources will be released to support reproduction of the reported results. For the external precedent bank, we will provide the source DOI, figure index, cropping coordinates, and reconstruction scripts for each retained figure, enabling the corpus to be rebuilt from the original publications.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo 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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Hangrui Bi, Hengyi Wang, Chence Shi, Connor Coley, Jian Tang, and Hongyu Guo. Non-autoregressive electron redistribution modeling for reaction prediction. In *International Conference on Machine Learning*, pp. 904–913. PMLR, 2021.

John Bradshaw et al. A generative model for electron paths. In *ICLR*, 2019.

Shuan Chen, Sunggi An, Ramil Babazade, and Yousung Jung. Precise atom-to-atom mapping for organic reactions via human-in-the-loop machine learning. *Nature Communications*, 2024.

Xingyu Dang, Haocheng Tang, Junmei Wang, and Yanjun Li. Learning mechanistic reasoning for chemical reactions with large language models. *arXiv preprint arXiv:2607.12771*, 2026.

Manajit Das, Ajnabiul Hoque, Mayank Baranwal, and Raghavan B Sunoj. Deepmech: a machine learning framework for chemical reaction mechanism prediction. *Chemical Science*, 2026.

EPAM Systems. Indigo toolkit. https://lifescience.opensource.epam.com/indigo/, 2026. Accessed: 2026-04-23.

Vincent Fan, Yujie Qian, Alex Wang, Amber Wang, Connor W Coley, and Regina Barzilay. Openchemie: an information extraction toolkit for chemistry literature. *Journal of Chemical Information and Modeling*, 64(14):5521–5534, 2024.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, volume 2025, pp. 61424–61449, 2025.

Joonyoung F Joung, Mun Hong Fong, Nicholas Casetti, Jordan P Liles, Ne S Dassanayake, and Connor W Coley. Electron flow matching for generative reaction mechanism prediction. *Nature*, 645(8079):115–123, 2025.

Matthew A Kayala and Pierre Baldi. Reactionpredictor: prediction of complex chemical reactions at the mechanistic level using machine learning. *Journal of chemical information and modeling*, 52(10):2526–2540, 2012.

Shi Xuan Leong, Sergio Pablo-García, Brandon Wong, and Alán Aspuru-Guzik. Mermaid: Universal multimodal mining of chemical reactions from pdfs using vision-language models. *Matter*, 8 (12), 2025.

Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, et al. Chemvlm: Exploring the power of multimodal large language models in chemistry area. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 415–423, 2025.

Ramil Nugmanov, Natalia Dyubankova, Andrey Gedich, and Joerg Kurt Wegner. Bidirectional graphormer for reactivity understanding: neural network trained to reaction atom-to-atom mapping task. *Journal of chemical information and modeling*, 62(14):3307–3315, 2022.

OpenAI. Openai api. `https://openai.com/api/`, 2026. Accessed: 2026-05-03.

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, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.

Philippe Schwaller, Benjamin Hoover, Jean-Louis Reymond, Hendrik Strobelt, and Teodoro Laino. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances*, 7(15):eabe4166, 2021.

Chence Shi, Minkai Xu, Hongyu Guo, Ming Zhang, and Jian Tang. A graph to graphs framework for retrosynthesis prediction. In *International conference on machine learning*, pp. 8818–8827. PMLR, 2020.

Shengyin Sun, Wenhao Yu, Yuxiang Ren, Weitao Du, Liwei Liu, Xuecang Zhang, Ying Hu, and Chen Ma. Gdiffretro: Retrosynthesis prediction with dual graph enhanced molecular representation and diffusion generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12595–12603, 2025.

Zhengkai Tu and Connor W Coley. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *Journal of chemical information and modeling*, 62 (15):3503–3513, 2022.

Chaochao Yan, Qianggang Ding, Peilin Zhao, Shuangjia Zheng, Jinyu Yang, Yang Yu, and Junzhou Huang. Retroxpert: Decompose retrosynthesis prediction like a chemist. *Advances in Neural Information Processing Systems*, 33:11248–11258, 2020.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. In *International Conference on Learning Representations*, volume 2025, pp. 21074–21098, 2025.

Xuanle Zhao, Xinyuan Cai, Xiang Cheng, Xiuyi Chen, and Bo Xu. Chemvlr: Prioritizing reasoning in perception for chemical vision-language understanding. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 24096–24115, 2026a.

Xuanle Zhao, Shuxin Zeng, Xinyuan Cai, Xiang Cheng, Duzhen Zhang, Xiuyi Chen, and Bo Xu. Tinychemvl: Advancing chemical vision-language models via efficient visual token reduction and complex reaction tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1596–1604, 2026b.

