ChemVA: Bridging Visual and Semantic Gaps in Multimodal LLM Chemical Diagram Reasoning

Anonymous ACL submission

Abstract

While Large Language Models (LLMs) have revolutionized scientific text processing, they exhibit a significant capability gap when interpreting chemical reaction diagrams. We identify two bottlenecks restricting current systems: a *Visual Deficit*, where generic vision encoders struggle to resolve the strict topological connectivity of dense molecular graphs, and a *Semantic Disconnect*, where standard linear strings (*e.g.*, SMILES), although structurally precise, are often poorly aligned with the chemical semantic abstractions used by LLMs for reasoning. To bridge these gaps, we propose the **Chemical Visual Activation (ChemVA)** framework, which employs a novel *Visual Anchor* mechanism to ground functional groups via hybrid-granularity detection, followed by a semantic alignment approach that translates recognized structures into language-accessible chemical representations for downstream LLM reasoning. We evaluate our approach on **OCRD-Bench**, a newly constructed benchmark spanning recognition, chemical knowledge, and mechanistic reasoning. Experiments show that ChemVA achieves 92.0% structural recognition accuracy. It further delivers a consistent performance gain of approximately 20 percentage points across 9 diverse LLMs, allowing open-weights models to approach strong proprietary direct-input systems on complex chemical reasoning tasks.

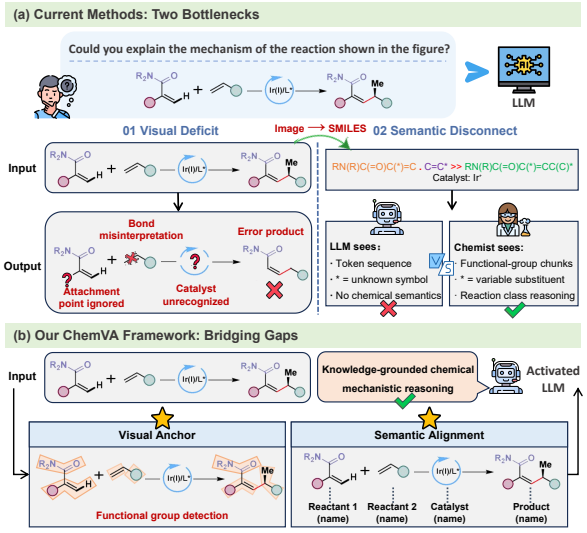


Figure 1: Overview of limitations in LLM-based chemical reaction diagram understanding and the proposed ChemVA framework. (a) Existing pipelines exhibit a Visual Deficit in Reaction Diagram Parsing and a Semantic Disconnect under SMILES-based representations. (b) ChemVA bridges these gaps via functional-group–grounded Parsing and Semantic Activation.

1 Introduction

The integration of Large Language Models (LLMs) into chemical research has accelerated progress in "AI for Chemistry", especially in domains that require complex reasoning, such as organic chemistry (Boiko et al., 2023; Zhang et al., 2024b). General-purpose models such as GPT-5 and Gemini2.5-pro have shown strong performance on text-centric chemical tasks, including synthesis planning and property prediction (Bran et al., 2025; Xian et al., 2025; Tran et al., 2025). However, a central modality of chemical knowledge in the literature remains visual, *i.e.*, molecular structures and reaction diagrams, and current state-of-the-art models still exhibit a substantial gap on this modality (Li et al., 2025; Guo et al., 2023). We trace this gap to two bottlenecks: a *Visual Deficit* in structural perception and a *Semantic Disconnect* in language-accessible chemical representations (Figure 1).

First, a critical **Visual Deficit** remains in chemical image understanding. Chemical structures are dense graph objects whose meaning is determined by topological connectivity and stereochemical geometry, rather than by the global semantic cues that CLIP-style encoders typically emphasize (Morin et al., 2023; Staker et al., 2019). Prior evaluations (Qian et al., 2023b; Clevert et al., 2021) show that

general-purpose models often hallucinate atoms, confuse bond types, and degrade under rotations or visually crowded graphs; even a single local error can change molecular identity and invalidate downstream analysis. Existing Optical Chemical Structure Recognition (OCSR) tools, including MolScribe(Qian et al., 2023b), Decimer(Rajan et al., 2020, 2023), and related methods(Chen et al., 2025; Oldenhof et al., 2021; Shah et al., 2025; Wang et al., 2025; Zhuang et al., 2025), have improved image-to-graph translation, but most formulations remain *atom-level*, representing molecules as sequences of nodes and edges. Such representations do not explicitly capture *substructure-level* perception, where functional groups serve as stable semantic units for expert interpretation.

Second, even when structure recognition yields a correct linear representation such as SMILES, downstream reasoning in LLMs can still be hindered by what we term a **Semantic Disconnect**. This term refers to the mismatch between chemically valid symbolic encodings and the semantic cues that LLMs rely on during inference. While SMILES provides a precise specification of molecular graphs, it is often a weak semantic interface for a language model and does not reliably expose the functional concepts that support chemical reasoning. Prior evidence indicates that SMILES alone is often less effective than natural-language chemical representations in eliciting chemically meaningful reasoning (Edwards et al., 2022). Conversely, chemical entity names (*e.g.*, IUPAC names(Heller et al., 2015)) provide explicit descriptors that more consistently elicit knowledge about reactivity and synthesis (Cui et al., 2025; Diao et al., 2023). Therefore, pipelines that follow "Image → SMILES → Reasoning" can underutilize the reasoning capacity of LLMs, because the intermediate representation is not aligned with the semantic abstractions that guide inference.

To bridge these gaps, we propose Chemical Visual Activation (**ChemVA**), a two-phase framework comprising **Reaction Diagram Parsing** and **Semantic Activation**. ChemVA trains a functional-group-grounded vision-language model (FG-VLM) for reaction diagram parsing, while using existing downstream LLMs as reasoning backends without modifying their parameters or architectures. Motivated by the "chunking" mechanism in human cognition (Gobet et al., 2001), the Reaction Diagram Parsing phase uses FG-VLM to ground functional groups as *Visual Anchors* through reaction deconstruction and hybrid-granularity visual anchor modeling, enabling topology-faithful structural reconstruction. The Semantic Activation phase then performs semantic alignment by resolving recognized structures into language-accessible chemical representations, such as entity names, to construct semantically grounded prompts for downstream LLM reasoning. We further construct **OCRD-Bench**, a unified multimodal benchmark derived from graduate-level chemistry entrance exams, to evaluate the full process from visual perception to chemical knowledge and mechanistic reasoning. Experiments on OCRD-Bench show that ChemVA consistently outperforms strong baselines: FG-VLM reduces error accumulation while Semantic Activation improves downstream reasoning by providing language-accessible chemical representations that better align with the semantic abstractions used by LLMs.

Our **contributions** can be summarized as follows: **(1)** We propose ChemVA, a two-phase framework that bridges the *Visual Deficit* and *Semantic Disconnect* through Reaction Diagram Parsing and Semantic Activation. **(2)** We develop FG-VLM, a hybrid-granularity OCSR model that grounds functional groups as *Visual Anchors* and reconstructs topology-faithful molecular graphs via Directional Vector Matching, going beyond purely atom-level serialization. **(3)** We construct OCRD-Bench, a challenging multimodal benchmark derived from graduate-level entrance exams, with broad reaction coverage and a hierarchical evaluation protocol spanning visual perception, chemical knowledge, and mechanistic reasoning. **(4)** We show that ChemVA consistently improves performance across a wide range of LLMs, yielding substantial gains on complex chemical reasoning tasks.

2 Related Work

2.1 Methods for Chemical Reaction Diagram Understanding

Prior work on chemical reaction diagram understanding mainly follows two directions: symbolic structure serialization and multimodal LLM-based reasoning. The first line converts raster diagrams into machine-readable molecular graphs or SMILES using OCSR systems such as MolScribe (Qian et al., 2023b), RxnScribe (Qian et al., 2023a), OpenChemIE (Fan et al., 2024), DECIMER (Rajan et al., 2020, 2023), and ReactionDataExtractor1.0/2.0 (Wilary and Cole, 2021, 2023). These

methods typically formulate recognition as atom-level structured prediction or graph-to-text encoding (Yu et al., 2024). Although effective on clean inputs, they primarily act as structure digitizers and do not explicitly model functional groups as higher-level semantic units for robust perception and downstream reasoning.

The second line couples visual encoders with LLMs or external resources. Implicit alignment methods such as ChemVLM (Li et al., 2024) tune generic vision-language backbones on chemical data, but suffer from structural hallucinations on dense molecular graphs. Retrieval-based systems such as MolMole (Chun et al., 2025) and Doc2SAR (Zhuang et al., 2025) improve factual grounding by retrieving molecular information from documents or databases, but they do not directly address topology-faithful visual parsing. In contrast, ChemVA combines a hybrid-granularity visual parser that grounds functional groups as *Visual Anchors* with a semantic alignment stage that converts recognized structures into language-accessible chemical representations for downstream LLM reasoning.

2.2 Benchmarks for Chemical Reaction Diagram Understanding

Existing benchmarks usually evaluate isolated parts of the reaction-diagram understanding pipeline. Recognition-centric benchmarks, such as USPTO-based settings (Oldenhof et al., 2021), focus on Image-to-SMILES accuracy without testing downstream chemical reasoning. Text-centric scientific reasoning benchmarks, including SciBench (Wang et al., 2023) and ChemBench (Zhang et al., 2024a), evaluate knowledge and problem solving but bypass visual perception. Recent multimodal benchmarks such as MMCR (Li et al., 2024) and MaCBench (Alampara et al., 2024) introduce visual reasoning, while molecular and reaction reasoning benchmarks further evaluate chemical understanding (Runcie et al., 2025; Xu et al., 2026, 2025; Bartmann et al., 2026). However, these resources do not specifically require models to parse reaction diagrams and generate open-ended mechanistic explanations. More general multimodal reasoning benchmarks (Yue et al., 2024; Lu et al., 2022) are also not tailored to chemical reaction diagrams. OCRD-Bench fills this gap by requiring end-to-end generative reasoning from visual reaction recognition to multi-step mechanism explanation.

3 Method

This section presents the proposed **ChemVA** framework for chemical reaction diagram understanding. Given a reaction image, ChemVA first performs **Reaction Diagram Parsing** with a trained functional-group-grounded vision-language model (FG-VLM), which recovers molecular topology through reaction deconstruction, hybrid-granularity visual anchor modeling, and directional vector matching. It then performs **Semantic Activation**, which converts recognized structures into language-accessible chemical representations for downstream LLM reasoning. We also summarize the data construction and training procedure used to equip FG-VLM with hybrid-granularity perception.

3.1 Reaction Diagram Parsing via FG-VLM

Given a chemical reaction diagram image I, our goal is to recover a structured reaction representation that supports downstream reasoning. Prior approaches typically cast diagram understanding as atom-level sequence generation, *i.e.*, emitting an ordered list of atoms V_{atom} and their bonds (Qian et al., 2023b). However, organic molecules contain recurring rigid motifs whose semantics are determined at the substructure level; purely atom-level generation therefore produces long outputs and is prone to local errors that propagate to global topology. To address this, we present functional-group-grounded vision-language model (FG-VLM).

3.1.1 Reaction Diagram Deconstruction

FG-VLM first acts as a layout parser that detects reaction constituents: FG-VLM first detects reaction constituents $\mathcal{O} = \{(c_i, \mathbf{b}_i)\}_{i=1}^{N}$, where c_i denotes the constituent class and $\mathbf{b}_i$ the corresponding bounding box. We use five classes: Reactant, Product, Condition, Arrow, and Plus. Reactant and product regions are cropped into molecular sub-images $\mathcal{I}_{crops}$ for fine-grained structure recognition.

