ChemVTS-Bench: Evaluating Visual–Textual–Symbolic Reasoning of Multimodal Large Language Models in Chemistry

Zhiyuan Huang[†1], Baichuan Yang[†2], Zikun He[†1], Yanhong Wu[†3], Fang Hongyu[4]
Zhenhe Liu[1], LIN DONGSHENG[4], Bing Su[1]
Renmin University of China[1], Beijing University of Posts and Telecommunications[2]
South China University of Technology[3], Gaotu Techedu Inc[4]
huangzhiyuan@ruc.edu.cn [*]

Abstract

Chemical reasoning inherently integrates visual, textual, and symbolic modalities, yet existing benchmarks rarely capture this complexity, often relying on simple image-text pairs with limited chemical semantics. As a result, the actual ability of Multimodal Large Language Models (MLLMs) to process and integrate chemically meaningful information across modalities remains unclear. We introduce ***ChemVTS-Bench****, a domain-authentic benchmark designed to systematically evaluate the Visual-Textual-Symbolic (VTS) reasoning abilities of MLLMs. ChemVTS-Bench contains diverse and challenging chemical problems spanning organic molecules, inorganic materials, and 3D crystal structures, with each task presented in three complementary input modes: (1) visual-only, (2) visual-text hybrid, and (3) SMILES-based symbolic input. This design enables fine-grained analysis of modality-dependent reasoning behaviors and cross-modal integration. To ensure rigorous and reproducible evaluation, we further develop an automated agent-based workflow that standardizes inference, verifies answers, and diagnoses failure modes. Extensive experiments on state-of-the-art MLLMs reveal that visual-only inputs remain challenging, structural chemistry is the hardest domain, and multimodal fusion mitigates but does not eliminate visual, knowledge-based, or logical errors, highlighting ChemVTS-Bench as a rigorous, domain-faithful testbed for advancing multimodal chemical reasoning. All data and code will be released to support future research.*

1. Introduction

Recent advances in Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) have demonstrated unprecedented potential across a wide range of scientific domains, including molecular generation, materials discovery, and scientific reasoning [3, 5, 14]. These models are no longer limited to textual understanding but are gradually acquiring the ability to perform cross-modal symbolic reasoning, enabling them to tackle complex cognitive tasks in fields such as physics, biology, and materials science. In many of these areas, MLLMs have already shown remarkable capabilities in integrating information across modalities, identifying patterns, and supporting hypothesis generation.

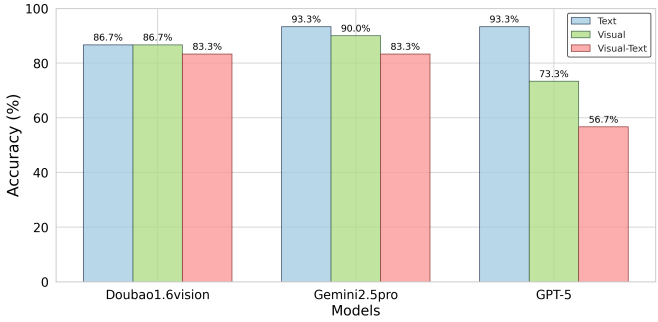

Figure 1. Comparison of model accuracy on ChemVTS-Bench for instances that provide all three input modalities (Text, Visual, and Visual-Text) simultaneously.

However, despite these encouraging developments, the application of multimodal models to certain scientific domains remains underexplored, particularly in chemistry. Chemical problems often involve highly specialized multimodal representations—including intricate organic molecular structures, three-dimensional crystal lattices requiring spatial reasoning, and reaction equations that demand deep domain expertise, as illustrated in Fig. 3. These characteristics make reasoning in chemistry substantially more challenging than conventional visual or textual understanding tasks, highlighting the need for dedicated evaluation frame-

[*†] Equal contribution. Correspondence to Bing Su: subingats@gmail.com.

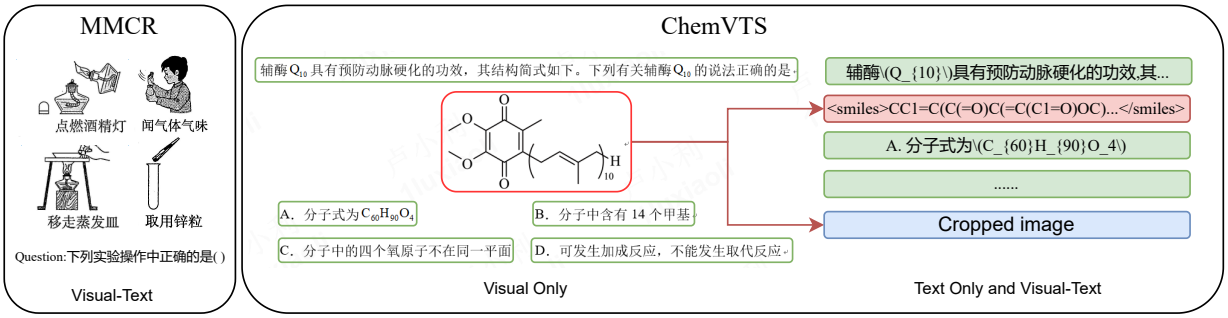

Figure 2. Left: an example from MMCR; Right: an example from ChemVTS. As shown, MMCR contains image prompts with weak chemical semantics—often resembling simple "describe-the-picture" tasks—which may overestimate a model's cross-modal reasoning ability. Moreover, MMCR provides only a visual–text input mode. In contrast, ChemVTS constructs text-only tasks (via OCR extraction and SMILES reconstruction) and visual–text tasks from the original visual-only data, enabling a more comprehensive and fine-grained multimodal evaluation.

works tailored to this domain.

Although several recent studies have introduced chemistry-related multimodal datasets [4, 15], the unique challenges of chemical reasoning mean that directly applying existing benchmarks may not accurately reflect models' true capabilities. Many of these efforts focus primarily on single image–text pairs and do not consider how different modality configurations—such as image, text, and SMILES(Simplified Molecular-Input Line-Entry System) [12]—can lead to divergent reasoning performance as shown in Fig. 1. In addition, the visual data used in some benchmarks [8] often lack rich chemical semantics (as illustrated in Fig. 2), which may inadvertently overestimate the actual reasoning abilities of current MLLMs in the chemistry domain. Consequently, there remains a need for a benchmark that can rigorously and systematically evaluate MLLMs' reasoning across multiple chemical modalities.