Zihan Zhao, Ziping Wan, Lu Chen, Xuanze Lin, Shiyang Yu, Situo Zhang, Da Ma, Zichen Zhu, Danyang Zhang, Huayang Wang, et al. Chemdfm-r: A chemical reasoning llm enhanced with atomized chemical knowledge. *arXiv preprint arXiv:2507.21990*, 2025.

A MechBench: A Literature-Derived Mechanism Benchmark

Motivation. Existing reaction-mechanism benchmarks are primarily derived from curated databases or educational sources. While these datasets provide well-controlled evaluation settings, they cover only a limited subset of the mechanisms encountered in the broader chemical literature. To evaluate zero-shot transfer under a more realistic distribution shift, we construct **MechBench**, a literature-derived benchmark collected from published chemistry articles. MechBench contains **2,184 complete mechanism pathways** comprising **9,146 elementary steps**, collected from **417 journal publications**. The benchmark is used exclusively for evaluation and is never used for model training or target-domain adaptation.

Literature collection. We manually collected publications from three chemistry journals: *Journal of the American Chemical Society* (JACS), *The Journal of Organic Chemistry* (JOC), and *Organic Letters* (OL). Within the 2000–2019 publication period, we first searched for articles whose titles contained the terms *synthesis* or *reaction*, and then manually screened the retrieved papers for explicit reaction-mechanism figures. We retained only figures from the main article that contained a complete, self-contained mechanistic pathway, such that the intermediates and elementary transformations required for pathway reconstruction were available directly from the figure without relying on Supporting Information. This procedure yielded mechanisms from 417 journal publications.

Mechanism curation and quality control. For each retained figure, GPT-6-Astra (OpenAI, 2026) was used to assist the extraction and transcription of the depicted mechanism into an ordered sequence of molecular intermediates and elementary transformations. Atom mappings between consecutive states were assigned using RxnMapper (Schwaller et al., 2021), from which the corresponding bond-order and atom-local electron changes were derived. The resulting pathways were subsequently checked against the original figures by three chemical experts until agreement. We retained only examples for which a complete and evaluable sequence of intermediates could be reconstructed, excluding incomplete, duplicated, or ambiguous pathways.

Held-out evaluation and leakage prevention. MechBench is treated as a fully held-out benchmark. No MechBench examples are used for representation learning, mechanism supervised fine-tuning, or target-domain adaptation. For the open-book setting, benchmark reactions are additionally excluded from the external precedent corpus, preventing retrieval of the evaluated mechanism itself. Thus, both closed-book and open-book results measure transfer to unseen mechanism instances rather than memorization of benchmark examples.

Comparison with FukuyamaBench. As shown in Fig. 4 and Tab. 7, MechBench complements FukuyamaBench, which contains 319 expert-curated mechanism pathways and 1,997 elementary steps from the Fukuyama Mechanism Book. Whereas FukuyamaBench provides a compact

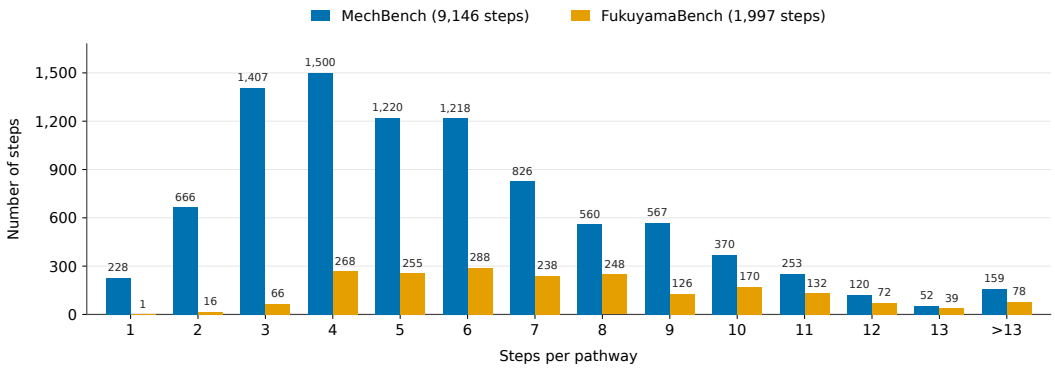


Figure 4: Distribution of elementary steps by pathway length in MechBench and FukuyamaBench. Bars indicate the total number of elementary steps contributed by pathways of each length.

Table 7: Comparison of the fully held-out mechanism benchmarks used in this work. Neither benchmark is used for training or target-domain adaptation.

Benchmark	**Source**	**# Publications**	**# Pathways**	**# Steps**
FukuyamaBench	Mechanism book	–	319	1,997
MechBench	Journal literature	417	2,184	9,146

textbook-derived evaluation, MechBench draws mechanisms from hundreds of journal publications and contains approximately $6.85\times$ as many pathways and $4.58\times$ as many elementary steps. The two benchmarks therefore provide complementary tests of zero-shot mechanism transfer beyond the source training distribution.