3.1.2 Hybrid-Granularity Visual Anchor Modeling

Hybrid-Granularity Nodes and Bonds. For each molecule crop I_{crop}, FG-VLM predicts a hybrid graph $G' = (V_{\text{atom}} \cup V_{\text{super}}, E')$, where V_{super} contains functional-group super-nodes, V_{atom} contains residual atoms, and E' encodes bonds among them. Each super-node g is represented as $(id_g, name_g, \mathbf{b}_g)$, where $name_g$ specifies the functional-group type and $\mathbf{b}_g$ is its bounding

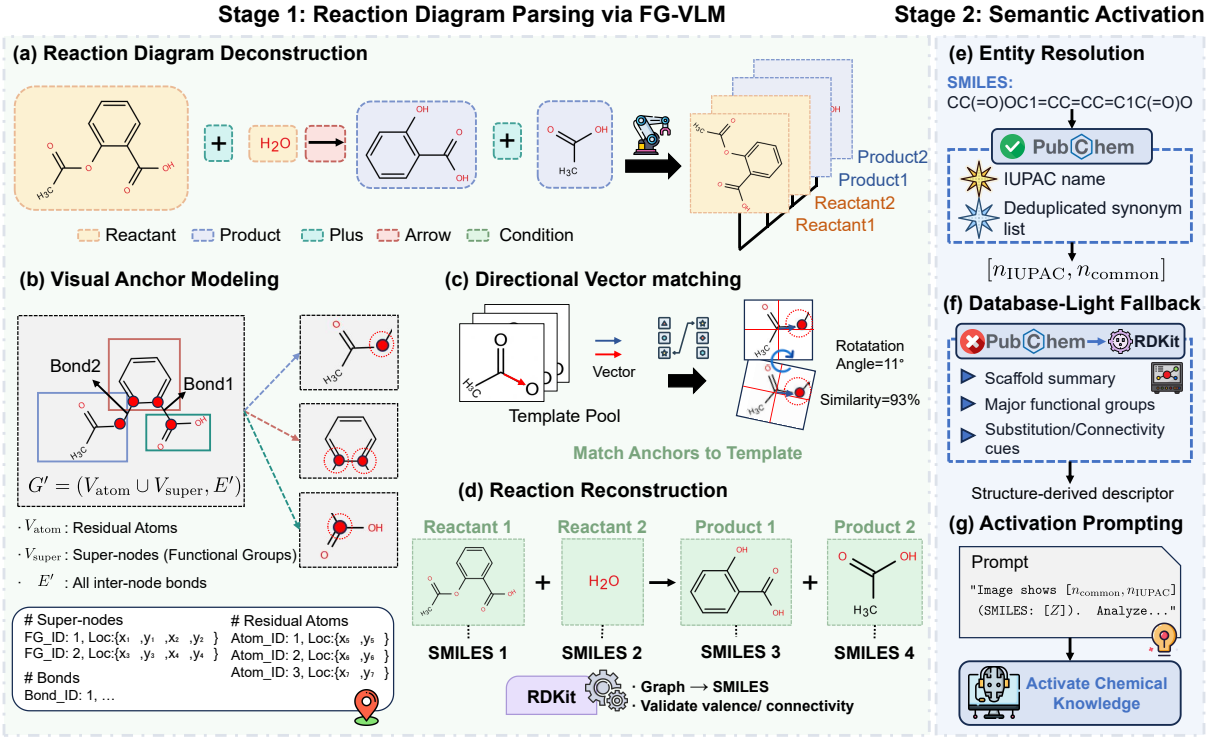


Figure 2: Overview of the ChemVA framework. **Stage 1: Reaction Diagram Parsing.** (a) **Reaction Diagram Deconstruction**: FG-VLM localizes diagram elements (*e.g.*, reactants, conditions, arrows) and crops molecular regions. (b) **Visual Anchor Modeling**: FG-VLM predicts a hybrid-granularity graph (functional groups and residual atoms) with bonds, and regresses group attachment anchors as pixel keypoints. (c) **Directional Vector Matching**: Group templates are oriented and attached to anchors to instantiate a topology-faithful molecular graph. (d) **Reaction Reconstruction**: The recovered graphs are converted to SMILES and parsed into a structured reaction sequence. **Stage 2: Semantic Activation.** (e) **Entity Resolution**: SMILES are resolved to IUPAC/common names via PubChem. (f) **Database-Light Fallback**: When PubChem matching fails, RDKit-derived structural descriptors are used as lightweight semantic representations. (g) **Activation Prompting**: The names are used to construct semantically grounded (name-first) prompts that more reliably elicit chemical knowledge for downstream reasoning.

box; each residual atom a is represented as $(id_a, elem_a, \mathbf{b}_a)$. Each bond stores its endpoints and bond order, covering atom–atom, atom–super-node, and super-node–super-node connections.

Visual Anchors. Bounding boxes alone are insufficient for rigid functional groups, because the model must determine which internal atoms participate in external bonds. For each functional-group node g, FG-VLM predicts a visual anchor set $\mathcal{A}_g = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k_g)}\}$, where each anchor $\mathbf{x}^{(j)} = (x, y)$ is the pixel coordinate of an internal attachment atom. Terminal groups usually have one anchor, while bridging or scaffold groups may have multiple anchors. These anchors are later matched to canonical attachment atoms in group templates through Directional Vector Matching.

3.1.3 Structural Reconstruction via Directional Vector Matching

Given predicted nodes, bonds, and visual anchors, we reconstruct the full molecular graph by matching each visual anchor to a canonical attachment atom in the corresponding FG template.

Let T_{c_g} denote the 2D template of group class c_g, with candidate attachment directions $\mathbf{u}_{c_g}^{(i)}$. For a predicted anchor $\mathbf{x}^{(j)}$, we compute its direction from the predicted group-box center $\mathbf{c}_{box}$:

$$\mathbf{v}_{pred}^{(j)} = \mathbf{x}^{(j)} - \mathbf{c}_{box}. \tag{1}$$

We map each anchor j to a template index by maximizing cosine similarity:

$$m_j^* = \arg \max_{i \in \{1,\ldots,N_{c_g}\}} \left(\frac{\mathbf{v}_{pred}^{(j)} \cdot \mathbf{u}_{c_g}^{(i)}}{\|\mathbf{v}_{pred}^{(j)}\| \|\mathbf{u}_{c_g}^{(i)}\|} \right). \tag{2}$$

This converts continuous pose estimation into a discrete matching problem and naturally handles rotational symmetries in rigid groups. After matching, we instantiate the oriented group templates, attach incident bonds to the matched sites, and combine them with residual atoms to obtain a topology-faithful molecular graph.

3.1.4 Reaction Reconstruction

We convert the reconstructed graph into a RDKit(Landrum et al., 2013) molecule and export a canonical SMILES string, using RDKit sanitization to validate valence and connectivity. Finally, we assemble the global reaction sequence by organizing reactant, product, and condition regions according to the parsed diagram layout.

3.2 Semantic Activation

Although FG-VLM provides topology-faithful SMILES and structured reaction representations, SMILES alone can be a weak semantic interface for LLM reasoning. We therefore introduce **Semantic Activation**, which converts recognized structures into language-accessible chemical representations.

Entity Resolution. When a recognized SMILES Z can be resolved in PubChem (Kim et al., 2023), we retrieve its IUPAC name and selected common synonyms: $\mathcal{M}(Z) = \{n_{\text{IUPAC}}, n_{\text{common}}\}$. Common names often align with natural-language chemistry corpora, while IUPAC names provide standardized identifiers. We combine the resolved names with the original SMILES to construct semantically grounded prompts for downstream LLM reasoning.

Database-Light Fallback. When PubChem-resolved names are unavailable, we use a lightweight fallback that verbalizes the recognized structure itself. Specifically, we use RDKit as a local graph-processing toolkit to summarize the coarse scaffold, major functional groups, and key substitution/connectivity cues from FG-VLM outputs. The resulting structure-derived descriptor is used together with SMILES as a compact semantic representation, without querying an external database.

Activation Prompting. The final prompt combines the recognized structure, SMILES, and available semantic representation:

$$P_{act} = \texttt{"Image shows [}r(Z)\texttt{]} \quad \texttt{(SMILES: [}Z\texttt{]). Analyze..."} \tag{3}$$

where $r(Z)$ denotes either PubChem-resolved names or the structure-derived fallback descriptor. This formulation treats Semantic Activation as a language-facing interface for downstream reasoning rather than a modification to the LLM itself.

3.3 Training Procedure for FG-VLM

FG-VLM is trained on **FG-SFT**, a specialized instruction-tuning dataset for hybrid-granularity recognition.

3.3.1 Dataset Construction

FG-SFT contains 100k molecule samples and 10k reaction samples; detailed statistics and source distributions are provided in (Appendix A2).

Molecule-Anchor Subset. We source unique molecular structures from PubChem (Kim et al., 2023). Starting from each `.mol` file, we construct the atom–bond graph with bond orders. We then apply a priority-driven greedy decomposition (Appendix A2) to partition atoms into functional-group instances $\{FG_i\}$; atoms not assigned to any FG_i are treated as residual atoms. This yields the hybrid-granularity node set (functional-group nodes + residual-atom nodes) together with the bond set inherited from the `.mol` connectivity.

To derive visual anchors, we inspect cross-boundary bonds between the atom set of FG_i and atoms outside FG_i (either residual atoms or atoms in other functional groups). The endpoint atoms inside FG_i are defined as attachment atoms and used as anchors.

Reaction-Layout Subset. We source chemically valid reaction records from the Open Reaction Database (ORD) (Kearnes et al., 2021) and synthesize diagrams in three literature-style layouts: *Linear Flow* (70%), *Multi-line* (15%), and *Tree/-Graph* (15%), with all constituents automatically annotated by bounding boxes.

3.3.2 Training Details

We fine-tune a single FG-VLM based on Qwen2.5-VL-32B (Bai et al., 2025) in a unified instruction-following formulation. Each example consists of a visual input $X \in \{I, I_{\text{crop}}, I_g\}$, a task instruction, and a structured target Y. FG-VLM is trained with three task formats: *Reaction Deconstruction* predicts diagram constituents and boxes, *Hybrid Structure Prediction* predicts the hybrid graph sequence $S = S_V \oplus S_E$, and *Anchor Prediction* predicts the anchor set $\mathcal{A}_g$ for each functional group. We use

Table 1: **Main Results on OCRD-Bench.** Comparison between the Baseline (Direct Input) and our ChemVA Framework (Ours). **Rec**: Recognition, **Know**: Knowledge, **Reas**: Reasoning. ChemVA shows consistent double-digit improvements across all LLMs.

Models	Vanilla LLM				LLM+ChemVA				Gain
	Rec	Know	Reas	**Total**	Rec	Know	Reas	**Total**	
Proprietary SOTA Models									
Gemini 2.5 Pro	68.0	59.0	54.0	58.80	92.0	74.0	71.0	**76.40**	+17.60
GPT-5 (Preview)	64.0	53.0	50.5	54.20	92.0	69.0	66.2	**72.48**	+18.28
Qwen3-Max	61.0	50.0	48.0	51.40	92.0	70.0	65.7	**72.68**	+21.28
GPT-4o	45.0	48.0	46.5	46.80	92.0	67.0	62.5	**70.20**	+23.40
Claude 4 Sonnet	49.0	47.0	45.2	46.68	92.0	66.0	62.5	**69.80**	+23.12
Open-Weights									
Yi-Lightning	47.0	41.0	38.7	41.28	92.0	58.0	54.7	**63.48**	+22.20
Llama-4-109B	44.0	38.5	36.7	38.88	92.0	55.0	52.0	**61.20**	+22.32
Intern-S1	54.0	43.5	41.2	44.68	92.0	56.5	53.5	**62.40**	+17.72
ChemVLM	36.0	27.5	25.7	28.48	92.0	32.5	30.5	**43.60**	+15.12

a curriculum-style schedule that first learns hybrid graph prediction, then adds anchor supervision, and finally mixes layout, graph, and anchor tasks for joint instruction tuning. During training, molecule and group crops are generated from ground-truth boxes; at inference, they are generated from FG-VLM predictions. Full objectives, task templates, sampling ratios, and hyperparameters are provided in Appendix A1.1.

4 Experiments

4.1 Experimental Setup

Implementation Details. We use Qwen2.5-VL-32B as the FG-VLM backbone and fine-tune it on FG-SFT for reaction deconstruction, functional-group grounding, and visual-anchor detection. For Semantic Activation, evaluation LLMs are queried with standardized name-first prompts using recognized SMILES and entity names. Hyperparameters (*e.g.*, learning rates, batch sizes) and the complete prompt templates are provided in Appendix.

Baselines. We compare ChemVA with proprietary and open-weight multimodal LLMs (Table 1).

Evaluation Dataset. **OCRD-Bench** is curated from graduate-level chemistry assessment materials and contains 100 reaction scenarios expanded into 500 hierarchical questions across eight reaction categories. It evaluates three tiers: *L1 (Visual Perception)*, *L2 (Chemical Knowledge)*, and *L3 (Mechanistic Reasoning)*, covering the process from diagram perception to mechanism explanation. Two chemistry PhD annotators independently verified the questions, reference answers, and scoring rubrics, with disagreements resolved through discussion.