To address these limitations, we introduce ChemVTS-Bench, a novel benchmark specifically designed to evaluate the Visual–Textual–Symbolic (VTS) reasoning capabilities of MLLMs in chemistry. Unlike existing datasets, ChemVTS-Bench systematically examines how different modality configurations influence a model's reasoning process. It comprises three complementary input settings—(1) pure image input, (2) image–text hybrid input, and (3) SMILES-based symbolic input—enabling a fine-grained analysis of modality-dependent reasoning behaviors. Furthermore, the benchmark covers a broad spectrum of chemical domains, including organic compounds, inorganic materials, and three-dimensional crystal structures, thereby assessing both conceptual understanding and spatial reasoning. By leveraging authentic, domain-specific visual representations rather than generic scientific imagery, ChemVTS-Bench provides a realistic and rigorous evaluation of MLLMs' chemical reasoning capabilities, effectively mitigating the overestimation observed in previous benchmarks. We will open-source all our data and code to promote the development of the community.

In summary, our work makes the following key contributions:

- **A domain-authentic and challenging benchmark**: We construct ChemVTS-Bench, a large-scale, high-quality dataset that spans diverse chemistry domains—including organic and inorganic compounds as well as 3D crystal structure. Each problem is designed to capture the intrinsic multimodal nature of chemical reasoning, offering three distinct input modes (image, image–text, and SMILES) for rigorous evaluation.
- **Systematic evaluation framework**: We develop an automated and efficient agent-based workflow that systematically evaluates the visual–textual–symbolic reasoning abilities of MLLMs across different modalities and reasoning pathways. The framework not only standardizes the assessment process but also performs fine-grained error diagnosis, enabling comprehensive analysis of models' failure modes.
- **Extensive experimental analysis**: We conduct thorough experiments on a suite of state-of-the-art MLLMs, revealing current limitations and providing insights into their strengths and weaknesses in chemical cross-modal reasoning.

2. Related works

2.1. Multimodal Large Language Models

In recent years, MLLMs have achieved remarkable progress across a variety of application domains. Researchers have begun introducing these models to chemical problems, particularly for representing organic molecular structures and

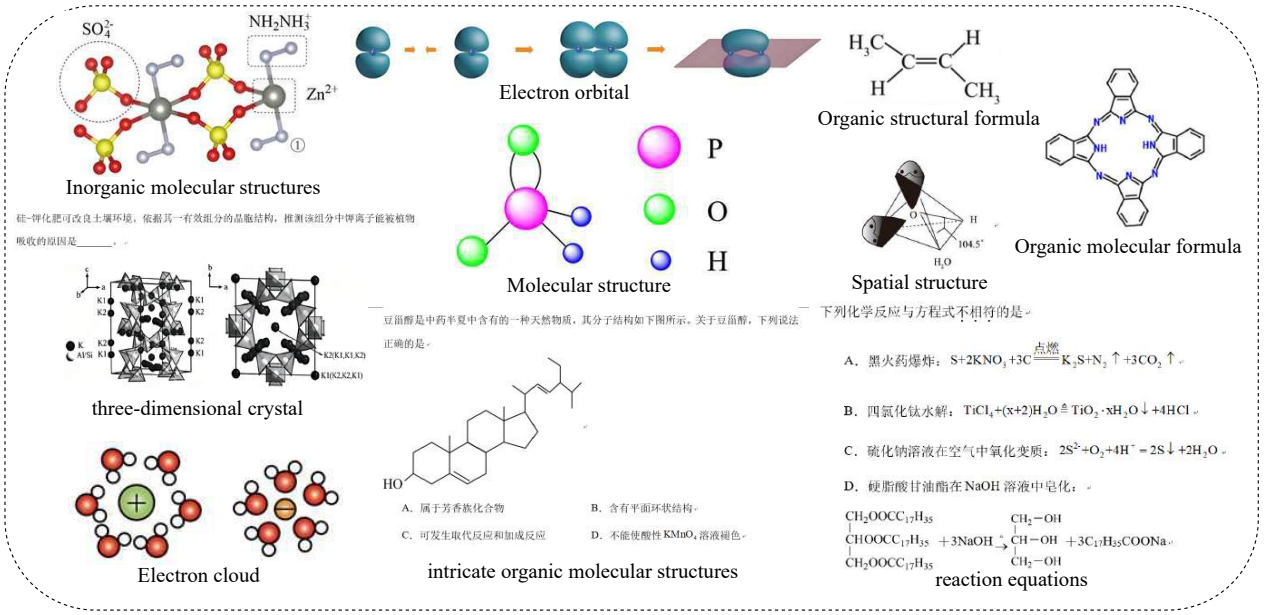

Figure 3. Representative visual examples from ChemVTS-Bench, highlighting the diverse and challenging chemical structures present in the benchmark.

reasoning about reaction mechanisms. Nevertheless, chemical MLLM still face substantial challenges in cross-modal inference, structural representation, and the design of complex reasoning chains [4]. Chemical tasks typically involve multiple intricate modalities, typically including chemical structure diagrams, SMILES strings, and textual descriptions. Different modalities may contain complementary information, which leads MLLM to struggle to fully align information across modalities, thereby undermining inference performance.

For example, contemporary vision–language models such as BLIP-3 [13] and InternVL3 [16] exhibit strong recognition capabilities in general vision–language tasks, while they tend to lack sufficient domain knowledge and reasoning capacity for complex chemical problems; as a result, these models may correctly recognize a question, but fail to answer more demanding questions. By contrast, general MLLM such as GPT-5 [7], Doubao1.6 vision [1], and Gemini2.5-pro [2] possess stronger foundational knowledge and reasoning abilities. However, these models must still meet the challenge of accurately capturing fine-grained details from visual modalities because many aspects of chemical problems are difficult to express purely in natural language.

2.2. Benchmarks in the Chemical Domain

Numerous benchmarks targeting Multimodal Large Language Models (MLLMs) in the chemical domain have been proposed in the academic literature. Among the representative existing benchmarks, ChemIQ [9] is a text-only dataset that represents chemical structures via SMILES strings and evaluates model performance through a question-and-answer format. ChemTable [15] is a dataset focused on recognizing and understanding tabular data from chemical literature, particularly assessing a model's ability to extract chemical data from tables in scientific documents. ChemBench [11] is a text-based chemistry benchmark that covers a wide range of chemistry topics, task types, and skill categories. It supports structured modalities such as SMILES or equations through semantic annotation but lacks truly multimodal inputs like images. MMCR-Bench [4] is mainly derived from the Chinese college entrance examination chemistry section's assessment model's ability to solve complex multimodal chemical reasoning (MMCR) problems, including molecular title and molecular property prediction tasks. However, as shown in Figure 2, many samples in MMCR-Bench have images that are unrelated to chemical structures and are not helpful for understanding the questions. Furthermore, MMCR-Bench only provides a visual-text hybrid modality.