B External Visual Precedent Bank

Literature collection and figure screening. We construct a fixed external visual precedent bank from five publication sources: *Journal of the American Chemical Society* (JACS), *The Journal of Organic Chemistry* (JOC), *Organic Letters* (OL), *Nature Communications*, and ChemRxiv. We first identify publications whose titles contain the terms *synthesis* or *reaction*. The publications are rendered as page images, from which candidate figures are cropped using VisualHeist (Leong et al., 2025). We then use Qwen2.5-VL-32B (Bai et al., 2025) to screen the candidate crops with explicit inclusion criteria. A figure is retained only when it depicts a complete and self-contained reaction mechanism, such that the molecular structures, relevant intermediates, and mechanistic transformations are available within the figure. The final bank contains **70,384 mechanism figures**.

Native visual representation. All selected figures are retained as their original figure crops. We do not convert them to SMILES, molecular graphs, or other symbolic representations, nor do we explicitly parse intermediates, electron-flow arrows, or textual annotations. Consequently, molecular structures, curved arrows, reaction annotations, and other visual information remain available to the model in their original form. The precedent bank provides no structured mechanism labels or query–precedent correspondence annotations and is used purely as external visual memory.

Deduplication and leakage prevention. We first remove visually near-duplicate figure crops using CLIP (Radford et al., 2021) similarity, discarding one item from pairs with cosine similarity above 0.95. We intentionally retain mechanistically related but distinct reactions, since such recurring precedents constitute useful external chemical knowledge. To prevent evaluation leakage, all publications used to construct MechBench are excluded from the precedent bank at the DOI level. We additionally perform reaction-instance overlap screening against the supervised mechanism datasets and held-out benchmarks. Qwen2.5-VL-32B is used to flag candidate overlaps based on the depicted molecular structures and reactions, and flagged cases are manually inspected before

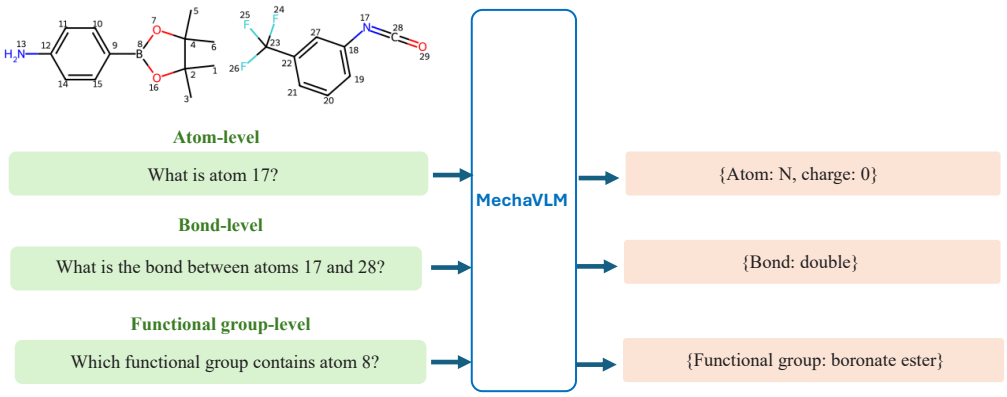


Figure 5: Example of multiscale chemical grounding.

inclusion. This procedure removes identical reaction instances while retaining related mechanistic patterns that differ in their molecular realization.

Fixed external memory. The precedent bank is constructed before mechanism training and remains fixed thereafter. The same bank is accessible during source-domain mechanism training and target-domain open-book inference, but its figures are never used as supervised mechanism targets. Thus, the model can learn how to use retrieved external evidence during training without receiving target-domain examples or updating the external memory at test time. In the closed-book setting, the precedent bank is disabled entirely.

Offline indexing and retrieval. After Stage-I representation learning, the frozen visual encoder $\mathcal{E}_{\theta_0}$ is used to pre-compute both the global descriptor $z(I_i^p)$ and dense visual features H_i^p for every precedent. These representations are cached and reused throughout training and inference, so only the current query image needs to be encoded online. Following Sec. 3.3, retrieval is performed in two stages. We first use global descriptor similarity to retrieve the top $L = 64$ candidates and then rerank them using dense late interaction, retaining the top $K = 4$ precedents for cross-image transfer. Unless otherwise specified, we use these values throughout the experiments.

C Multiscale Chemical Grounding

The multiscale grounding objective in Sec. 3.2 is generated automatically from the structured reactants of the corresponding source-domain training set. No additional molecular corpus, product information, reaction-center labels, or mechanism annotations are used. The objective grounds visual representations at three complementary scales: individual atoms, local connectivity, and functional-group environments.