Evaluation Metric. For L1, we report strict binary accuracy based on the Tanimoto similarity of Morgan fingerprints. For L2 and L3, we adopt a rubric-based LLM-as-a-Judge protocol using Gemini 2.5 Pro (Zheng et al., 2023), which assigns a fine-grained 0–100 score evaluating both answer correctness and explanation validity against reference answers. To validate judge reliability, two chemistry PhD annotators independently scored 100 sampled L3 responses on a coarser 0–3 ordinal scale (0: incorrect, 1: partially correct, 2: mostly correct, 3: fully correct). For cross-scale comparison, LLM 0–100 scores are binned into the same 0–3 scale (0–25, 26–50, 51–75, 76–100) before computing inter-rater agreement. Annotators achieved substantial agreement ($\kappa = 0.876$), and the LLM judge agreed well with both ($\kappa = 0.952/0.919$, Spearman $\rho = 0.974/0.881$), with all cases differing by at most one ordinal point. Metric definitions, judge prompts, binning details, score aggregation formulas and reliability statistics are provided in Appendix A3.

4.2 Main Results

Table 1 compares ChemVA with the Vanilla LLM baseline. The Total gain reported in Table 1 can be decomposed into two contributions: (i) the *recognition contribution* provided by FG-VLM, which standardizes the Recognition score to 92.0% across all backbones (see Appendix Table A2, which reports per-model results under the "SMILES Only" condition where FG-VLM provides recognition but

Table 2: Scale Robustness on Qwen2.5-VL. Vanilla LLM vs. ChemVA across different parameter scales within the same model family. ChemVA yields stable gains, indicating low sensitivity to backbone size.

Backbone	Vanilla LLM				LLM+ChemVA				Gain
	Rec	Know	Reas	**Total**	Rec	Know	Reas	**Total**	
Qwen2.5-VL-3B	11.0	9.0	3.1	7.04	92.0	23.0	11.6	32.24	+25.20
Qwen2.5-VL-7B	21.0	18.5	13.5	17.00	92.0	35.0	22.1	41.24	+24.24
Qwen2.5-VL-32B	39.0	28.5	24.7	29.08	92.0	44.5	39.7	52.08	+23.00
Qwen2.5-VL-72B	45.0	36.0	32.2	36.28	92.0	51.0	45.6	57.04	+20.76

Semantic Activation is disabled), and (ii) the *semantic activation contribution* from entity-name prompts, visualized as the green portion of each bar in Figure 3.

Overcoming the Visual Deficit. Direct-input baselines confirm the *Visual Deficit*: even strong proprietary models struggle with molecular topology. While Gemini-2.5-Pro achieves a recognition score of 68%, other large-scale models such as GPT-4o perform poorly, attaining only 45% accuracy in structural identification via direct input. Specialized models such as ChemVLM also exhibit limitations, with a Recognition score of 36%, which restricts their effectiveness in this domain. In contrast, our ChemVA framework explicitly models functional groups and visual anchors, elevating the Recognition score to 92% across all backbones. This structural correction provides a consistent and accurate foundation for downstream tasks.

Performance Leap in Knowledge and Reasoning. ChemVA also improves downstream reasoning across architectures. The gains are especially clear for open-weight models: Llama-4-109B improves from 38.88% to 61.20%, and Yi-Lightning from 41.28% to 63.48%. Crucially, this comprehensive improvement enables open-weights models to compete with proprietary direct-input systems. With the proposed framework, Llama-4-109B (61.2%) outperforms the direct input baseline of Gemini 2.5 Pro (58.8%). This empirically demonstrates that providing structured, high-quality inputs that incorporate both precise visual graphs and semantic entities is critical for realizing the reasoning potential of these models.

4.3 Robustness to Backbone Scale

To isolate the impact of backbone size, we conduct a controlled scaling study using the Qwen2.5-VL family, with parameter counts ranging from 3B to 72B. Each model is evaluated under two configurations: (i) *Vanilla LLM*, in which the model generates responses directly from raw reaction images via its native vision encoder; and (ii) *LLM+ChemVA*, in which the model relies on topology-faithful structures and entity names generated by FG-VLM and the Semantic Activation pipeline.

As presented in Table 2, ChemVA achieves a consistent absolute improvement ranging from +20.76 to +25.20 points in the Total score across all model sizes. Although the performance of the Direct Input baseline improves with scale (increasing from 7.04 to 36.28), it remains significantly constrained by visual perception capabilities. Conversely, ChemVA elevates the recognition score to 92.0% for all backbones and consistently enhances performance in both the Knowledge and Mechanistic Reasoning tasks. These results indicate that the primary bottleneck is unreliable visual grounding rather than insufficient linguistic capacity. Consequently, employing topology-faithful recognition combined with semantic activation offers an efficient alternative to simply increasing parameter scale.

4.4 Ablation Studies

To rigorously validate the effectiveness of our framework components, we conduct ablation studies focusing on two critical aspects: the impact of incorporating semantic entities (Semantic Activation) and the necessity of high-fidelity structure recognition (FG-VLM).

Bridging the Semantic Disconnect. For qualitative case studies and detailed performance metrics, please refer to Appendix A4. As illustrated in Figure 3, the results demonstrate that *Semantic Activation* serves as an effective solution to the semantic disconnect. By incorporating Entity Names, consistent performance gains are observed across the entire spectrum of models. This effect is particularly substantial in Open-Weights models, where improvements exceed 10%. For instance, the score of Llama-4-109B improves from a base-

Table 3: **Recognizer Substitution on OCRD-Bench (L1).** We compare several structure recognizers as drop-in replacements for FG-VLM.

Model Type	Model Name	OCRD-Bench (L1 only)
Transformer	RxnScribe	86%
	Decimer	83%
VLM	ChemDFM-X-13B	25%
	ChemVLM-26B	36%
Ours	**FG-VLM**	**92%**

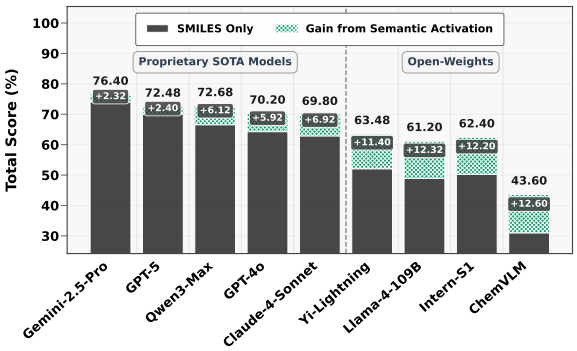


Figure 3: Effectiveness of Semantic Activation. The stacked bar chart illustrates the Total Scores across all models. The gray base represents the performance with *SMILES Only*, while the green top section highlights the Gain from Semantic Activation. The contrast clearly shows that gains are universal, with all models demonstrating significant improvement.

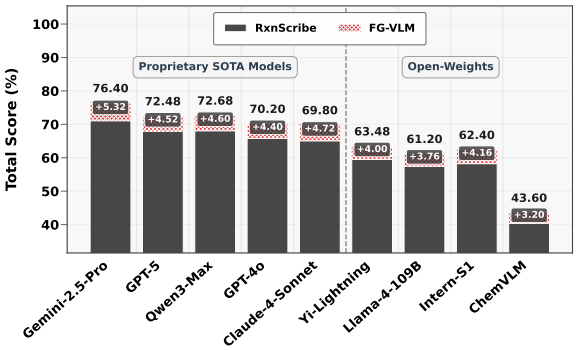


Figure 4: Superiority of FG-VLM over RxnScribe as the recognition frontend. The chart illustrates the Total Score comparison on OCRD-Bench across 9 models. The gray bars represent the performance using RxnScribe as the recognition tool, while the red sections indicate the explicit performance gains achieved by our FG-VLM.

line of 48.88% (see Appendix Table A2) to 61.2% (+12.32%), and that of Yi-Lightning increases from 52.08% to 63.48% (+11.40%). This consistent enhancement confirms that ChemVA effectively bridges the semantic gap, enabling diverse architectures to accurately align visual representations with latent chemical reasoning.

Replacing the Recognition Frontend. To further verify that the observed gains are driven by accurate structure recognition, the FG-VLM module is replaced with a range of off-the-shelf recognizers. These include transformer-based OCSR tools (*e.g.*, RxnScribe (Qian et al., 2023a) and DECIMER (Rajan et al., 2020, 2023)) and instruction-tuned chemical VLMs (*e.g.*, ChemVLM (Li et al., 2024)). Table 3 presents the L1 recognition accuracies on OCRD-Bench. While transformer-based OCSR baselines reach 83–86%, VLM-based recognizers exhibit severe hallucinations (25–36%), highlighting the limitations of generic encoders. In contrast, the proposed FG-VLM achieves 92%, surpassing all alternatives. Furthermore, the most competitive external baseline (RxnScribe) is integrated into the pipeline to evaluate end-to-end performance across all nine LLM backbones. As depicted in Figure 4 and Appendix Table A3, the recognition ceiling of RxnScribe (86%) constrains the final Total Scores (*e.g.*, Llama-4-109B: 57.44 with RxnScribe vs. 61.20 with FG-VLM), demonstrating that topology-faithful recognition is a prerequisite for reliable downstream reasoning.

5 Conclusion

In this work, we present a framework for chemical visual reasoning that bridges the critical gaps of visual deficit and semantic disconnect via hybrid-granularity anchoring and semantic activation. Built upon this framework, we develop a specialized visual module (FG-VLM) and a semantic alignment pipeline, capable of precise functional group grounding and improved interfacing between chemical diagrams and LLM reasoning. Moreover, we contribute two high-quality resources to the community: FG-SFT, a large-scale hybrid-granularity dataset for fine-grained visual instruction tuning, and OCRD-Bench, a challenging evaluation benchmark derived from graduate-level exams requiring end-to-end mechanistic reasoning. Experiments demonstrate that ChemVA substantially improves both structural recognition (reaching 92.0% accuracy) and downstream chemical reasoning (an average gain of ~20 points across 9 diverse LLMs), enabling open-weight models to approach, and sometimes surpass, strong proprietary direct-input baselines.

Limitations

Several limitations remain. First, the *Visual Anchor* mechanism relies on a predefined functional-group vocabulary. Although an auxiliary coverage analysis in Appendix A1.1 suggests broad coverage of common molecular patterns, rare ionic motifs, atypical bonding patterns, or molecules dominated by unseen groups may require heavier atom-level fallback, reducing the robustness of anchor-based matching.

Second, the current evaluation mainly uses standardized reaction diagrams. Robustness to noisier literature figures, low-resolution scans, complex layouts, and hand-drawn structures remains an important direction for future work.

Acknowledgment

We use an AI assistant for paraphrasing support during the writing of the manuscript.

References

Nawaf Alampara, I. Mandal, P. Khetarpal, H. S. Grover, and 1 others. 2024. Macbench: A multimodal chemistry and materials science benchmark. In *NeurIPS 2024 Workshop AI for Materials*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Christoph Bartmann, Johannes Schimunek, Mykyta Ielanskyi, Philipp Seidl, Günter Klambauer, and Sohvi Luukkonen. 2026. Molecularіq: Characterizing chemical reasoning capabilities through symbolic verification on molecular graphs. *arXiv preprint arXiv:2601.15279*.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.

Andres M Bran, Theo A Neukomm, Daniel P Armstrong, Zlatko Jončev, and Philippe Schwaller. 2025. Chemical reasoning in llms unlocks steerable synthesis planning and reaction mechanism elucidation. *arXiv e-prints*, pages arXiv–2503.

Kexin Chen, Yuyang Du, Junyou Li, Hanqun Cao, Menghao Guo, Xilin Dang, Lanqing Li, Jiezhong Qiu, Guangyong Chen, and Pheng Ann Heng. 2025. Chemminer: A large language model agent system for chemical literature data mining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7595–7603.

Sehyun Chun, Jiye Kim, Ahra Jo, Yeonsik Jo, Seungyul Oh, Seungjun Lee, Kwangrok Ryoo, Jongmin Lee, Seung Hwan Kim, Byung Jun Kang, and 1 others. 2025. Molmole: Molecule mining from scientific literature. *arXiv preprint arXiv:2505.03777*.

Djork-Arné Clevert, Tuan Le, Robin Winter, and Floriane Montanari. 2021. Img2mol - accurate molecular structure estimation from images. *Chemical Science*, 12(42):14174–14181.

Dong-Xu Cui, Shi-Yu Long, Yi-Xuan Tang, Yue Zhao, and Qiao Li. 2025. Can reasoning power significantly improve the knowledge of large language models for chemistry?based on conversations with llms. *Journal of Chemical Information and Modeling*.

Yanyan Diao, Feng Hu, Zihao Shen, and Honglin Li. 2023. Macfrag: Segmenting large-scale molecules to obtain diverse fragments. *Bioinformatics*, 39.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

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

Fernand Gobet, Peter CR Lane, Steve Croker, Peter CH Cheng, Gary Jones, Iain Oliver, and Julian M Pine. 2001. Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, 5(6):236–243.