While these benchmarks provide basic task evaluation frameworks for MLLMs in chemistry, they have several shortcomings in assessing multimodal capabilities and reasoning abilities. First, most existing benchmarks lack visual data related to molecular and atomic structures, despite the fact that many chemical properties are not adequately represented by simple molecular formulas but are implicitly encoded in two-dimensional or three-dimensional structures.

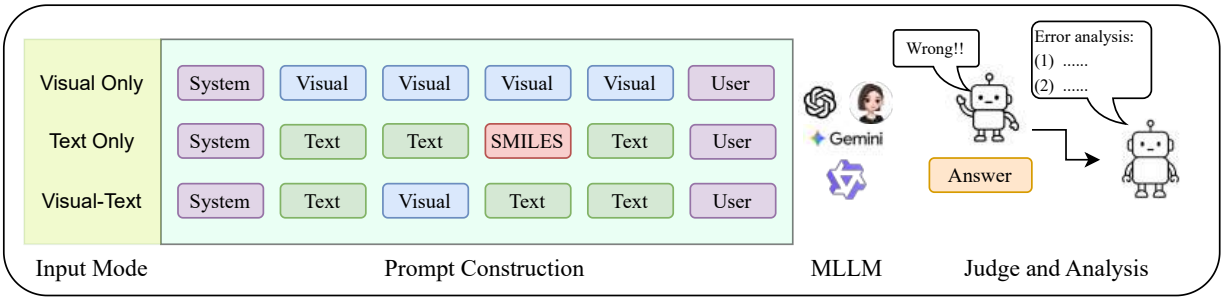

Figure 4. Our Evaluation Pipeline. Visual Text SMILES are used as input tokens for the problem. The prompts for the System and the User follow a unified input format with slight variations. The answers generated by the MLLM are then processed through a two-stage agent workflow to obtain evaluation outcomes and error analysis, respectively.

Benchmark	Vision	SMILES	Reasoning	Structure Rep.
ChemIQ	✗	✓	✓	✗
ChemTable	✓	✗	✓	✗
ChemBench	✗	✓	✗	✗
MMCR-Bench	✓	✗	✓	✗
ChemVTS	✓	✓	✓	✓

Table 1. Comparison of Benchmarks, Rep. denotes representation

Furthermore, current benchmarks do not cover all essential modalities. Visual, textual, and SMILES representations are all crucial for chemical problem-solving, yet existing single-modality benchmarks fail to evaluate a model's integrated performance across all modalities. Finally, the tasks in current benchmarks are overly simplistic, focusing on rote knowledge recall, basic image captioning, and formula transcription, which do not exceed the average high school level of difficulty. This lags behind the progress of multimodal large models in more general domains.

3. ChemVTS-Bench

Composition and Categorization The high-quality chemical questions in ChemVTS-Bench are drawn from a large corpus of high-school chemistry exam items. We first employed the Qwen3-VL model to filter questions containing images, after which professional chemistry teachers were invited to conduct a secondary selection. Their review ensured that the accompanying images were highly relevant to the chemical content of each question and played a critical role in comprehension. Based on their subject-matter expertise, the teachers determined that these questions typically involve chemical structures that are difficult to convey through text alone and therefore require visual or SMILES-based representations. The selected items also demand a certain degree of reasoning and fall within a moderate difficulty range. The question formats include both multiple-choice and fill-in-the-blank types.

Following the high-school chemistry curriculum standards and exam-design guidelines, we categorized the filtered questions into 3 groups , structural chemistry, organic chemistry, and others , resulting in a total of 445 questions. Furthermore, leveraging the recognition capabilities of Qwen3-VL and the teachers' manual verification, ChemVTS-Bench provides fine-grained annotations for each question according to the specific visual skills and chemical knowledge points required.

Finally, ChemVTS-Bench includes standard answers and reference explanations, enabling researchers to assess the correctness of model outputs and to evaluate whether the reasoning process is accurate.

We further conducted extensive experimental validation to assess the coverage of the ChemVTS-Bench dataset. A group of chemistry teachers—who did not participate in the earlier question-screening process—were invited to randomly select a full set of comprehensive high-school chemistry exam papers. They analyzed each item in the set to determine whether its corresponding category and knowledge points were represented in ChemVTS-Bench. The results demonstrate that ChemVTS-Bench provides strong coverage of the major content areas in the high-school chemistry curriculum.

Data Processing As illustrated in Fig. 2, we extracted the selected image-based questions from the original item bank using a combination of OCR techniques and manual annotation, separating each question into its textual and visual components. The visual component contains all structures and diagrams relevant to chemical content, and each distinct chemical structure was segmented into an independent image. We subsequently attempted to convert these chemical structures into SMILES representations. Due to limitations in SMILES expressiveness, 95 questions ultimately contain both visual and SMILES modalities. We anticipate that providing multiple modalities for the same ques-

tion will enable researchers to analyze model performance across different input types and thereby pursue more targeted improvements.

In the textual component, certain LaTeX expressions are retained to represent equations and molecular formulas, and special tokens are inserted to indicate the relative positions of chemical images within the text. Through this processing pipeline, we obtained a high-quality dataset with fine-grained labels on chemical knowledge categories.

4. Experiments

4.1. Experimental Setup

Evaluation Models To comprehensively evaluate the capabilities of Multimodal Large Language Models (MLLMs) in the domain of chemical reasoning, we assess six representative models on the ChemVTS-Bench benchmark. The selection of these models is designed to cover a diverse spectrum of the current MLLM landscape, including both the latest closed-source systems from leading industrial labs and prominent open-source alternatives that are widely accessible. The evaluated closed-source models comprise GPT-5 [7], renowned for its advanced reasoning capabilities; Gemini-2.5-Pro [2], recognized for its strong performance on scientific and multimodal tasks; and Doubao 1.6 Vision [1], a powerful model from a major Chinese tech company. On the open-source front, we evaluate Qwen3-VL [10], known for its competitive vision-language integration; InternVL 3 [16], which emphasizes scaling vision foundation models to match large language models; and Llama 3.2 [6], a recent iteration of a highly influential open-weight language model family, often used as a base for multimodal fine-tuning. This comparative analysis allows us to benchmark the chemical reasoning prowess of state-of-the-art MLLMs, providing insights into their abilities to interpret molecular structures, understand chemical phenomena, and solve problems based on textual and visual information.