Visual indexing. All rendered heavy atoms across all reactant molecules are assigned globally unique visual indices. The indices serve only as visual anchors for referring to specific chemical units and are independently randomized for every rendering. Original atom-mapping identifiers are never exposed to the model, preventing persistent indices from becoming shortcuts across examples or rendering variants. Implicit hydrogens are not assigned indices.

Grounding tasks. Given an indexed reactant image, we construct the three instruction-based tasks summarized in Fig. 5. Atom grounding queries a single indexed atom and requires its elemental identity and formal charge. Bond grounding queries a pair of indexed atoms and predicts their connectivity and bond type, with bonded and non-bonded pairs sampled at approximately equal frequency. Functional-group grounding also queries a single indexed atom, but asks the model to identify a functional group containing that atom. Functional-group labels are obtained automatically through SMARTS matching. Because chemical motifs may overlap, an atom can have multiple valid functional-group labels. All matched labels are treated as correct.

Functional-group vocabulary. Functional-group supervision is generated using a fixed SMARTS-based vocabulary covering common local chemical environments in organic reactions. For each

queried atom, we collect all vocabulary entries whose SMARTS matches contain that atom. When multiple patterns overlap, all corresponding labels are retained as valid answers. During training, one valid label is randomly sampled as the textual target for each grounding instance. During evaluation, a prediction is considered correct if it matches any valid label. Atoms that do not participate in any predefined functional group are assigned `none`. The vocabulary is summarized in Tab. 8.

Table 8: Functional-group vocabulary used for atom-anchored functional-group grounding. Multiple entries may be valid for the same queried atom when SMARTS patterns overlap.

Category	**#**	**Canonical vocabulary**
Unsaturated / aromatic	4	alkene, alkyne, conjugated diene, aromatic ring
Oxygen-containing	8	alcohol, phenol, ether, epoxide, peroxide, acetal, hemiacetal, silyl ether
Carbonyl / acyl	15	aldehyde, ketone, carboxylic acid, ester, lactone, amide, lactam, imide, anhydride, acyl fluoride, acyl chloride, acyl bromide, carbonate, carbamate, urea
Nitrogen-containing	13	primary amine, secondary amine, tertiary amine, quaternary ammonium, imine, enamine, oxime, nitrile, nitro, azo, diazonium, azide, isocyanate
Sulfur-containing	9	thiol, thioether, disulfide, sulfoxide, sulfone, sulfonamide, sulfonate ester, thioester, sulfonyl halide
Boron-containing	3	boronic acid, boronate ester, organotrifluoroborate
Halogen-containing	12	alkyl fluoride, alkyl chloride, alkyl bromide, alkyl iodide, vinyl fluoride, vinyl chloride, vinyl bromide, vinyl iodide, aryl fluoride, aryl chloride, aryl bromide, aryl iodide
Phosphorus-containing	4	phosphine, phosphine oxide, phosphonium, phosphate ester
Total	**68**	

Training supervision. All grounding tasks use the standard text output of the VLM. Grounding instances are sampled from chemical units naturally occurring in each reactant image, with the number of queries per rendering capped to prevent large molecules from dominating training. For functional-group grounding, one valid matched label is randomly selected as the supervision target whenever multiple labels apply to the queried atom. The same underlying chemical unit can therefore appear with different visual indices, layouts, rendering styles, and valid functional-group descriptions across epochs.

D Training and Evaluation Details

Mechanism training. Closed-book and open-book models are initialized from the same Stage-I checkpoint and independently trained on the same source-domain mechanism data for 2 epochs. The visual encoder remains frozen, while the language backbone and output head are fully fine-tuned; the open-book model additionally trains randomly initialized cross-image transfer and gated-fusion modules with access to the fixed precedent bank. We use AdamW with a learning rate of 3×10^{-6} and cosine decay, with identical batch size and optimization budget for both settings.

State rendering and rollout. During training, we use ground-truth state conditioning: the ground-truth molecular state $S_q^{t-1,*}$ is rendered to construct the visual input for step t, and the corresponding electron edit $\Delta_q^{t,*}$ is used as the supervision target. During inference, the predicted edit is instead applied recursively,

$$S_q^t = S_q^{t-1} \oplus \widehat{\Delta}_q^t, \qquad I_q^{t+1} = \rho(S_q^t). \tag{15}$$

Atom indices are kept consistent throughout a mechanism rollout, and a fixed rendering style is used at inference time. In the open-book setting, reaction-level precedents are retrieved only once from the initial reactants; subsequent states re-read the same precedents through the step-conditioned local-transfer mechanism in Sec. 3.3.

Decoding. We use deterministic beam search with a beam size of 5. At each mechanism step, candidate structured electron edits are decoded and propagated through their corresponding molecular states. Decoding terminates when the model produces the terminal no-edit output, indicating that

no further electron transformation is predicted. The retained beams are used to compute Top-1 and Top-5 results without stochastic sampling.