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Stephen R Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. 2015. Inchi, the iupac international chemical identifier. *Journal of cheminformatics*, 7(1):23.

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. 2021. The open reaction database. *Journal of the American Chemical Society*, 143(45):18820–18826.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, and 1 others. 2023. Pubchem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380.

Greg Landrum and 1 others. 2013. Rdkit: Open-source cheminformatics. http://www.rdkit.org. Accessed: 2025-05-20.

Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, and 1 others. 2025. Chemvlm: Exploring the power of multimodal large language models in chemistry area. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 415–423.

Junxian Li, Di Zhang, Dongzhan Zhou, and 1 others. 2024. Chemvlm: Exploring the power of multimodal large language models in chemistry area. *arXiv preprint arXiv:2408.07246*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Lucas Morin, Martin Danelljan, Maria Isabel Agea, Ahmed Nassar, Valery Weber, Ingmar Meijer, Peter Staar, and Fisher Yu. 2023. Molgrapher: Graph-based visual recognition of chemical structures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19552–19561.

Martijn Oldenhof, Adam Arany, Yves Moreau, and Jaak Simm. 2021. Self-labeling of fully mediating representations by graph alignment. In *Benelux Conference on Artificial Intelligence*, pages 46–65. Springer.

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

Yujie Qian, Jiang Guo, Zhengkai Tu, Zhening Li, Connor W. Coley, and Regina Barzilay. 2023b. Molscribe: Robust molecular structure recognition with image-to-graph generation. *Journal of Chemical Information and Modeling*, 63(18):5833–5844.

K Rajan, H O Brinkhaus, M I Agea, A Zielesny, and C Steinbeck. 2023. Decimer.ai: An open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. *Nature communications*, 14(1):5045.

Kohulan Rajan, Achim Zielesny, and Christoph Steinbeck. 2020. Decimer: towards deep learning for chemical image recognition. *Journal of Cheminformatics*, 12(1):65.

Nicholas T Runcie, Charlotte M Deane, and Fergus Imrie. 2025. Assessing the chemical intelligence of large language models. *Journal of Chemical Information and Modeling*, 66(1):216–227.

Ayush Kumar Shah, Abhisek Dey, Leo Luo, Bryan Amador, Patrick Philippy, Ming Zhong, Siru Ouyang, David Mark Friday, David Bianchi, Nick Jackson, and 1 others. 2025. Multimodal search in chemical documents and reactions. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4030–4034.

Joshua Staker, Kyle Marshall, Robert Abel, and Carolyn M McQuaw. 2019. Molecular structure extraction from documents using deep learning. *Journal of chemical information and modeling*, 59(3):1017–1029.

Tuan Tran, Uyen Le, and Victor Phan. 2025. Evaluating the accuracy and educational potential of generative ai models in pharmacy education: A comparative analysis of chatgpt and gemini across bloom's taxonomy. *Pharmacy*, 14(1):1.

Jingchao Wang, Haote Yang, Jiang Wu, Yifan He, Xingjian Wei, Yinfan Wang, Chengjin Liu, Lingli Ge, Lijun Wu, Bin Wang, and 1 others. 2025. Gtr-cot: Graph traversal as visual chain of thought for molecular structure recognition. *arXiv preprint arXiv:2506.07553*.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models.

Damian M Wilary and Jacqueline M Cole. 2021. Reactiondataextractor: A tool for automated extraction of information from chemical reaction schemes. *Journal of chemical information and modeling*, 61(10):4962–4974.

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

Ziting Xian, Jiawei Gu, Lingbo Li, and Shangsong Liang. 2025. Molrag: unlocking the power of large language models for molecular property prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15531.

Qiang Xu, Shengyuan Bai, Yu Wang, He Cao, Leqing Chen, Yuanyuan Liu, Bin Feng, Zijing Liu, and Yu Li. 2026. Reactbench: A benchmark for topological reasoning in mllms on chemical reaction diagrams. *Preprint*, arXiv:2604.15994.

Ruiling Xu, Yifan Zhang, Qingyun Wang, Carl Edwards, and Heng Ji. 2025. omebench: Towards robust benchmarking of llms in organic mechanism elucidation and reasoning. arXiv preprint arXiv:2510.07731.

Zhaoning Yu, Xiangyang Xu, and Hongyang Gao. 2024. G2t-llm: Graph-to-tree text encoding for molecule

generation with fine-tuned large language models. *arXiv preprint arXiv:2410.02198*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, and 1 others. 2024a. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*.

Qiang Zhang, Keyang Ding, Tianwen Lyu, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, and 1 others. 2024b. Scientific large language models: A survey on biological and chemical domains. *arXiv preprint arXiv:2401.14656*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Jiaxi Zhuang, Kangning Li, Jue Hou, Mingjun Xu, Zhifeng Gao, and Hengxing Cai. 2025. Doc2sar: A synergistic framework for high-fidelity extraction of structure-activity relationships from scientific documents. *arXiv preprint arXiv:2506.21625*.

A1 Implementation Details: Prompts and Instruction Tuning

A1.1 Functional-Group Vocabulary Coverage Analysis

To better characterize the coverage of our predefined functional-group vocabulary, we conduct an auxiliary analysis on a large PubChem sample. We randomly sample 100,000 molecules from PubChem and obtain their SMILES strings. For each molecule, we compare two decomposition systems: our predefined vocabulary of 66 structural units and checkmol's broader set of 204 functional-group / structural-element categories. We find that 96.4% of sampled molecules show consistent decomposition between the two systems, suggesting that the 66-unit vocabulary captures the dominant structural organization of most common molecules.

We further analyze the relationship between the 204 checkmol categories and our 66 structural units. Specifically, for each checkmol category, we test whether it contains at least one of our predefined units as a substructure. The result shows that 98.0% of the checkmol categories contain at least one covered unit, indicating that many apparently unseen functional groups are not entirely novel visual patterns but are composed of substructures already represented in our vocabulary. The remaining unmatched cases mainly involve ionic categories, such as cations and anions, and a small number of special nitrogen-centered bonding patterns.

To estimate the effect of fallback-heavy cases, we use the number of isolated atom-level residues in each molecule as a proxy for fallback intensity. With a threshold of five isolated atoms, molecules with at most five isolated atoms achieve 36/37 recognition accuracy, whereas molecules with more than five isolated atoms achieve 56/63 accuracy. This confirms that recognition performance decreases when the model relies more heavily on atom-level fallback. However, the degradation is moderate rather than catastrophic, since covered substructures can still be recognized at the functional-group level while only uncovered motifs fall back to atom-level prediction.

Overall, this analysis suggests that the current vocabulary provides broad practical coverage for common molecular structures, while molecules dominated by rare ionic motifs, atypical bonding patterns, or heavily uncovered groups remain challenging.

A1.2 Reaction Diagram Deconstruction Prompt

To initiate the pipeline, we employ the following prompt to decompose the raw reaction image into semantic components. Beyond simple object detection, this prompt enforces a **logic-aware grouping mechanism** via the `reaction_id` field. This design is critical for handling complex scenarios such as multi-step synthesis (*e.g.*, $A \rightarrow B \rightarrow C$) or disconnected reaction variants, ensuring that shared intermediates are correctly associated with their respective reaction steps. The structured JSON output containing bounding boxes is subsequently used to crop individual molecular images for fine-grained recognition.

Reaction Parsing Prompt

System Prompt
You are an expert chemical reaction diagram analysis assistant. Your task is to analyze chemical reaction diagrams and parse them

into structured component lists. You must accurately detect all visual elements, classify their roles, and group them by reaction pathways. Handle complex layouts including linear flows, multi-step synthesis, and divergent/convergent pathways. Respond strictly in the specified JSON format.

User Prompt
Task Description
Analyze the provided chemical reaction diagram image and extract all semantic components. For each detected object, provide: 1. **Reaction ID**: An integer identifier grouping components belonging to the same reaction step. If multiple distinct reactions are present (*e.g.*, Step 1 and Step 2), assign distinct IDs. Shared components (*e.g.*, an intermediate serving as product of R1 and reactant of R2) should be listed for *each* reaction they participate in. 2. **Role**: Classify the object into one of: "Reactant", "Product", "Condition", "Arrow", "Plus". 3. **BBox**: The bounding box coordinates [x1, y1, x2, y2] normalized to 0-1000.
Parsing Rules
Role Definitions:
- *Reactant*: Starting materials or substrates appearing before the reaction arrow.
- *Product*: Final compounds or intermediates appearing after the reaction arrow.
- *Condition*: Text or small formulas above/below arrows (reagents, solvents, temperature, time).
- *Arrow*: The graphical arrow indicating reaction direction.
- *Plus*: The "+" sign separating multiple reactants or products.

Grouping Logic:
- A "Reaction" consists of a set of Reactants + Conditions + Arrow $\rightarrow$ Products. - In multi-step reactions ($A \rightarrow B \rightarrow C$), "B" is the Product of Reaction 1 and the Reactant of Reaction 2. - In disconnected layouts (multiple independent reactions in one image), treat them as separate Reaction IDs.
Output Format (STRICT)
Respond with a valid JSON list of objects. Do not include markdown code blocks (“‘json). [{"reaction_id": 1, "role": "Reactant", "bbox": [x1, y1, x2, y2]}, {"reaction_id": 1, "role": "Arrow", "bbox": [x1, y1, x2, y2]}, {"reaction_id": 1, "role": "Condition", "bbox": [x1, y1, x2, y2]}, {"reaction_id": 1, "role": "Product", "bbox": [x1, y1, x2, y2]}, {"reaction_id": 2, "role": "Reactant", "bbox": [x1, y1, x2, y2]}, ...]
Input Image
{input_image}

Now begin your structured response:

A1.3 Hybrid-Granularity Structure Recognition Prompt

This prompt drives the core hybrid-granularity recognition. The prompt enforces a **Top-Down Visual Perception** strategy. We instruct the model to prioritize the detection of holistic functional group patterns based on the provided Priority List , treating atoms merely as residual elements to fill the gaps. This "Chunking" strategy ensures topological robustness against visual noise.

Hybrid-Granularity Structure Recognition

System Prompt
You are an expert chemical structure recognition assistant. Your task is to convert a cropped molecule image into a **Hybrid-Granularity Graph**. You must employ a **Top-Down Visual Perception** strategy: 1. First, identify large, rigid **Super-nodes** (Functional Groups) based on global visual patterns. 2. Second, identify **Isolated Atoms** only in the remaining empty regions. 3. Finally, determine the connectivity (**Bonds**) between all identified nodes.
User Prompt
Task Description
Analyze the molecular image and generate a structured JSON representation containing `supernodes`, `atoms`, and `bonds`.
Decomposition Constraints (CRITICAL)
1. **Visual Priority (Top-Down)**: - *Rule*: Prioritize the detection of Functional Group Patterns (from {priority_list_file}) over individual atoms. - *Strategy*: Treat the image primarily as a composition of Functional Groups. Identify these rigid structures based on their global visual features. - *Conflict Resolution*: If a visual region matches a Functional Group template, you MUST output it as a Super-node. Do NOT

decompose it into internal atoms.
2. **Reference Priority List**: Strict adherence to the provided file: **{priority_list_file}**. - *Greedy Matching*: Scan and match functional groups strictly in the order they appear in the file. - *Closed-Set Constraint*: Only recognize Super-nodes explicitly listed in {priority_list_file}. If a substructure is not in the list, treat it as distinct Isolated Atoms.
3. **Residual Atom Identification**: - *Rule*: Only scan for **Isolated Atoms** in the image regions that are **NOT** covered by the bounding boxes of identified Super-nodes. - *Logic*: Atoms are "residual" elements. If an atom is visually encapsulated within a Super-node's region, it is implicitly handled by the Super-node template and must not be listed here.
4. **Super-node Exclusivity (No Overlap)**: - *Rule*: Super-nodes must NOT share any common atoms. - *Logic*: While bounding boxes of adjacent groups may spatially overlap due to crowding, the underlying atomic membership is mutually exclusive. No single atom can belong to more than one Super-node. If ambiguity exists, assign ownership to the group appearing earlier in the Priority List. ### Output Schema Definitions ###
1. **SuperNodes**: - `id`: Unique identifier (*e.g.*, "fg_1"). - `label`: Must match a name from {priority_list_file}. - `bbox`: [x1, y1, x2, y2].
2. **Atoms** (Residual): - `id`: Unique identifier (*e.g.*, "atom_1"). - `symbol`: Element symbol (*e.g.*, "C", "N", "O"). - `bbox`: [x1, y1, x2, y2].
3. **Bonds**: - `source`: Node ID (Atom or SuperNode). - `target`: Node ID (Atom or SuperNode). - `type`: Bond order ("SINGLE", "DOUBLE", "TRIPLE", "AROMATIC", "WEDGE", "DASH"). *Note: Connects any pair of nodes (Atom-Atom, SuperNode-SuperNode, or Atom-SuperNode).*
Output Format (STRICT)
Respond with a single valid JSON object.

```
{
"supernodes": [ ... ],
"atoms": [ ... ],
"bonds": [ ... ]
}
```

Input Image
{cropped_molecule_image}

Functional Group Priority List (File X)
{priority_list_content}

Now begin your analysis, prioritizing visual patterns from the list:

A1.4 Fine-Grained Visual Anchor Prediction Prompt

While the Hybrid-Granularity Recognition stage (Section 3.1.2) establishes the topological connectivity, it lacks the precise geometric alignment required for reconstructing the chemical drawing. To bridge this gap, we employ the following Visual Anchor Prediction prompt, which queries the VLM for the precise pixel coordinates of attachment atoms within each detected functional group. This prompt takes the bounding box and class label of a detected super-node as input and queries the VLM for the precise coordinates of its **Visual Anchors**—the specific atoms within the group that serve as attachment points to the external skeleton. These coordinates are the prerequisites for the subsequent *Directional Vector Matching* algorithm.