4.2. Evaluation Pipeline

As illustrated in the Fig. 4, we ensure both fairness and practicality during evaluation by adopting a consistent prompting strategy across all models, with only minimal adjustments to accommodate different input modalities. Specifically, for each input mode—visual-only, text-only, and visual–text hybrid—we design dedicated prompt templates that are carefully tailored to effectively elicit the model's reasoning process while preserving consistency in instruction phrasing and response format across all experimental conditions. This approach minimizes confounding factors introduced by prompt design, allowing for a more direct comparison of model capabilities. All inference runs are conducted with the temperature parameter set to 0 to minimize the stochasticity inherent in generative MLLMs, thereby enhancing the reproducibility and reliability of our results. After obtaining the model's output for a given problem instance, we employ a two-stage automated agent-based workflow to systematically evaluate and analyze performance. In the first stage, a Doubao-based [1] evaluation agent compares the model's response with the ground-truth solution. This agent is specifically prompted to assess the correctness of the final answer while considering permissible variations in expression or intermediate reasoning steps, ensuring a fair and nuanced judgment. The agent outputs a binary decision indicating whether the answer is correct or incorrect.

In the second stage, for responses that are deemed incorrect, another specialized Doubao-based [1] analysis agent is invoked to perform a detailed error diagnosis. This agent categorizes the failure into one of several predefined error types—such as *Visual Recognition Error*, *Symbolic Parsing Error*, *Numerical Computation Error*, *Logical Reasoning Erro*, or *Knowledge-based Erro*—based on a set of explicit classification guidelines. This two-step agent workflow not only enables efficient and accurate answer verification at scale but also provides fine-grained diagnostic insights into the specific weaknesses and failure patterns of each model. By automating both the scoring and the error analysis, this pipeline greatly reduces the burden of manual evaluation while offering rich interpretability into model behavior. Additional details regarding the implementation of the agent prompts and the error taxonomy are provided in the Error Analysis section and the Appendix.

4.3. Main Results

The main experimental results are presented in Tab. 2. Owing to the multiple input modalities per question and the diverse chemical domains covered, our benchmark enables more detailed and fine-grained analyses than prior benchmarks:

Modality-specific Consistency Analysis To assess model robustness under varying input modalities, we compute answer consistency, defined as the proportion of instances for which a model outputs responses that are simultaneously correct or incorrect across the available modalities of each question. This metric reflects whether a model can maintain stable reasoning when presented with visual-only, text-only, and visual–text fused inputs.

As shown in Tab. 3, consistency varies substantially across models and modality groups. Proprietary models such as Gemini-2.5-Pro exhibit strong cross-modal robustness, achieving consistently high agreement across all modality pairs. In contrast, open-source models demonstrate more pronounced variability: Qwen3-VL and InternVL 3 show relatively strong alignment in VT&T (text-only vs. visual–text), yet their performance degrades in set-

Model	Visual-only				Texts-only				Visual-Texts			
	SC	OC	Others	Average	SC	OC	Others	Average	SC	OC	Others	Average
	Closed-source											
GPT-5 [7]	65.20	66.00	73.58	66.29	79.49	79.59	100	81.05	54.40	62.50	72.73	57.67
Gemini-2.5-Pro [2]	72.51	84.00	88.68	75.73	89.74	79.59	100	85.26	70.40	87.50	90.91	74.85
Doubao 1.6 Vision [1]	58.77	88.00	84.91	65.17	69.23	79.59	100	76.84	60.00	81.25	81.82	65.03
	Open-source											
Qwen3-VL [10]	62.87	72.00	81.13	66.07	69.23	67.35	85.71	69.47	55.20	81.25	77.27	60.74
InternVL 3 [16]	30.12	50.00	47.17	34.38	48.72	48.98	85.71	51.58	31.20	75.00	27.27	34.97
Llama 3.2 [6]	33.04	42.00	30.19	33.71	58.97	55.10	71.43	57.89	41.60	75.00	59.09	47.24

Table 2. Performance of various multimodal large language models on ChemVTS-Bench. The reported values indicate accuracy. SC stands for Structural Chemistry, and OC stands for Organic Chemistry.

Model	VTS	V&T	VT&V	VT&T
GPT-5 [7]	53.33	72.63	71.17	56.67
Gemini-2.5-Pro [2]	86.67	90.53	84.66	90.00
Doubao 1.6 Vision [1]	73.33	82.11	82.21	76.67
Qwen3-VL [10]	70	78.95	79.14	86.67
InternVL 3 [16]	66.67	75.79	73.01	83.33
Llama 3.2 [6]	50.00	66.32	61.96	80

Table 3. Answer consistency results on ChemVTS-Bench. "VTS" denotes questions that contain all three input modes; "V&T" denotes those with both visual-only and text-only modes; "VT&V" denotes those with both visual-only and visual–text modes; and "VT&T" denotes those with both text-only and visual–text modes.

tings involving visual-only inputs (VTS and VT&V). This gap suggests that visual grounding remains a key bottleneck for current open-source MLLMs. Notably, models such as Llama 3.2 display large discrepancies across modality combinations, indicating unstable reasoning pathways when the input representation changes.

Overall, these results highlight that—even when accuracy appears comparable—models may differ significantly in their ability to produce modality-invariant predictions, underscoring the importance of modality-specific consistency analysis for evaluating true multimodal robustness.

Domain-wise Analysis. The domain-level results further reveal distinct reasoning challenges across different chemistry subfields.

For **Structural Chemistry**, the overall poor performance can be attributed to the intrinsic visual–symbolic complexity of the tasks. Many questions involve crystal lattices, organic frameworks, or novel compound structures that models have not encountered during pretraining. In the *text-only* setting, such structural information cannot be exploited effectively, as these problems rely heavily on spatial or topological cues. Meanwhile, the *visual-only* models struggle to recognize fine-grained atomic arrangements or bond geometries. In addition, numerous problems contain molecular diagrams that are not easily captured by OCR systems, causing models to misinterpret the visual input or even hallucinate nonexistent structural features, leading to erroneous reasoning.

In contrast, **Organic Chemistry** questions are generally more approachable for MLLMs. When problems focus on fundamental concepts—such as reaction types, functional groups, or substitution rules—models can often derive the correct answers directly from textual reasoning. Many organic chemistry problems involve molecular structures that can be represented using SMILES strings, which enables models to semantically parse and reason over the full molecular structure. For problems involving structures that cannot be fully represented by SMILES, the model can sometimes rely on textual descriptions to infer the correct answer, especially when the image is not essential for reasoning. However, in a few cases involving newly synthesized or uncommon organic compounds, even advanced models fail due to a lack of prior chemical knowledge and limited ability to comprehend unseen molecular structures. When models succeed on such questions, their success often stems from leveraging textual cues or analogical reasoning—identifying structural similarities between the described compound and known molecules.