Evaluation protocol. We report Step and Pathway Top-k accuracy using a common evaluation pipeline across all retrained methods. FlowER and ReactMech models use their official data splits, while FukuyamaBench and MechBench are used exclusively for zero-shot evaluation and never for hyperparameter tuning or target-domain adaptation. All trainable baselines are retrained from their official implementations on the same source-domain splits. Vision-language baselines receive the same reactant-only image inputs and are adapted using the same mechanism-supervision protocol as MechaVLM. For open-book comparisons, all retrieval-based methods access the same external precedent bank and retrieve the same number of final precedents.

Metrics. Following (Joung et al., 2025), we report *Top-k step* and *Top-k pathway accuracy*, with $k \in \{1, 5\}$. Because the two metrics use different units of averaging, namely individual steps and complete pathways, pathway accuracy is not necessarily bounded by step accuracy. In particular, a small number of long and difficult pathways can contribute many incorrect steps and lower the step-level average, while each complete pathway contributes only one instance to pathway accuracy.

E Implementation Details

Backbone and visual features. We implement MechaVLM on Qwen2.5-VL-3B (Bai et al., 2025). We treat its visual tower together with the multimodal projector as the visual encoder $\mathcal{E}_\theta$. The dense representation H corresponds to the projected visual tokens immediately before they are consumed by the language backbone. This representation is shared across chemical grounding, precedent retrieval, and cross-image transfer. For reaction-level retrieval, we obtain the global descriptor by mean-pooling the dense visual tokens followed by ℓ_2 normalization,

$$z(I) = \mathrm{Norm}\left(\frac{1}{N_I}\sum_{j=1}^{N_I} h_j\right). \tag{16}$$

After Stage-I representation pretraining, the complete visual encoder $\mathcal{E}_{\theta_0}$ is frozen.

Precedent retrieval. Because the external precedent bank is fixed, we pre-compute and cache both the normalized global descriptor $z(I_i^p)$ and dense visual features H_i^p for every precedent using $\mathcal{E}_{\theta_0}$. At inference, only the current query image is encoded online. For the coarse stage, we compute exact cosine similarities against all cached global descriptors and retain the top $L = 64$ candidates. We then perform exact late-interaction reranking using the ℓ_2-normalized dense visual features, following Sec. 3.3, and retain the top $K = 4$ precedents for mechanism transfer. The same retrieval configuration is used during mechanism training and open-book evaluation.

Cross-image transfer and fusion. For open-book prediction, the dense visual tokens of the retrieved precedents are concatenated and accessed through a single multi-head cross-attention module. We use the current query features as queries and the precedent features as keys and values. The attention module uses 8 heads, with its hidden dimension matched to that of the projected visual features. The transferred representation is projected back to the query feature space and incorporated using the token-wise gate defined in Sec. 3.3. The cross-attention, output projection, and gating layers are optimized during open-book mechanism training, while the visual encoder and cached precedent representations remain fixed.

Closed- and open-book models. Both settings use the same visual backbone and mechanism-decoding interface. The closed-book model bypasses the external precedent branch and directly decodes from the current query representation H_q^t. The open-book model additionally applies precedent retrieval, cross-image transfer, and gated fusion to obtain $\widetilde{H}_q^t$ before decoding. Structured electron-edit serialization and execution are described separately in the following section.

OCSR-based structured retrieval baseline. For the structured retrieval baseline, we preprocess the external precedent bank with OpenChemIE (Fan et al., 2024), converting each literature figure into structured representations of the depicted molecular structures and reaction connectivity. The resulting symbolic outputs are used to construct the structured retrieval index, while figures for which OpenChemIE fails to produce a valid extraction are excluded. This pipeline provides an OCSR-based alternative to directly indexing the precedent bank in its native visual form.

F Molecular Image Representation Details

Standard molecular renderings Joung et al. (2025) often suffer from visual overlaps (e.g., overlapping groups and occluded bonds) that hinder accurate reasoning. To create clear and reliable visual inputs, we apply three rendering optimizations:

Implicit hydrogen omission. We omit predictable implicit hydrogens to substantially reduce visual clutter while preserving their mechanistic information. When an implicit hydrogen participates in an electron-transfer step, it is represented through the heavy atom to which it is attached rather than by a separate visual index. For example, if an omitted hydrogen is bonded to atom `C:137`, we denote it as `H:(C:137)`, which uniquely identifies the hydrogen through its attachment site. Electron changes involving this hydrogen can therefore still be specified and mapped back to the corresponding heavy atom, so removing its explicit rendering does not discard the information required for mechanism prediction. We implement this using a modified RDKit `RemoveHs` procedure that removes implicit hydrogens while retaining chemically explicit hydrogen species, such as H–H and H^+.