Prompt : Visual Anchor Prediction

System Prompt
You are a fine-grained visual grounding assistant for chemical structures. Your task is to identify the precise "Visual Anchors" within a specific functional group. **Definition**: A Visual Anchor is an atom *belonging to the functional group* that forms a chemical bond with an atom *outside* the group. - **Terminal Groups** (*e.g.*, -OH, -Cl): Typically have 1 anchor. - **Linker Groups** (*e.g.*, Phenyl, Pyridine): Can have 2 or more anchors depending on the substitution pattern (para-, meta-, ortho-).
User Prompt
Task Description
Focus exclusively on the functional group located at the bounding box: **{target_bbox}**. The class of this group is: **{target_label}**.
Your goal is to find the attachment points to the external molecule skeleton. 1. Observe the specified region. 2. Identify the specific atom(s) inside this group that connect to external bonds. 3. Return the center coordinates [x, y] of these anchor atoms (normalized 0-1000).

Output Format (STRICT)
Respond with a JSON object containing a single list of coordinates.
{
"anchors": [
[x1, y1],
[x2, y2]
]
}
Input Image
{cropped_molecule_image}

Now, locate the anchors for the **{target_label}** at **{target_bbox}**:

A2 Data Construction Details

A2.1 Image Synthesis Configuration

Standard datasets typically offer binary modalities: either high-level text captions (*e.g.*, molecule names) or low-level dense coordinates (*e.g.*, pure ".mol" files). Neither format is sufficient to teach a model the expert-like cognitive process of identifying functional groups.

To address this, we constructed **FG-SFT**, a specialized dataset designed to instill hybrid-granularity perception into the model. Unlike simple Image-SMILES pairs, FG-SFT provides a structured annotation layer that sits between raw atoms and semantic entities. It comprises 100k Molecule samples and 10k Reaction samples, where each entry explicitly maps visual regions to functional group and defines their specific *attachment points* (anchors). This unique format compels the model to learn both the semantic identity of chemical substructures and their precise geometric connectivity within the global reaction context. To ensure visual robustness, we synthesized images using 8 distinct templates, varying parameters such as bond thickness, font style, and atom label visibility. For reaction diagrams, we leveraged the Open Reaction Database (ORD) (Kearnes et al., 2021).

FG-SFT COCO-style Annotation(Lin et al., 2014) (Example)


```
{
  "images": [
    { "id": 0, "width": 438, "height": 149 },
    { "id": 1, "width": 94, "height": 94 },
    { "id": 2, "width": 26, "height": 22 },
    { "id": 3, "width": 82, "height": 66 },
    { "id": 4, "width": 45, "height": 43 }
```


Table A1: **Statistics of the FG-SFT Training Dataset.** The dataset is balanced across molecular recognition and reaction reasoning tasks.

Data Subset	Source	Sample Count
Reaction-Linear	†	7,000
Reaction-MultiLine	†	1,500
Reaction-Graph	†	1,500
Molecule	‡	100,000

Note: †: Sourced from ORD/Open-Source Datasets. ‡: Sourced from PubChem/Open-Source Datasets.

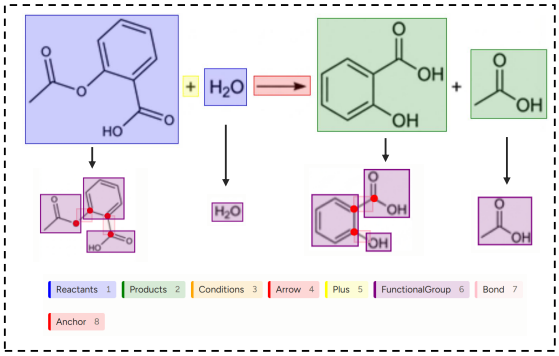


Figure A1: **FG-SFT example.** A raw reaction diagram with our FG-SFT annotations, including component-level boxes (Reactants, Products, Conditions, Arrow, Plus) and fine-grained labels (functional-group boxes, bond boxes, and anchor keypoints).


```
  ],
  "categories": [
    { "id": 0, "name": "Arrow" },
    { "id": 1, "name": "Conditions" },
    { "id": 2, "name": "FunctionalGroup" },
    { "id": 3, "name": "Plus" },
    { "id": 4, "name": "Products" },
    { "id": 5, "name": "Reactants" },
    { "id": 6, "name": "Bond" },
    { "id": 7, "name": "Anchor" }
  ],
  "annotations": [
    { "id": 0, "image_id": 0, "category_id":
        5, "bbox": [9.842696426713314,
        24.11469943321664, 123.0337081689187,
         102.85602809804595] },
    { "id": 1, "image_id": 0, "category_id":
        5, "bbox": [154.03820224719104,
        66.93033707865169, 33.95730337078655,
         29.528089887640455] },
    { "id": 2, "image_id": 0, "category_id":
        4, "bbox": [245.0831458649156,
        26.575280898876393,
        103.84044964070257,
        90.55278370922488] },
    { "id": 3, "image_id": 0, "category_id":
        0, "bbox": [193.90112359550562,
        72.83595505617978,
        45.276404494382035,
        15.256179775280895] },
    { "id": 4, "image_id": 0, "category_id":
        4, "bbox": [368.6089887640451,
        51.67419100942299,
```

```
        61.024719101123544,
        55.119067417543256] },
    { "id": 5, "image_id": 0, "category_id":
        3, "bbox": [136.81348314606745,
        72.83595505617978,
        12.303370786516838,
        16.240449438202244] },

    { "id": 6, "image_id": 1, "category_id":
        2, "bbox": [42.772914430818844,
        11.654662055488608,
        38.71857656274783,
        39.11501050858761] },
    { "id": 7, "image_id": 1, "category_id":
        2, "bbox": [1.617676919769829,
        27.48886108930232,
        38.586436511857926,
        36.339972887709216] },
    { "id": 8, "image_id": 1, "category_id":
        2, "bbox": [46.158279007543406,
        62.07354685827545, 45.59014143850223,
         20.218239461161858] },
    { "id": 9, "image_id": 1, "category_id":
        6, "bbox": [36.22133942935225,
        41.036336247198605,
        14.663592154970253,
        12.256130208981267] },
    { "id": 10, "image_id": 1, "category_id":
         6, "bbox": [61.60886594821314,
        47.71122806061365,
        10.942943747285096,
        17.946525800088715] },
    { "id": 11, "image_id": 1, "category_id":
         7, "bbox": [69, 65, 0, 0], "
        keypoints": [69, 65, 2] },

    { "id": 12, "image_id": 2, "category_id":
         2, "bbox": [1.4982595024672898,
        4.630889830131542, 24.50174049753269,
         14.664484462083207] },

    { "id": 13, "image_id": 3, "category_id":
         2, "bbox": [43.831353712647655,
        1.6233789731164145,
        35.714393151953395,
        37.74365447296713] },
    { "id": 14, "image_id": 3, "category_id":
         2, "bbox": [45.454723666369084,
        51.94825561870752, 19.48059585701531,
         14.051744381292487] },
    { "id": 15, "image_id": 3, "category_id":
         2, "bbox": [2.4350940336716946,
        21.509836637025675,
        36.52609768035893,
        40.58456165346862] },
    { "id": 16, "image_id": 3, "category_id":
         6, "bbox": [35.53113426627757,
        23.018826070221383,
        14.995638526775984,
        11.748175672830737] },
    { "id": 17, "image_id": 3, "category_id":
         6, "bbox": [35.91312944139563,
        49.95370849328054, 12.51232708738311,
         9.073811275111346] },
    { "id": 18, "image_id": 3, "category_id":
         7, "bbox": [52, 24, 0, 0], "
        keypoints": [52, 24, 2] },

    { "id": 19, "image_id": 4, "category_id":
         2, "bbox": [3.0908032596041917,
        2.409778812572759, 41.90919674039581,
         36.69455735869402] }
  ],
  "relations": []
}
```


A2.2 Priority-driven Greedy Decomposition Protocol

We employ a specialized decomposition algorithm to resolve semantic ambiguity in functional group identification. The protocol consists of three specific steps:

1. Priority Hierarchy Definition. We constructed a hierarchical dictionary where functional groups are ranked by heavy atom count, topological complexity, and semantic weight. High-complexity groups (*e.g.*, Carboxyl $-$COOH, Amide $-CONH_2$) are assigned higher matching priority than their constituents (*e.g.*, Carbonyl $C = O$, Hydroxyl $-OH$).

2. Recursive Matching with Exclusivity. For each SMILES string, we perform recursive substructure matching using RDKit (Landrum et al., 2013). Crucially, we enforce an *Atom-wise Exclusivity Constraint*: once an atom is assigned to a high-priority "super-node" (*e.g.*, the Carbon in $-$COOH), it is locked and explicitly excluded from subsequent scans. This prevents the redundant labeling of a carboxyl carbon as a ketone, ensuring a bijective mapping between visual anchors and semantic tokens.

3. Residual Atom Handling. After the greedy matching process, any remaining atoms (typically satisfying the saturated alkane skeleton) are retained as atomic tokens. This *hybrid-granularity* approach ensures that the model captures both high-level functional semantics and low-level structural details. We then calculate the 2D bounding box and the precise anchor coordinates (x_{anc}, y_{anc}) for each identified group.

priority_list_file:

1. Quaternary Ammonium
2. Carboxylate
3. Boronic Acid
4. Boronate Ester
5. Carboxylic Anhydride
6. Orthoester
7. Carbonate Ester
8. Lactone
9. Acetal/Ketal
10. Hemiacetal/Hemiketal
11. Epoxide
12. Imide
13. Lactam
14. Carbamate
15. Amidine
16. Carboxylic Acid
17. Ester
18. Amide
19. Acyl Halide
20. Carbothioic O-Acid
21. Carbothioic S-Acid
22. Carbodithioic Acid
23. Thiolester
24. Thionoester
25. Isocyanate
26. Isothiocyanate
27. Cyanate
28. Thiocyanate
29. Nitrile / Cyano
30. Isonitrile
31. Nitro
32. Nitroso
33. Oxime
34. Azide
35. Azo
36. Hydrazine
37. Primary Amine
38. Secondary Amine
39. Tertiary Amine
40. Imine
41. Sulfo (Sulfonic Acid)
42. Sulfonate
43. Sulfonyl
44. Sulfino
45. Disulfide
46. Thiol/Sulfhydryl
47. Sulfide(Thioether)
48. Sulfinyl(Sulfoxide)
49. Aldehyde
50. Acetyl
51. Ketone/Carbonyl
52. Phenol
53. Alcohol/Hydroxyl
54. Hydroperoxy
55. Peroxy
56. Ether
57. Methylenedioxy
58. Aryl Halide
59. Alkyl Halide
60. Alkynyl
61. Alkenyl
62. Pyridyl
63. Imidazole
64. Aryl
65. Carbonyl
66. Halo

A3 OCRD-Bench Framework: Design and Metrics

To comprehensively evaluate Multimodal Large Language Models (MLLMs) on organic chemistry reasoning, we designed **OCRD-Bench**, a hierarchical benchmark covering 8 major reaction categories (see Figure A2). The evaluation is structured into three cognitive tiers, ranging from visual perception to deep mechanistic reasoning. Here, we detail the task design and the corresponding metric protocol for each level.