For the **Others** category, the relatively high accuracy can be explained by the nature of the questions, which mostly assess chemical common sense or fundamental principles such as periodic trends. These problems require little domain-specific visual understanding, allowing models to perform well even when the image content is partially unrecognized. When questions involve more complex visual elements—such as electrochemical equations for organic materials—closed-source models exhibit greater robustness, likely because they can implicitly encode molec-

ular information from SMILES representations of structurally similar compounds and leverage such analogies to better interpret the reactants and underlying reaction mechanisms. For element inference tasks, the high success rate can be attributed to the fact that most elements involved are main-group atoms (atomic number below 30), whose properties are well covered in pretraining corpora. Consequently, models can often derive the correct answers directly from textual reasoning without relying on image understanding.

4.4. Error Analysis

Based on the comparative performance of different MLLMs across various modalities and domains, we further categorize and analyze the causes, types, and distributions of errors observed during model evaluation. To this end, the errors exhibited by MLLMs are systematically classified into the following five categories:

- **Visual Recognition Error (VRE)** refers to failures in the model's perception or interpretation of visual inputs. This includes cases where the model falsely denies the presence of an existing object, misidentifies one structure or entity as another, or provides inaccurate spatial or compositional descriptions of image elements.
- **Symbolic Parsing Error (SPE)** denotes incorrect interpretation or generation of symbolic notations such as chemical equations, structural formulas (e.g., line-angle diagrams), or linear notations (e.g., SMILES). These errors usually arise from syntactic misplacement, tokenization mistakes, or semantic confusion within symbolic systems.
- **Numerical Computation Error (NCE)** involves inaccuracies in mathematical reasoning or quantitative calculations, including arithmetic mistakes, rounding or approximation errors, and inconsistencies in intermediate computational steps.
- **Knowledge-based Error (KBE)** occurs when the model exhibits factual inaccuracies or misconceptions about established chemical knowledge, physical principles, or empirical data. This also includes fabricating nonexistent compounds, reactions, or properties, or introducing extraneous information not supported by the given context.
- **Logical Reasoning Error (LRE)** refers to flaws in inferential reasoning or argumentative structure, such as broken causal chains, contradictory premises, or inconsistencies across reasoning steps. These reflect failures in maintaining logical coherence and internal consistency during the reasoning process.

Furthermore, we define each error type with tailored prompts to guide the agent's classification. For instance, the *Visual Recognition Error* prompt is designed to detect specific keywords. The presence of phrases such as *in the figure*, *in the image*, or *as shown in the figure*, alongside action verbs like *identified as*, *misidentified as*, *mistaken for*, *overlooked*, or *misinterpreted*, will trigger the agent to assign this error category. The underlying logic is that if an MLLM's erroneous solution contains terms like *in the figure* while arriving at an incorrect conclusion, it strongly indicates a failure in the visual perception process. Similarly, the use of verbs that explicitly denote visual recognition, such as *mistaken for* or *overlooked*, in the context of an error, is also classified as a VRE.

The above definitions distinguish the five error categories clearly and minimize potential overlaps among them. Based on these refined definitions, as shown in the Fig. 5, we employ the high-accuracy agent which was introduced before to evaluate the errors produced by different MLLMs under various modalities, identifying both the cause of each error and the corresponding correct answer. This significantly reduce the manual workload of error identification and analysis. Full distribution of error causes for the six MLLMs is illustrated in the Appendix.

Error distributions vary across different modalities. For text-only problem-solving tasks, we found that MLLMs generally achieve higher accuracy than in visual or Visual-Text settings. Among all models, *Gemini-2.5-pro* and *InternVL3* exhibited the best and worst overall performances, with average accuracy 85.26 and InternVL 3, respectively. The majority of these errors originated from issues in textual understanding and analysis, which will not be further discussed here.

In the case of visual-only tasks, the best and worst performing models were again *Gemini-2.5-pro* and *Llamma 3.2*, with average error accuracy of 75.73 and 33.71, respectively. The detailed error type distribution of all models is shown in the Appendix. Visual recognition errors(VRE) were the most prevalent across all models, suggesting that current MLLMs still struggle with the accurate recognition of chemical imagery such as crystal structures, molecular diagrams, and schematic representations. In contrast, models performed relatively well in symbolic parsing and numerical computation, indicating that the current architectures are already capable of handling chemical text and mathematical expressions effectively. However, logical reasoning and knowledge-based errors varied significantly among models. Most commercial MLLMs demonstrated strong factual consistency, while some open-source models exhibited severe hallucinations, likely due to limitations in training data quality or domain coverage. Fundamentally, both knowledge-based and logical reasoning errors stem from the model's training corpus and architectural design, rather than from the input modality itself.

For Visual-Text problem-solving tasks, all models showed an increasement in average error accuracy. As illus-

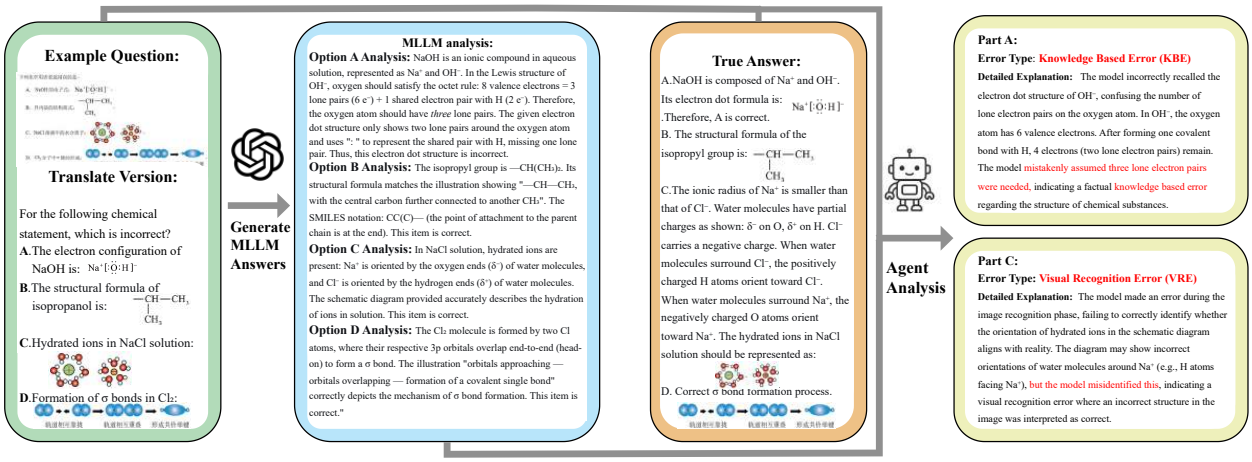

Figure 5. This figure depicts a specific example of the pipeline for MLLM-generated solution evaluation and Agent-based error diagnosis. The process begins by inputting a ground-truth problem into the MLLM to obtain its solution. Subsequently, the Agent receives the original problem, the reference solution, and the MLLM's solution as inputs, finally producing a diagnosis of the error type.