Adjacent indexing: We place atom indices next to the symbols instead of on top of them. This prevents numbers from blocking bonds and keeps local structures clear. This is implemented by enabling the *addAtomIndices* option when rendering molecular graphs with RDKit.

Adaptive layout optimization. For complex molecules, we dynamically adjust bond lengths, bond angles, and fragment placements until all remaining spatial overlaps are resolved. We formulate this step as a constrained energy minimization problem,

$$E = w_1 E_{\text{overlap}} + w_2 E_{\text{cross}} + w_3 E_{\text{bond}} + w_4 E_{\text{angle}} + w_5 E_{\text{ring}}, \tag{17}$$

where E_{overlap} is a quadratic penalty on non-bonded atom pairs within a cutoff $d_{\min}$, E_{cross} counts intersections between non-adjacent bond segments, E_{bond} and E_{angle} are harmonic deviations from ideal bond lengths l_0 and from ideal bond angles θ_0 (set per hybridization state, e.g. 120° for sp^2, 109.5° for sp^3), and E_{ring} regularizes ring atoms toward their canonical regular-polygon positions. We set $w_1, w_2 \gg w_3, w_4, w_5$, treating overlap and bond crossing as primary objectives while canonical geometry is enforced as a soft prior.

The optimization proceeds in four stages, releasing degrees of freedom from coarse to fine. *(i) Multi-seed initialization.* We generate a pool of candidate layouts by combining CoordGen with RDKit's stochastic 2D embedder under multiple random seeds and structural configurations, and select the lowest-E candidate as the starting point. *(ii) Rigid-fragment reorientation.* Each ring system and each substituent fragment is rotated as a rigid body about its connecting bond over a 30° grid $\{0^\circ, \pm 30^\circ, \pm 60^\circ, \pm 90^\circ, \pm 120^\circ, \pm 150^\circ, 180^\circ\}$ and translated along 12 radial offset directions, accepting the configuration that minimizes E. This stage resolves the majority of inter-fragment crossings without altering any internal geometry. *(iii) Branch-wise local relaxation.* For every atom of degree ≥ 3, we refine its outgoing single-bond branches in two passes: bond angles are first perturbed within $\pm 15^\circ$ of θ_0 in 5° steps, then bond lengths are scaled within $[0.9, 1.4]\, l_0$, with the range escalated to $[1.4, 1.8]\, l_0$ only when the local overlap remains unresolved. Double, triple, and ring bonds are held rigid throughout to preserve canonical chemical geometry. *(iv) Coordinated escape from local minima.* Greedy single-branch updates can stagnate when overlaps require two or more branches to move in concert. Whenever no improving move is found for two consecutive rounds, we jointly search the top $k{=}3$ most crowded branches over their combined offset space, and additionally apply simulated-annealing-style random perturbations accepted with probability $\min(1, \exp(-\Delta E/T))$, with the temperature T decayed geometrically across rounds.

The procedure terminates once no non-bonded pair violates $d_{\min}$ and no bonds cross, or when a patience budget of 5 consecutive non-improving rounds is exhausted.

G Transferable Chemical Representations Beyond Mechanism Prediction

Beyond its primary objective, we demonstrate that vision-based mechanism prediction provides a highly effective prior for broader reaction-related tasks, specifically atom mapping and reaction center prediction. To evaluate this transferability, we render standard molecular inputs into 2D images (see Fig. 6) to reformulate these tasks into visual reasoning tasks. We then fine-tune the model with and without mechanistic training to explicitly quantify the performance gains.

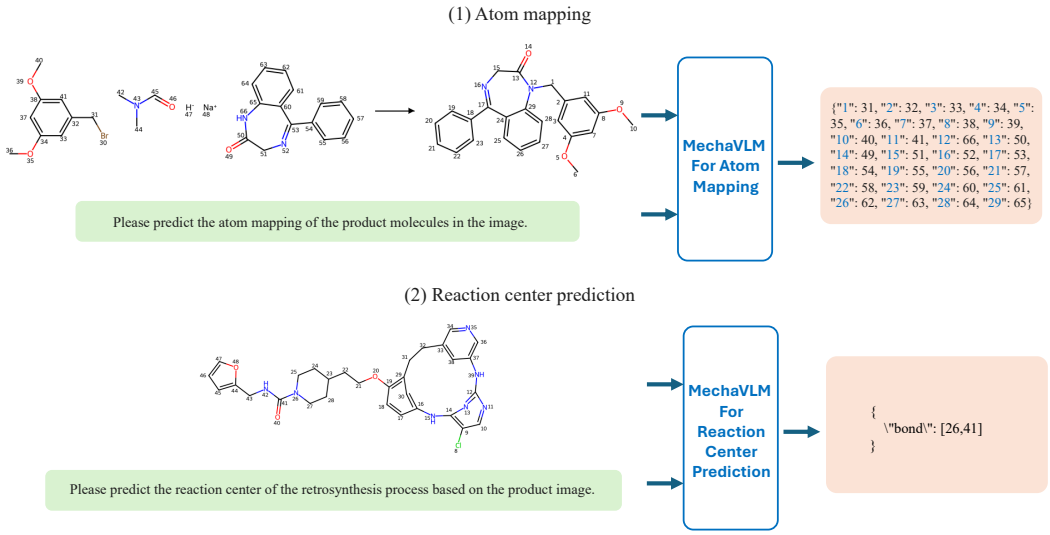


Figure 6: Example of input/output of the image-based atom mapping and reaction center prediction.