A3.1 Level 1: Visual Perception (L1)

Task Definition (Q1–Q2). This tier tests the fundamental ability to ground visual features to symbolic representations. Taking *Friedel-Crafts Alkylation* as an example, the model is prompted: *"Identify the substrate (Benzene) and reagent (Chloroethane) in the image and encode them into SMILES."*

Metric: Strict Structural Equivalence. Standard text-generation metrics (e.g., Levenshtein distance) fail to capture chemical validity. Instead, we utilize RDKit to generate Morgan Fingerprints (radius=2, 2048 bits) for both the predicted (P) and ground-truth (G) SMILES. The score is calculated as a strict binary metric:

$$\mathrm{S}_{L1} = \mathbb{1}(\text{Tanimoto}(P, G) = 1.0) \times 100 \quad \text{(A1)}$$

where $\mathbb{1}$ is the indicator function. This strict criterion ensures that any hallucinated atom, incorrect bond type, or stereochemical error results in a zero score, prioritizing topological fidelity over partial character matches.

A3.2 Level 2: Chemical Knowledge (L2)

Task Definition (Q3–Q4). This tier assesses the retrieval of static knowledge regarding reactivity, selectivity (regio-/chemo-), and reaction conditions. Example task: *"Compare the reactivity of chloroethane vs. chlorobenzene in this reaction system. Explain the scope and limitations regarding carbocation rearrangement."*

Metric: LLM-as-a-Judge (Knowledge with Rationale). While L2 tasks follow a single-choice or multiple-choice format, we do not rely solely on option accuracy to prevent "lucky guesses." We employ a strong proprietary model (Gemini 2.5 Pro) as an impartial judge to assess two components:

- **Option Correctness:** Whether the selected answer matches the key.
- **Rationale Validity:** Whether the generated explanation logically derives the answer using correct chemical principles.

A high score is awarded only when both the choice and the reasoning are correct.

A3.3 Level 3: Mechanistic Reasoning (L3)

Task Definition (Q5). This tier tests deep reasoning capabilities. The model must generate a valid causal chain of intermediates rather than simply hallucinating the final product. Example task: *"Provide the complete reaction mechanism, describing the electron flow and intermediate structures using SMILES sequences."*

Metric: LLM-as-a-Judge (Completeness & Consistency). For open-ended mechanism generation, the judge compares the model's response against the standard exam key based on:

- **Completeness:** Coverage of all key intermediates (e.g., formation of the Wheland intermediate) and electron flow steps.
- **Consistency:** Logical flow and absence of chemically impossible steps.

The judge assigns a fine-grained scalar score (0–100) reflecting the semantic alignment with the ground truth mechanism, enabling discrimination

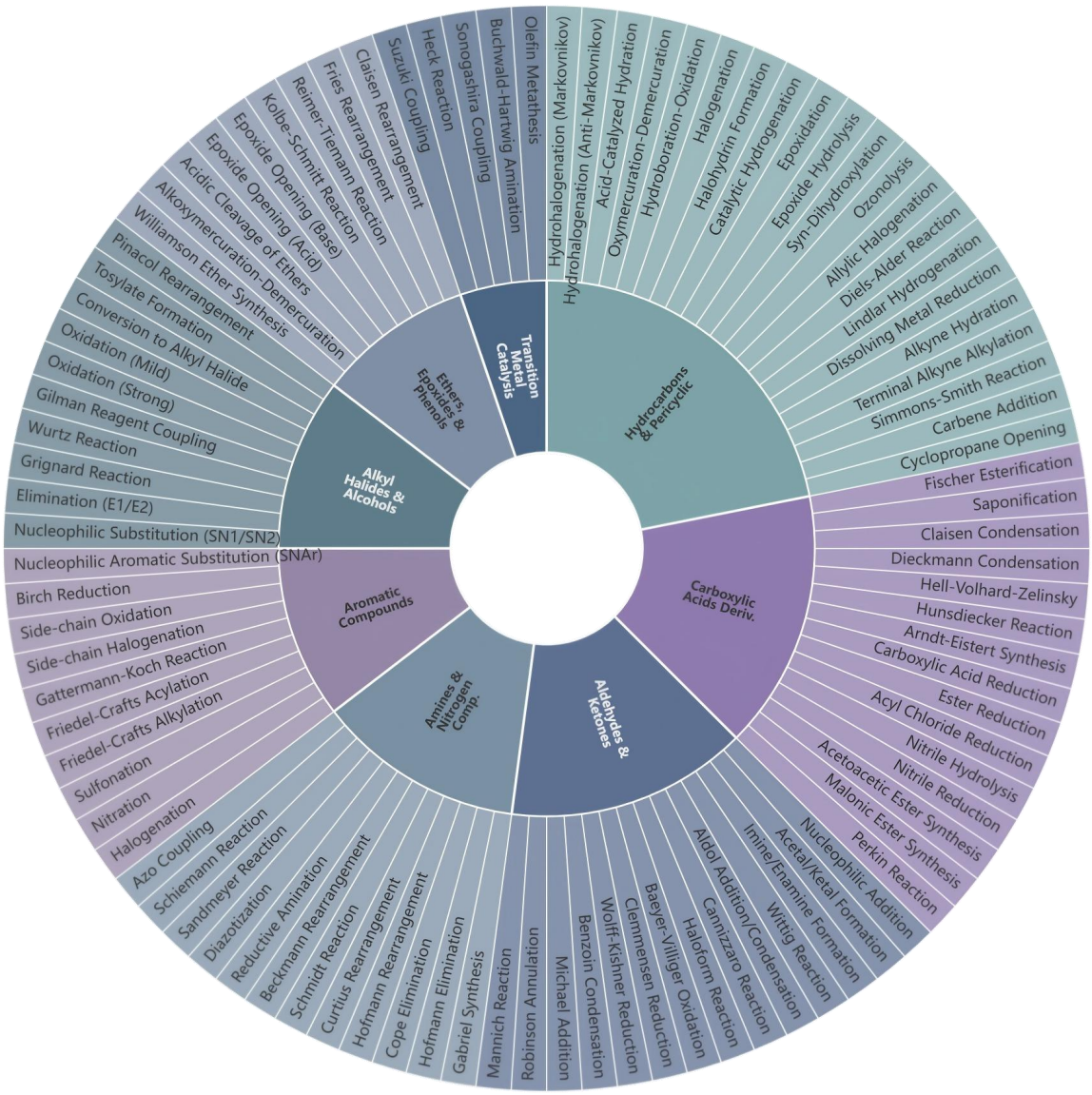


Figure A2: **Taxonomy of OCRD-Bench.** The benchmark spans 8 major categories of organic reactions, ranging from fundamental Hydrocarbons to complex Transition Metal Catalyzed Reactions. This diversity ensures a robust evaluation of model generalization across different chemical domains.

between models with similar but non-identical performance levels across the 500-question evaluation set.

Cross-Scale Alignment for Human Validation. While the LLM judge operates on the 0–100 scale described above for fine-grained ranking, the judge is prompted to output scores in increments of 10 (i.e., $\{0, 10, 20, \ldots, 100\}$), which provides 11 discrete quality levels–sufficient for discriminating between models while remaining stable across judge calls.Human reliability validation is performed on a coarser 0–3 ordinal scale, since human annotators cannot reliably distinguish 100 quality levels for complex chemical reasoning. The human rubric is defined as:

- **0 (Incorrect):** The mechanism contains fundamental chemical errors (e.g., wrong reaction type, impossible electron flow), or the response is irrelevant to the question.
- **1 (Partially correct):** The model identifies the reaction type correctly but misses major intermediates or contains significant mechanistic errors.
- **2 (Mostly correct):** The mechanism is largely valid with minor omissions or imprecisions (e.g., missing one intermediate, incomplete electron arrow description).
- **3 (Fully correct):** The mechanism covers all

key intermediates and electron flow steps with no chemical errors.

For cross-scale comparison, the LLM's 0–100 scores are binned into the 0–3 ordinal scale using a uniform mapping: $[0, 25] \rightarrow 0$, $(25, 50] \rightarrow 1$, $(50, 75] \rightarrow 2$, $(75, 100] \rightarrow 3$. All quadratic weighted κ statistics reported in Section 4 are computed on this binned ordinal scale, while Spearman ρ values are computed on the original continuous 0–100 scale to preserve fine-grained ranking information. Two chemistry PhD annotators independently scored the same 100-question subset; they achieved substantial agreement at $\kappa = 0.876$, and the LLM judge showed high agreement with both ($\kappa = 0.952/0.919$, $\rho = 0.974/0.881$), with all human-LLM disagreements bounded within one ordinal point.

A3.4 Detailed Performance Data

Table A2 provides the fine-grained metrics for the baseline setting, serving as the reference point for the gains illustrated in the main text.

A3.5 Score Aggregation

The Total Score reported aggregates the three evaluation tiers using a weighted average:

$$\text{Total} = w_{L1} \cdot S_{L1} + w_{L2} \cdot S_{L2} + w_{L3} \cdot S_{L3} \quad \text{(A2)}$$

where S_{L1}, S_{L2}, S_{L3} are the average scores on Level 1 (Visual Perception), Level 2 (Chemical Knowledge), and Level 3 (Mechanistic Reasoning) respectively. We set $w_{L1} = 0.2$, $w_{L2} = 0.4$, $w_{L3} = 0.4$ to reflect the increasing cognitive complexity from perception to reasoning: L1 contributes a smaller weight because it captures only foundational structural recognition, while L2 and L3 contribute equal weights as they together represent the chemical reasoning capabilities that are the primary focus of this work.

A4 Case Study

To demonstrate the efficacy of our method in handling complex chemical reasoning, we conduct a detailed qualitative analysis on a representative sample from the OCRD-Bench (Transition Metal Catalysis). We select the Suzuki-Miyaura Coupling reaction, a cornerstone of organic synthesis, which requires the model to navigate through a multi-step reasoning chain: from structural recognition to condition extraction, and finally to mechanistic inference.

Table A2: Detailed Breakdown of Baseline Performance (SMILES Only). This table lists the fine-grained metrics (Rec, Know, Reas) for the visual-only setting. The specific gains achieving the final Total Scores are visualized in Figure 3.

Model	**SMILES Only**			
	Rec	Know	Reas	**Total**
Proprietary SOTA Models				
Gemini 2.5 Pro	92.0	72.0	67.2	74.08
GPT-5 (Preview)	92.0	67.5	61.7	70.08
Qwen3-Max	92.0	61.0	59.4	66.56
GPT-4o	92.0	59.5	55.2	64.28
Claude 4 Sonnet	92.0	61.0	50.2	62.88
Open-Weights				
Yi-Lightning	92.00	44.0	40.2	52.08
Llama-4-109B	92.00	40.0	36.2	48.88
Intern-S1	92.00	41.5	38.0	50.20
ChemVLM	92.00	16.5	15.0	31.00

Through this case study, we aim to highlight two critical failure modes prevalent in traditional Multimodal Large Language Models (MLLMs):

- **Visual Deficit:** The inability to accurately resolve fine-grained structural details (e.g., bond orders, stereochemistry, and connectivity) in dense molecular diagrams.
- **Semantic Disconnect:** The representation-knowledge mismatch. Since LLMs acquire the vast majority of their chemical knowledge from textbooks—where entities are primarily referenced by names or aliases (e.g., "Aryl halide") rather than linear strings—SMILES often fails to effectively "awaken" (activate) the model's deep-seated domain knowledge.

A4.1 Visual Deficit in General-Purpose VLMs

To explicitly demonstrate the *Visual Deficit* discussed in Section 1, we present a qualitative comparison using a Suzuki-Miyaura coupling reaction sample from the OCRD-Bench. As shown in Figure A3, the input image contains both dense molecular graphs (reactants/products) and textual reaction conditions.

Vanilla LLM: Qwen3-Max

Context: *Single-step Reaction Parsing Prompt*

Q1, Q2
System Prompt:
You are an expert organic chemist and an advanced Optical Chemical Structure Recogni-

Table A3: Detailed Performance Benchmark with Rxnscribe. This table lists the fine-grained metrics (Rec, Know, Reas) when using the external SOTA tool (RxnScribe) as the visual frontend. Note that the Recognition score is standardized to the tool's performance ceiling (86.0%).