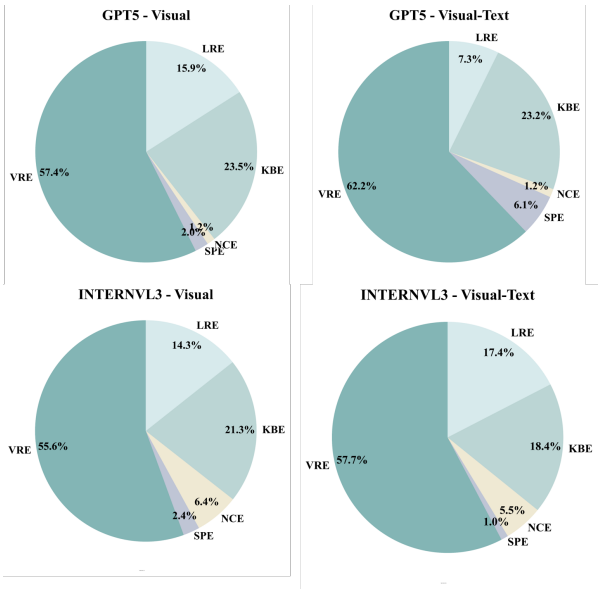

Figure 6. **Distribution of GPT-5 and InternVL3 Error Types.** For the five predefined categories of error causes, we present GPT-5 and InternVL3's error distribution as illustrative examples, including the count and proportion of each error type as well as the total errors.

trated in Fig. 6, the shifting distribution of error types that the MLLM's capability in processing visual information from the problems remains inadequate, even with the integration of textual-visual hybrid data. This deficiency results in a persistent, and even increased, proportion of vision-related errors. For the remaining four error categories, the shift in input modality did not lead to a substantial performance gain. This is because reasoning and factual accuracy are primarily determined by model architecture and training corpus rather than input format. Similarly, symbolic parsing and numerical computation errors remained consistently low across modalities, suggesting that modern MLLMs already possess mature capabilities in symbolic reasoning and quantitative computation.

5. Conclusion

We introduced ChemVTS-Bench, a domain-authentic benchmark for evaluating the visual–textual–symbolic (VTS) reasoning abilities of MLLMs in chemistry. Through three complementary input modalities, ChemVTS-Bench enables fine-grained analysis of multimodal robustness and reasoning pathways. Our experiments reveal substantial gaps across models: closed-source systems exhibit strong cross-modal consistency, whereas open-source models struggle particularly with visual-only inputs, highlighting persistent weaknesses in chemical visual grounding. Domain-wise results show that structural chemistry remains the most challenging, while organic chemistry tasks benefit from symbolic (SMILES) representations. Error analysis further demonstrates that visual recognition errors dominate in unimodal settings and decrease notably with multimodal inputs, whereas knowledge-based and logical reasoning errors persist across modalities. Together, these findings show that current MLLMs lack stable, domain-faithful chemical reasoning, establishing ChemVTS-Bench as a rigorous testbed for future progress.

References

[1] ByteDance. Seed1.6 tech introduction. https://seed.bytedance.com/en/seed1_6, 2025. Accessed: November 2025. 3, 5, 6

[2] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 3, 5, 6

[3] Wei Liu Di Zhang, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Dongzhan Zhou, Shufei Zhang, et al. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*, 9, 2024. 1

[4] Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, et al. Chemvlm: Exploring the power of multimodal large language models in chemistry area. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 415–423, 2025. 2, 3

[5] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024. 1

[6] Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, 2024. 5, 6

[7] OpenAI. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/, 2024. 3, 5, 6

[8] Jiacheng Ruan, Dan Jiang, Xian Gao, Ting Liu, Yuzhuo Fu, and Yangyang Kang. Mme-sci: A comprehensive and challenging science benchmark for multimodal large language models. *arXiv preprint arXiv:2508.13938*, 2025. 2

[9] Nicholas T Runcie, Charlotte M Deane, and Fergus Imrie. Assessing the chemical intelligence of large language models. *arXiv preprint arXiv:2505.07735*, 2025. 3

[10] Qwen Team. Qwen3 technical report, 2025. 5, 6

[11] T Walker, CM Grulke, D Pozefsky, and A Tropsha. Chembench: a cheminformatics workbench. *Bioinformatics*, 26 (23):3000–3001, 2010. 3

[12] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988. 2

[13] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024. 3

[14] Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. A comprehensive survey of scientific large language models and their applications in scientific discovery. *arXiv preprint arXiv:2406.10833*, 2024. 1

[15] Yitong Zhou, Mingyue Cheng, Qingyang Mao, Yucong Luo, Qi Liu, Yupeng Li, Xiaohan Zhang, Deguang Liu, Xin Li, and Enhong Chen. Benchmarking multimodal llms on recognition and understanding over chemical tables. *arXiv preprint arXiv:2506.11375*, 2025. 2, 3

[16] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 3, 5, 6

ChemVTS-Bench: Evaluating Visual–Textual–Symbolic Reasoning of Multimodal Large Language Models in Chemistry

Supplementary Material

6. ChemVTS Item Selection Criteria

The construction of visual items in ChemVTS was carried out by specialists in chemistry education, following a rigorous framework grounded in **China's national high-school chemistry curriculum**. The selection criteria are outlined below:

1. **Curriculum-aligned knowledge boundaries.** Guided by the official curriculum standards, the core concepts, instructional emphases, and known learning difficulties of each module were systematically reviewed. These boundaries determined the scope of image selection. For example, within the module "Ion Equilibria in Aqueous Solutions," images related to acid–base ionization equilibria, hydrolysis equilibria, and precipitation–dissolution equilibria were prioritized.
2. **Textbook-grounded content with pedagogical extensions.** Beyond images explicitly presented in standard Chinese high-school chemistry textbooks, the dataset also incorporates visualizations of concepts that are mentioned but not illustrated, or that typically require strengthened conceptual scaffolding. For instance, although "crystal structures" appear only briefly in textbooks, detailed structural models were supplemented to enhance visual comprehension.
3. **Level-based stratification of conceptual and visual complexity.** To reflect the cognitive progression expected in the curriculum, visual items were organized into three levels of complexity:
   - *Level 1 — Foundational Concepts and Basic Visuals:* Includes introductory conceptual diagrams and simple experimental images, such as matter classification charts, basic laboratory apparatus, and elementary reaction energy profiles.
   - *Level 2 — Principle-oriented and Moderately Complex Representations:* Contains images that require deeper understanding of chemical principles, including electrochemical cell schematics, illustrative chemical process diagrams, and organic synthesis route maps.
   - *Level 3 — Integrative, Cross-topic, and Application-oriented Visuals:* Covers advanced, multi-concept diagrams such as industrial process flowcharts, multi-topic knowledge networks, and problem-driven schematic representations that integrate multiple domains of chemistry.
4. **Emphasis on abstract, challenging, or frequently misunderstood topics.** For conceptual areas where students commonly experience difficulty—such as electrochemical principles, chemical equilibrium, and organic reaction mechanisms—priority was given to concrete and highly visual representations (e.g., microscopic diagrams or stepwise mechanistic flowcharts). Additionally, the dataset includes discriminative or contrastive images targeting high-frequency exam pitfalls (e.g., ion coexistence analysis, pH calculation steps, correct vs. incorrect laboratory operations).
5. **Balanced coverage across major content domains.** The dataset maintains proportional coverage of the major knowledge areas mandated in the Chinese curriculum, including fundamental principles, inorganic and organic chemistry, laboratory skills, structural chemistry, and STSE-related content. This ensures comprehensive representation without overemphasizing or underrepresenting any module.

7. Others Experiment Results

7.1. Heterogeneity Across Input Modalities

As shown in Fig. 7, different models exhibit heterogeneous performance across input modalities, and the degree of heterogeneity varies by model. This divergence reflects the differences in modality-specific reasoning capabilities among models. Existing benchmarks do not provide such fine-grained modality-aligned problem sets, which may lead to either overestimation or underestimation of MLLM chemical reasoning capabilities. In contrast, ChemVTS enables detailed modality-level capability analysis, facilitating a more comprehensive and interpretable evaluation.

7.2. Error Analysis

In this section, we elaborate on the construction of an accurate and efficient agent for root cause analysis, designed to diagnose errors in the solutions generated by various MLLMs. The potential errors have been pre-defined into a taxonomy of five distinct types: (Visual Recognition Error (VRE), Symbolic Parsing Error (SPE), Numerical Computation Error (NCE), Knowledge-based Error (KBE), Logical Reasoning Error (LRE)), each with specific definitions and associated keywords.

Subsequently, we present three concrete examples to elucidate the criteria for classifying these five error types, as shown in Figs. 8 to 10 .

To validate the agent's efficacy, we employed a set of 30 representative error instances generated by various

MLLMs. Each instance was processed by the agent for root-cause analysis and subsequently subjected to manual verification. The agent achieved a 100% consistency rate with human judgments, confirming that it is an accurate, efficient, and resource-conserving tool for large-scale error diagnosis.

8. Evaluation Setting

8.1. Prompt Design for Three Input Modes

To ensure a fair and consistent evaluation across different input settings, we design three prompts corresponding to the three evaluation modes used in ChemVTS-Bench: (1) Image-only, (2) Text-only, and (3) Multimodal (Text + Image). All three prompts follow a unified structure consisting of (i) a shared system prompt and (ii) a mode-specific user prompt. The system prompt remains largely identical across modes, while the user prompts differ according to the input modality.

Shared System Prompt (Common Across All Three Modes)

Shared System Prompt

You are an experienced high school chemistry teacher with deep expertise in chemistry knowledge. Your task is to solve the given chemistry problem thoroughly and accurately following the rules below:
1. Content Extraction You must faithfully transcribe all original question content without any modification. This includes question numbers, textual descriptions, chemical equations, molecular/structural representations, crystal diagrams, and any other expressible chemical information.
2. Solution Procedure - In the **"analysis"** section, provide detailed reasoning including chemical logic, derivations, and intermediate steps. Use LaTeX for chemical equations and SMILES notation for complex organic structures when necessary. - In the **"answer"** section, give concise final answers. Use A/B/C/D for multiple-choice questions.
3. Handling Multi-part Questions List the reasoning and answers for all sub-questions sequentially inside a single output object.
4. Output Format (strict)

```
{
 "analysis":["Overall reasoning: ...",
             "Sub-question (1): ...",
             "Sub-question (2): ..."],
 "answer":["Sub-question (1): ...",
           "Sub-question (2): ..."]
}
```

The following sections describe the mode-specific differences in the user prompt.

8.1.1. Visual-only Mode

User Prompt for Image-only Mode

User Message:
The problem is given as an image: {{IMAGE_CHEMISTRY_QUESTION}}. Please provide the detailed reasoning and final answer following the rules above.

8.1.2. Text-only Mode

User Prompt for Text-only Mode

User Message:
Here is the chemistry problem text: {{CHEMISTRY_QUESTION}}. Please provide the detailed reasoning and final answer following the rules above.

8.1.3. Visual-Text Mode

User Prompt for Multimodal Mode

User Message:
Here is the problem with both text and accompanying images: {{TEXT_IMAGE_CHEMISTRY_QUESTION}}.
Please provide the detailed reasoning and final answer following the rules above.

9. Advantages of ChemVTS-Bench in Structural Chemistry

We argue that what makes chemical problems particularly challenging for multimodal large language models (MLLMs) is the presence of intra- and intermolecular planar and three-dimensional structures that are difficult to interpret. Such phenomena are ubiquitous in chemistry—for example, electronic clouds, bond angles, and stereochemistry in three-dimensional space. These concepts are hard to communicate purely through text and therefore pose substantial challenges to the visual modality of MLLMs. As discussed in related work, existing benchmark datasets exhibit clear deficiencies in the domain of structural chemistry. Among prior efforts, ChemIQ and ChemBench do not include a visual modality, which implies that a large number of chemical structures cannot be faithfully represented. ChemTable focuses on tabular formats, in which the content primarily encodes the relative positions of textual entries rather than chemical structures themselves; for instance, the following figures from ChemTable illustrate this limitation:Fig. 11

MMCR-Bench suffers from similar issues: its visual modality is not specifically tailored to chemical features, and the coverage of chemical structures is limited. For example, the following figures are sample questions drawn from MMCR-Bench :Fig. 12

These questions almost exclusively test the ability of MLLMs to recognize text embedded in images, rather than their capacity to understand and reason about chemical structures. Although the content is presented in visual form, the informative signal is essentially confined to the textual elements within the images. Consequently, these questions provide only a one-sided evaluation of a model's multimodal capability: they primarily probe the MLLM's OCR skills and fail to adequately assess its understanding of genuine, purely structural chemical information.