Atom Mapping. We finetune MechaVLM on USPTO-FULL from Chen et al. (2024), and the USPTO-50K test set is created by randomly choosing 2k samples not seen in USPTO-FULL (both the training set of atom mapping and the reaction mechanism prediction). This guarantees that there is no data leakage in our experimental setting. In Tab. 3, Schneider test set is also the randomly chosen 2k samples from 50k samples in Schwaller et al. (2021). Jaworski contains annotated 624 samples from Schwaller et al. (2021). These test dataset has no overlap with the training set. Our mechanistic pre-training yields substantial gains over the from-scratch baseline, enabling our vision-centric framework to surpass specialized tools.

Reaction Center Prediction. Tab. 4 further validates this dual benefit for identifying reaction sites on the USPTO-50K dataset. We employ an official split to partition this dataset into training, validation, and test sets. To explicitly prevent data leakage, we implemented a strict reaction-level isolation: all specific reaction instances present in the USPTO-50K test set were completely excluded from the entire MechaVLM training corpus (including the USPTO-FULL dataset). This rigorous isolation ensures that the model is evaluated on entirely unseen molecular structures.

H Additional Experiments and Analysis

We provide additional analyses complementing Sec. 4.3, covering prediction validity and conservation, grounding granularity, retrieval frequency, and scaling with external mechanistic knowledge. Unless otherwise specified, all models are trained on FlowER.

Validity and conservation. As shown in Tab. 9, MechaVLM produces highly valid and chemically consistent predictions, achieving 95.12% validity and conservation across atoms, protons, and electrons. This substantially exceeds sequence-based baselines and slightly surpasses FlowER, indicating that image-based mechanism prediction does not sacrifice chemical validity despite operating directly from molecular diagrams.

Table 9: Comparison of prediction validity and conservation (%).

		Conservation		
Method	**Validity**	Atom	Proton	Electron
Graph2SMILES Tu & Coley (2022)	76.31	30.72	18.67	17.19
Graph2SMILES+H Tu & Coley (2022)	78.80	27.66	19.73	19.02
Molecular Transformer Schwaller et al. (2019)	70.16	39.11	33.69	33.04
FlowER Joung et al. (2025)	94.94	94.94	94.94	94.94
MechaVLM	**95.12**	**95.12**	**95.12**	**95.12**

Effect of Grounding Granularity. We further examine whether the transfer gains from chemically grounded pretraining arise only from atom-level recognition or from jointly modeling chemical units at multiple scales. Starting from the same backbone, we progressively add atom-, bond-, and functional-group-level supervision during Stage-I pretraining while keeping the subsequent mechanism training unchanged. As shown in Tab. 10, atom-level grounding provides the largest initial improvement by anchoring visual features to elemental identity and formal charge. Adding bond supervision further improves transfer by capturing local connectivity and bond order, while functional-group grounding provides additional gains by encoding higher-level recurring chemical environments. Notably, these improvements are much larger on the three OOD benchmarks than on FlowER, suggesting that multiscale chemical grounding mainly improves the transferability of learned representations rather than source-domain fitting.

Table 10: Effect of chemical grounding granularity (%). All variants use the same Stage-I training data and subsequent mechanism training protocol. Only Top-1 Step/Pathway accuracy is reported.

	FlowER (ID)		**ReactMech**		**FukuyamaBench**		**MechBench**	
Grounding supervision	Step	Path.	Step	Path.	Step	Path.	Step	Path.
None	88.41	91.96	52.32	32.47	15.87	1.57	6.82	0.60
Atom	88.86	92.51	57.18	37.06	17.38	2.19	7.54	0.78
Atom + Bond	89.31	93.08	60.74	42.26	18.63	3.13	8.16	0.96
Atom + Bond + Functional Group	**89.63**	**93.42**	**62.74**	**45.58**	**19.58**	**3.76**	**8.58**	**1.10**

Reaction-Level vs. Step-Wise Retrieval. We compare our reaction-level retrieval with step-wise re-retrieval, while keeping all other components unchanged. As shown in Tab. 11, fixing the precedent set per reaction consistently performs better across all three OOD benchmarks. Re-retrieval from evolving intermediates can introduce precedent drift, replacing globally relevant pathways with locally similar but less relevant figures. In contrast, MechaVLM preserves reaction-level context while dynamically selecting step-relevant evidence from a fixed precedent set, and reduces retrieval cost from T queries to one for a T-step pathway.