Model	RxnScribe			
	Rec	Know	Reas	**Total**
Proprietary SOTA Models				
Gemini 2.5 Pro	86.0	68.0	66.7	71.08
GPT-5 (Preview)	86.0	64.5	62.4	67.96
Qwen3-Max	86.0	66.0	61.2	68.08
GPT-4o	86.0	63.0	58.5	65.80
Claude 4 Sonnet	86.0	61.0	58.7	65.08
Open-Weights				
Yi-Lightning	86.0	54.5	51.2	59.48
Llama-4-109B	86.0	52.0	48.6	57.44
Intern-S1	86.0	52.5	50.1	58.24
ChemVLM	86.0	29.5	28.5	40.40

tion (OCSR) system. Your task is to accurately extract chemical information from reaction scheme images.

User Prompt:
Please analyze the attached image of a chemical reaction. Your goal is to identify the chemical structures and the text conditions, and extract them into a structured format.

Instructions:

- **Reactants:** Identify the molecules to the left of the reaction arrow. Convert their structures into SMILES. If there are multiple reactants, separate them with a period (.).
- **Products:** Identify the molecules to the right of the reaction arrow. Convert their structures into SMILES.
- **Conditions:** Read the text above and below the reaction arrow. Based on your chemical knowledge, categorize these substances into "Reagents" (including catalysts, bases, ligands, additives) and "Solvents".

Output Format: Please strictly follow this format (do not output markdown code blocks, just the raw text):

```
Reactants: [SMILES string]
Products: [SMILES string]
Reagents: [List of reagents, separated by comma]
Solvents: [List of solvents, separated by comma]
```

Answer:
```
Reactants: COC(=O)c1ccc(c2cc(Br)ccc2c1)c1ccccc1.OB(O)(O)c1ccccc1
Products: COC(=O)c1ccc(c2ccccc2c1)c1ccccc1
Reagents: Pd(OAc)2, PPh3, TBAB
Solvents: Toluene, H2O
```

FG-VLM Answer

Answer:
```
Reactants: COC(=O)c1ccc(-c2ccc(Br)cc2)cc1.OB(O)c1ccccc1
Products: COC(=O)c1ccc(-c2ccc(-c3cccc3)cc2)cc1
Reagents: Pd(OAc)2, PPh3, TBAB
Solvents: Toluene, H2O
```

The Illusion of Understanding. General-purpose VLMs, such as Qwen3-max, exhibit a deceptive capability profile. As observed in the detailed model outputs, Qwen3-max successfully retrieves the textual reaction conditions, correctly identifying the catalyst system ("$Pd(OAc)_2$, PPh_3") and solvents ("Toluene, H_2O"). This aligns with the model's strong OCR capabilities for standard text.

However, a critical failure occurs in the structural interpretation of the molecular components (Q1). The ground truth reactant is a biaryl bromide (*methyl 4'-(bromophenyl)-benzoate*), composed of two linearly connected benzene rings. In contrast, Qwen3-max generates a hallucinatory SMILES string:

```
COC(=O)c1ccc(c2cc(Br)ccc2c1)c1ccccc1
```

This output reveals a severe topological error: it incorrectly introduces an additional phenyl ring ("c1ccccc1") and disrupts the biaryl connectivity, effectively fabricating a molecule that does not exist in the image. This exemplifies the *Visual Deficit*: the model perceives the visual tokens but fails to resolve the strict graph topology, resulting in structural hallucination.

Resolving Connectivity via Visual Anchors. In contrast, our FG-VLM correctly reconstructs the reactant SMILES by leveraging the proposed *Hybrid-Granularity Visual Anchor Modeling*. Instead of treating the molecule as a sequence of atomic characters, FG-VLM identifies the rigid functional groups (e.g., the phenyl rings and ester group) as stable visual anchors. By explicitly modeling the directional vectors between these anchors, our method enforces topological fidelity, successfully recognizing the correct two-ring system and the precise position of the bromine substituent. This case underscores that bridging the visual deficit requires more than scaling generic vision encoders; it demands domain-specific grounding of chemical substructures.

A4.2 Beyond Correct Answers: Grounding Knowledge Reasoning in Visual Reality

While quantitative metrics (Accuracy) indicate the final performance, a qualitative inspection of the reasoning process reveals the profound impact of our ChemVA framework in bridging the *Semantic Disconnect*. We analyze two representative knowledge tasks (Q3 and Q4) to demonstrate the difference between *parametric retrieval* and *visually grounded reasoning*.

Know Task

Context: *Prompt with Knowledge Injection*

System Prompt:
You are an expert organic chemist and academic researcher. Your task is to analyze chemical reaction images and answer reasoning questions based on your domain knowledge of reaction mechanisms, reagents, and catalytic cycles.
CRITICAL OUTPUT RULES:

1. You must output the response STRICTLY in the following format.
2. Do not include any introductory text or concluding remarks.
3. For "answer": Output the single letter (e.g., A) or letters separated by commas.
4. For "explanation": Provide a concise but professional chemical justification.

Standard Output Format: `answer: [Option Letter(s)]`
`explanation: [Your detailed reasoning content]`

User Prompt:
Task: Analyze the provided chemical reaction image and the supplementary information, and then answer the following question.
Supplementary information:[]
Question:[]
Question Type:[]

SMILES Only
Supplementary Information (Input): `Reactant1:COC(=O)c1ccc(-c2ccc(Br)cc2)cc1,Reagents2:OB(O)c1ccccc1,Product1:COC(=O)c1ccc(-c2ccc(-c3ccccc3)cc2)cc1,Reagents:Pd(OAc)2,PPh3,TBAB;solvents:Toluene,H2O`
Q3
Model Output:

```
answer: B
explanation: The reaction
depicted is a Suzuki-Miyaura
cross-coupling between an
aryl bromide and
phenylboronic acid, catalyzed
by Pd(OAc)2 with PPh3 as
ligand and TBAB as
phase-transfer catalyst in a
toluene/water mixture.
Although the base (e.g.,
K2CO3) is not explicitly
shown in the diagram, it is
essential for the catalytic
cycle. The primary role of
the base is to deprotonate
the boronic acid, forming a
more nucleophilic boronate
species [R-B(OH)3]⁻, which
facilitates the
transmetallation step—where
the aryl group from the
palladium complex transfers
to the boronate, enabling
coupling. This step is
crucial for efficient
coupling. Option A is
incorrect because the
reduction of Pd(II) to Pd(0)
is typically achieved by the
ligands or solvent, not the
base. Option C is incorrect
because benzyne formation is
```

not involved; this is not a elimination-based mechanism. Option D is incorrect because the base actively participates in the transmetallation step, not merely neutralizing acid. Thus, B is correct.

Q4
Model Output:

answer: D
explanation: The reaction shown is a Suzuki-Miyaura cross-coupling, which proceeds via a Pd(0)/Pd(II) catalytic cycle. The correct chronological sequence of elementary steps is: (1) Oxidative Addition—where the Pd(0) species adds across the C–Br bond of the aryl bromide to form a Pd(II) complex; (2) Transmetallation—where the organoboron reagent (phenylboronic acid, activated by base) transfers its phenyl group to the palladium center, displacing the halide; and (3) Reductive Elimination—where the two aryl groups on palladium couple to form the biaryl product and regenerate the Pd(0) catalyst. This sequence is well-established in palladium-catalyzed cross-coupling reactions. Option A is incorrect because transmetallation cannot occur before oxidative addition, as no Pd–C bond exists yet. Option B reverses the order of transmetallation and reductive elimination, which is mechanistically invalid. Option C starts with reductive elimination, which requires prior formation of the Pd(II) intermediate, making it impossible as the first step. Thus, D is the only correct sequence.

ChemVA

Supplementary Information (Input): Reactant1:COC(=O)c1ccc(-c2ccc(Br)cc2)cc1,Synonyms:methyl4'-bromo[1,1'-biphenyl]-4-carboxylate,IUPACName:methyl4-(4-bromophenyl)benzoate,Reagents2:OB(O)c1ccccc1,Synonyms:Phenylboronic acid,IUPACName:phenylboronicacid,Product1:COC(=O)c1ccc(-c2ccc(-c3ccccc3)cc2)cc1,Synonyms:,IUPACName:methyl4-(4-phenylphenyl)benzoate,Reagents:Pd(OAc)2,PPh3,TBAB;solvents:Toluene,H2O

Q3
Model Output:

answer: B
explanation: The reaction shown is a Suzuki–Miyaura cross-coupling between methyl 4-(4-bromophenyl)benzoate and phenylboronic acid, catalyzed by $Pd(OAc)_2/PPh_3$ in the presence of TBAB, under biphasic conditions (toluene/H_2O). In this catalytic cycle, the base (e.g., K_2CO_3) plays a critical role during the transmetallation step. It deprotonates the boronic acid to form a more nucleophilic boronate species $[R\text{-}B(OH)_3]^-$, which facilitates the transfer of the organic group from boron to palladium. This step is essential because the neutral boronic acid is less reactive toward transmetallation with the Pd(II) or Pd(0) intermediate. Option A is incorrect because the reduction of Pd(II) to Pd(0) is typically facilitated by phosphine ligands or solvent-mediated processes, not the base. Option C is incorrect as benzyne formation requires strong bases and specific substrates (e.g.,

```
ortho-dihalides), which are
not present here. Option D is
incorrect because the base is
not merely a stoichiometric
neutralizer; it actively
participates in the mechanism
via coordination and
activation of the boronic
acid. Thus, the primary
mechanistic role of the base
is to generate the reactive
boronate complex, making
option B correct.
```

Q4
Model Output:

```
answer: D
explanation: The reaction
depicted is a Suzuki-Miyaura
cross-coupling between an
aryl bromide (methyl
4-(4-bromophenyl)benzoate)
and phenylboronic acid,
catalyzed by Pd(OAc)2 in the
presence of PPh3 and TBAB in
toluene/water. This
transformation proceeds via a
well-established Pd(0)/Pd(II)
catalytic cycle. The correct
chronological sequence of
elementary steps is: (1)
Oxidative Addition, where the
Pd(0) species adds across the
C-Br bond of the aryl halide
to form a Pd(II)
intermediate; (2)
Transmetallation, in which
the organoboron reagent
transfers its phenyl group to
the palladium center,
displacing the bromide and
forming a Pd(II)-aryl
complex; and (3) Reductive
Elimination, where the two
aryl groups coupled on the
Pd(II) center eliminate to
form the biaryl product and
regenerate Pd(0). Option D
correctly lists this
sequence: Oxidative Addition
→ Transmetallation →
Reductive Elimination.
Options A, B, and C are
incorrect because they
misorder these essential
mechanistic
steps—transmetallation
cannot occur before oxidative
addition, and reductive
elimination must follow
transmetallation to form the
new C-C bond.
```

Q3: Mechanistic Fidelity vs. Hallucinated Logic. In Question 3 (Role of the Base), both the baseline (SMILES Only) and ChemVA correctly selected Option B. However, a close examination of the generated explanations reveals a critical divergence in chemical understanding.

- The Baseline's Failure: Although the baseline selected the correct option, its explanation contains a fundamental mechanistic error. It states: *"the aryl group from the palladium complex transfers to the boronate."* This is chemically inverted; in Transmetallation, the organic group transfers *from* the activated boron species *to* the palladium center. This "Right Answer, Wrong Reason" phenomenon suggests the baseline relies on superficial keyword matching or hallucinated logic rather than a true understanding of the catalytic cycle.

- ChemVA's Precision: In contrast, ChemVA accurately describes the directionality of the group transfer (Boron → Palladium) and correctly identifies the nucleophilic activation process. This confirms that our Visual Anchor mechanism not only aids in structure recognition but also enforces logical consistency in downstream reasoning.

Q4: Instance-Level Awareness vs. Generic Templates. Question 4 asks for the sequence of the catalytic cycle. While the sequence is standard for all Suzuki couplings, the quality of the response depends on how well the model integrates the specific context of the input image.

- Generic Retrieval (Baseline): The baseline refers to the reactant simply as an "aryl bromide." While factually true, this generic label implies the model is retrieving a pre-

trained template for "Suzuki Reaction" without attending to the specific molecular details shown in the image.

- Contextual Grounding (ChemVA): ChemVA exhibits superior context awareness by explicitly naming the complex substrate: *"methyl 4-(4-bromophenyl)benzoate."* This precision stems directly from our Semantic Activation mechanism, which translates visual features into explicit textual anchors (*e.g., IUPAC names and synonyms*) before reasoning. It also correctly enumerates the specific reaction conditions ($Pd(OAc)_2$, PPh_3, TBAB) visible in the image. This proves that ChemVA is not merely reciting a textbook definition but is performing *instance-specific reasoning* grounded in the visual evidence extracted via the Semantic Activation module.

Reas Task

Context: *Open-Ended Reasoning Prompt*

System Prompt:
You are an expert organic chemist and synthetic consultant. Your task is to analyze chemical reaction images and answer advanced reasoning questions regarding substrate scope, reaction kinetics, and condition optimization.
CRITICAL OUTPUT RULES:

1. You must output the response STRICTLY in the following format.
2. Do not include any introductory text.
3. For "answer": Provide a concise, direct conclusion (1-2 sentences).
4. For "explanation": Provide a comprehensive analysis including Theoretical Basis, Mechanism Analysis, and Synthetic Strategy.

Standard Output Format: `answer: [Direct Conclusion/Prediction] explanation: [Detailed Mechanism Analysis and Synthetic Proposal]`

User Prompt:
Task: Analyze the provided chemical reaction image. Based on the specific reaction type and conditions shown, answer the following hypothetical reasoning question.
Question:[]
Supplementary information:[]
Question Type: Open-Ended Reasoning / Reaction Optimization

SMILES Only
Supplementary Information (Input): `Reactant1:COC(=O)c1ccc(-c2ccc(Br)cc2)cc1,Reagents2:OB(O)c1ccccc1,Product1:COC(=O)c1ccc(-c2ccc(-c3ccccc3)cc2)cc1,Reagents:Pd(OAc)2,PPh3,TBAB;solvents:Toluene,H2O`
Model Output:

```
answer: The reaction rate
would decrease significantly,
or the reaction might fail to
proceed under the original
mild conditions. Chemical
Reason: The C-Cl bond is
stronger than the C-Br bond.
Consequently, the oxidative
addition step—where the Pd(0)
catalyst inserts into the
carbon-halogen bond—becomes
much more difficult and acts
as the rate-determining step.

explanation:
Theoretical Basis: The bond
dissociation energy (BDE) of
a C-Cl bond (∼339 kJ/mol) is
significantly higher than
that of a C-Br bond (∼276
kJ/mol), making the cleavage
of the C-Cl bond in oxidative
addition less favorable
thermodynamically and
kinetically. This increased
bond strength results in a
higher activation energy
barrier for the oxidative
addition step, which is the
rate-determining step in
Suzuki-Miyaura coupling
reactions.

Mechanism Analysis: In the
Suzuki coupling, the Pd(0)
species undergoes oxidative
addition with the aryl halide.
For bromides, this step
proceeds readily due to the
weaker C-Br bond and better
polarizability. However,
```

when the bromine is replaced by chlorine, the stronger C–Cl bond resists oxidative addition, leading to slow or incomplete reaction. Additionally, the lower nucleophilicity and poorer leaving-group ability of chloride compared to bromide further hinder the process.

Synthetic Strategy: To overcome this challenge, electron-rich and sterically bulky phosphine ligands are essential. These ligands increase the electron density at the palladium center, enhancing its nucleophilicity and facilitating oxidative addition into the strong C–Cl bond. Specific examples include Buchwald-type ligands such as SPhos, XPhos, or RuPhos, which are known for their high activity in aryl chlorides. Alternatively, N-heterocyclic carbene (NHC) ligands like IPr or SIPr can be used, as they provide strong σ-donation and stability to the Pd(0) complex. The use of a more active catalyst system such as $Pd(OAc)_2$ with these advanced ligands, along with elevated temperatures or longer reaction times, may be required to achieve satisfactory yields.

ChemVA (Ours)

Supplementary Information (Input): `Reactant1:COC(=O)c1ccc(-c2ccc(Br)cc2)cc1,Synonyms:methyl4'-bromo[1,1'-biphenyl]-4-carboxylate,IUPACName:methyl4-(4-bromophenyl)benzoate,Reagents2:OB(O)c1ccccc1,Synonyms:Phenylboronic acid,IUPACName:phenylboronicacid,Product1:COC(=O)c1ccc(-c2ccc(-c3ccccc3)cc2)cc1,Synonyms:,IUPACName:methyl4-(4-phenylphenyl)benzoate,Reagents:Pd(OAc)2,PPh3,TBAB;solvents:Toluene,H2O`

Model Output:

answer: The reaction rate would decrease significantly, or the reaction might fail to proceed under the original mild conditions. Chemical Reason: The C-Cl bond is stronger than the C-Br bond. Consequently, the oxidative addition step—where the Pd(0) catalyst inserts into the carbon-halogen bond—becomes much more difficult and acts as the rate-determining step.

explanation:
Theoretical Basis: The bond dissociation energy (BDE) of a C–Cl bond ($\sim$339 kJ/mol) is significantly higher than that of a C–Br bond ($\sim$285 kJ/mol), making oxidative addition of Pd(0) into a C–Cl bond thermodynamically less favorable and kinetically slower. This increased bond strength directly impacts the rate of the first step in the Suzuki-Miyaura coupling mechanism, which is oxidative addition. Additionally, chlorine is a poorer leaving group than bromine due to its lower polarizability and greater electronegativity, further reducing the rate of this step.

Mechanism Analysis: In the standard Suzuki coupling mechanism, oxidative addition is typically the rate-determining step when using aryl halides. With an aryl chloride instead of an aryl bromide, the electron-deficient palladium center struggles to cleave the stronger C–Cl bond. The PPh_3 ligand used in the original conditions is moderately electron-rich but

not sufficiently so to stabilize the transition state for oxidative addition of a C-Cl bond. Moreover, PPh_3 has a moderate cone angle ($\sim$145°), which may not favor formation of the active monoligated Pd(0) species required for efficient oxidative addition.

Synthetic Strategy: To overcome this limitation, electron-rich and sterically bulky phosphine ligands must be employed. Buchwald-type ligands such as SPhos, XPhos, or RuPhos are ideal because they increase electron density at the palladium center, facilitating oxidative addition of the strong C-Cl bond. These ligands also possess large cone angles ($>$150°), promoting the formation of low-coordinate, highly reactive Pd(0) species. Alternatively, N-heterocyclic carbene (NHC) ligands like IPr or SIPr can be used due to their strong σ-donating ability, enhancing the nucleophilicity of Pd(0). A recommended optimized system would include $Pd(OAc)_2$ with XPhos (or SPhos) in a mixture of dioxane/water or toluene/water at elevated temperature (80-100 °C) to achieve successful coupling.

Question 5 serves as a stress test for synthetic planning, asking the model to predict reactivity changes for an Aryl Chloride substrate and propose an optimization strategy.

- Baseline: The baseline correctly identifies the primary thermodynamic barrier: the higher Bond Dissociation Energy (BDE) of C–Cl vs. C–Br. However, its reasoning is *unidirectional*, focusing solely on the substrate's limitations. It fails to explain specifically why the current ligand (PPh_3) is insufficient beyond generic electronic arguments, offering only a standard textbook response.
- ChemVA: It analyzes not just the substrate, but the specific incompatibility between the substrate and the ligand's steric properties.
  1. *Mechanistic Depth:* It explicitly cites the Cone Angle ($\sim$145°) of PPh_3, reasoning that it is insufficient to promote the formation of the highly active *monoligated* Pd(0) species required for difficult oxidative additions. This incorporates advanced organometallic concepts (steric bulk vs. coordination number) absent in the baseline.
  2. *Operational Precision:* While the baseline suggests generic ligand classes, ChemVA proposes a complete experimental protocol, including specific solvents (dioxane/water) and temperature ranges (80–100 °C). This reflects a "tactical" understanding of reaction conditions that is crucial for wet-lab success.

A4.3 The Role of Semantic Activation

The qualitative superiority of ChemVA across Q3 (Mechanism), Q4 (Context), and Q5 (Optimization) highlights the critical function of **Semantic Activation**.

Visual encodings (or SMILES strings) alone often trap Large Language Models in a "surface-level retrieval" mode, where they generate generic facts based on pattern matching. By translating visual features into explicit, knowledge-rich textual descriptors (*e.g., IUPAC names, specific reagents like* PPh_3), our method unlocks the model's latent "expert subspace."

For instance, in Q5, the token "PPh_3"—when explicitly presented via Semantic Activation—triggers associated concepts like *Cone Angles* and *Tolman parameters* from the model's pre-training corpus. This allows the model to move beyond simple property retrieval (BDE values) to complex, multi-variable reasoning (steric-electronic fits), effectively bridging the *Semantic Disconnect* between visual perception and deep chemical reasoning.

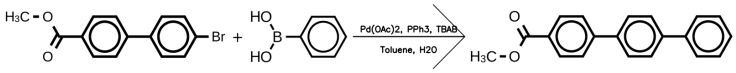
A4.4 Artifact Statement

Artifact Licenses and Intended Use. We use publicly available scientific artifacts in accordance with

their stated licenses and terms of use. PubChem is used for molecular records and entity-name resolution; ORD reaction data are publicly released under CC-BY-SA, and ORD software tools are released under Apache-style open-source terms. Qwen2.5-VL is used under its released model license. The artifacts created in this work, including FG-SFT and OCRD-Bench, are intended for research use in chemical reaction diagram understanding and evaluation.

Sample: Transition Metal Catalysis -> Suzuki Coupling

Reaction Diagram:



Q1: Please analyze the provided reaction image. Convert the reactant and product molecules into SMILES strings.
A1: Reactant1: COC(=O)c1ccc(cc1)c2ccc(Br)cc2, Reagents2: OB(O)c1ccccc1, Product1: COC(=O)c1ccc(cc1)c2ccc(cc2)c3ccccc3
Q2: List all reaction conditions (catalysts, solvents, reagents) textually present in the image.
A2: Reagents: Pd(OAc)2, PPh3, TBAB ; solvents: Toluene, H2O

Q3: In the reaction shown in the image, a base (e.g., K2CO3) is essential. Which of the following best describes the primary mechanistic role of the base in this catalytic cycle?
A. It acts as a reducing agent to convert the Pd(II) precatalyst into the active Pd(0) species.
B. It coordinates with the organoboron reagent to form a boronate complex [R−B(OH)3]−, which facilitates the transmetallation step.
C. It deprotonates the α-carbon of the aryl halide to generate a benzyne intermediate.
D. It serves solely to neutralize the acid (HX) byproduct and drive the equilibrium, without participating in the intermediate steps.
A3: B,
Explanation: In the Suzuki-Miyaura mechanism, the neutral organoboronic acid is not nucleophilic enough to undergo transmetallation with the intermediate Pd(II) complex. The base coordinates to the boron atom, forming a four-coordinate boronate anion($[R-B(OH)_3]^-$). This negatively charged species is sufficiently nucleophilic to transfer the organic group to the palladium center (Transmetallation). While the base does neutralize acid (Option D), that is a thermodynamic consequence, not the primary kinetic role required for the cycle to proceed.

Q4: The reaction proceeds via a standard Palladium(0)/Palladium(II) cross-coupling mechanism. Select the correct chronological sequence of the three fundamental elementary steps in the catalytic cycle.
A. Transmetallation → Oxidative Addition → Reductive Elimination
B. Oxidative Addition → Reductive Elimination → Transmetallation
C. Reductive Elimination → Oxidative Addition → Transmetallation
D. Oxidative Addition → Transmetallation → Reductive Elimination
A4: D
Explanation: The standard catalytic cycle for cross-coupling follows this order:
Oxidative Addition: The active Pd(0) species inserts into the C-X bond of the aryl halide.
Transmetallation: The organic group is transferred from the activated boron species to the Pd(II) center.
Reductive Elimination: The two organic ligands on Pd(II) couple to form the product, regenerating the Pd(0) catalyst.

Q5: If the reactant Aryl Bromide were replaced with the corresponding Aryl Chloride, predict how the reaction rate would likely change. Explain the underlying chemical reason based on bond dissociation energies. Propose a specific synthetic strategy (e.g., a specific class of ligands or catalyst system) that would effectively promote the reaction for the Aryl Chloride substrate.
A5: Prediction: The reaction rate would decrease significantly, or the reaction might fail to proceed under the original mild conditions. Chemical Reason: The C-Cl bond is stronger than the C-Br bond. Consequently, the oxidative addition step—where the Pd(0) catalyst inserts into the carbon-halogen bond—becomes much more difficult and acts as the rate-determining step. Optimization Strategy: The reaction requires the use of electron-rich, bulky phosphine ligands. Specific Examples: Buchwald ligands or trialkylphosphines. Mechanism: Electron-rich ligands increase the electron density on the palladium center, facilitating the difficult oxidative addition into the C-Cl bond. Steric bulk promotes the formation of the active monoligated Pd(0) species.

Figure A3: **An Example Case from OCRD-Bench.** We use the *Suzuki-coupling* to illustrate our hierarchical evaluation design. **L1 (Q1, Q2)** assesses visual perception and structure grounding. **L2 (Q3, Q4)** evaluates static knowledge retrieval. **L3 (Q5)** tests deep mechanistic reasoning.