Table 11: Effect of precedent retrieval frequency (%).

	ReactMech		**FukuyamaBench**		**MechBench**	
Retrieval strategy	Step	Path.	Step	Path.	Step	Path.
Re-retrieve at every step	74.68	58.93	25.87	7.47	12.08	2.15
Reaction-level retrieval (ours)	**76.21**	**61.52**	**27.14**	**8.15**	**12.72**	**2.52**

Scaling with External Mechanistic Knowledge. We evaluate how open-book transfer scales with the size of the external precedent bank by keeping model parameters fixed and restricting retrieval to nested subsets of the full corpus. As shown in Tab. 12, performance improves consistently as more precedents become accessible, with the largest gain appearing when moving from closed-book prediction to partial external memory and smaller gains thereafter. This trend suggests that MechaVLM benefits from broader mechanistic coverage rather than a small set of favorable precedents, and that additional external knowledge can improve transfer at inference time without retraining or target-domain supervision.

Table 12: Scaling with the external visual precedent bank (%).

		FukuyamaBench		**MechBench**	
Bank fraction	**# Figures**	Step	Path.	Step	Path.
0%	0	22.28	5.96	9.59	1.56
25%	17,596	24.29	6.58	10.74	1.88
50%	35,192	25.74	7.52	11.58	2.15
100%	70,384	**27.14**	**8.15**	**12.72**	**2.52**

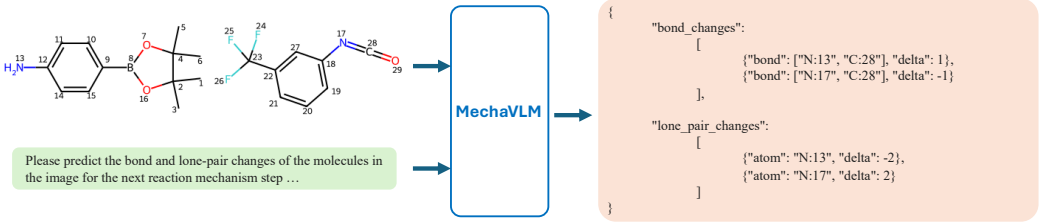


Figure 7: Example of input and output format of MechaVLM in closed-book setting.

I Input and Output Format of MechaVLM

Fig. 7 illustrates an example of complete input and output of MechaVLM in closed-book setting.

J More Visualization of OOD prediction

Fig. 7 shows more examples of MechaVLM's OOD prediction in closed-book/open-book setting.

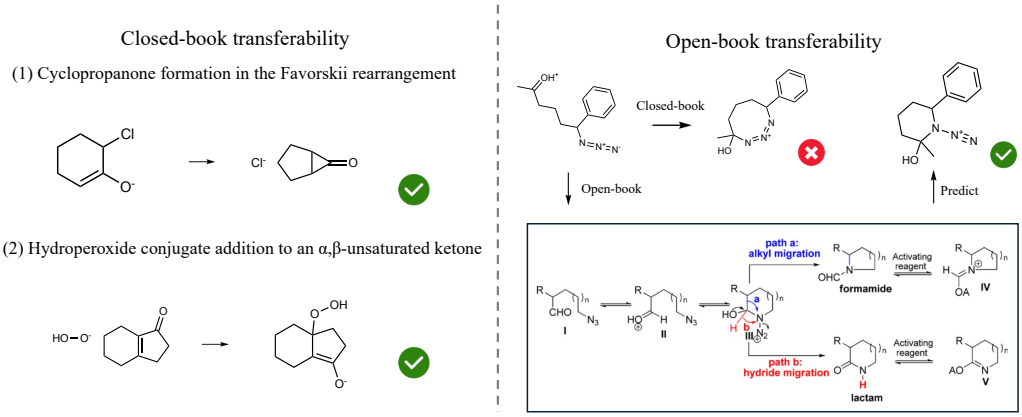


Figure 8: Example of input and output format of MechaVLM in closed-book setting.

K Model Details

MechaVLM adopts Qwen2.5-VL-3B as the backbone VLM. Its main image-encoder and language-model configurations are summarized in Tab. 13.

Table 13: Architecture details of the Qwen2.5-VL-3B backbone.

Component	Qwen2.5-VL-3B
Image Encoder	
Architecture	Vision Transformer
Layers	32
Hidden Size	1280
Attention Heads	16
Patch Size	14
Language Model	
Architecture	Transformer LLM
Layers	36
Hidden Size	2048
KV Heads	2
Head Size	128
Positional Encoding	Rotary