In ChemVTS-Bench, we place particular emphasis on evaluating MLLMs' ability to comprehend chemical structures. For instance, the dataset in ChemVTS-Bench contains questions of the following types:Fig. 13

These questions involve complex chemical structures. For some organic molecules, the structures can be represented in SMILES format, and we provide the corresponding converted data; others can only be expressed through the visual modality. In ChemVTS-Bench, MLLMs must not only understand the textual components, but also analyze purely structural chemical depictions in order to grasp the meaning conveyed by the figures. This imposes a novel and high-intensity challenge on the multimodal capabilities of MLLMs, and it corresponds to a core ability required for handling complex chemical problems.

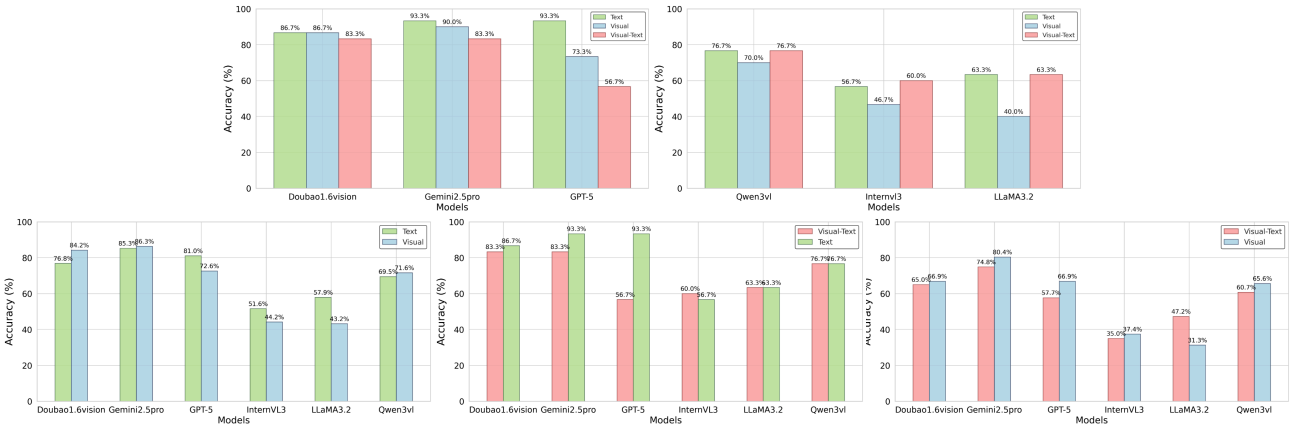

Figure 7. Model accuracies on ChemVTS subsets with different modality availability. The top row shows results on problems containing all three modalities, and the bottom row shows results on subsets containing only two modalities (Text + Visual, Visual-Text + Text, and Visual-Text + Visual).

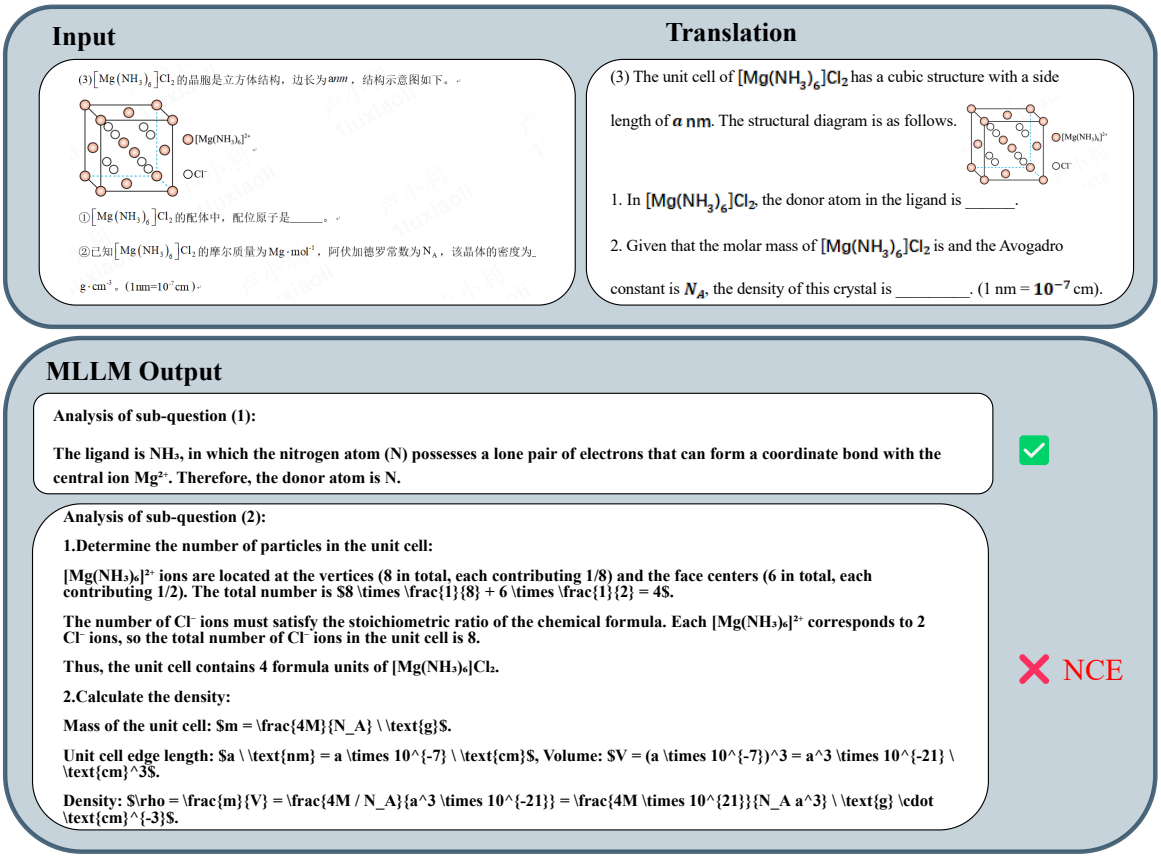

Figure 8. Case Study 1: Root-Cause Analysis The agent diagnosed the MLLM's answer as containing Numerical Computation Error (NCE).

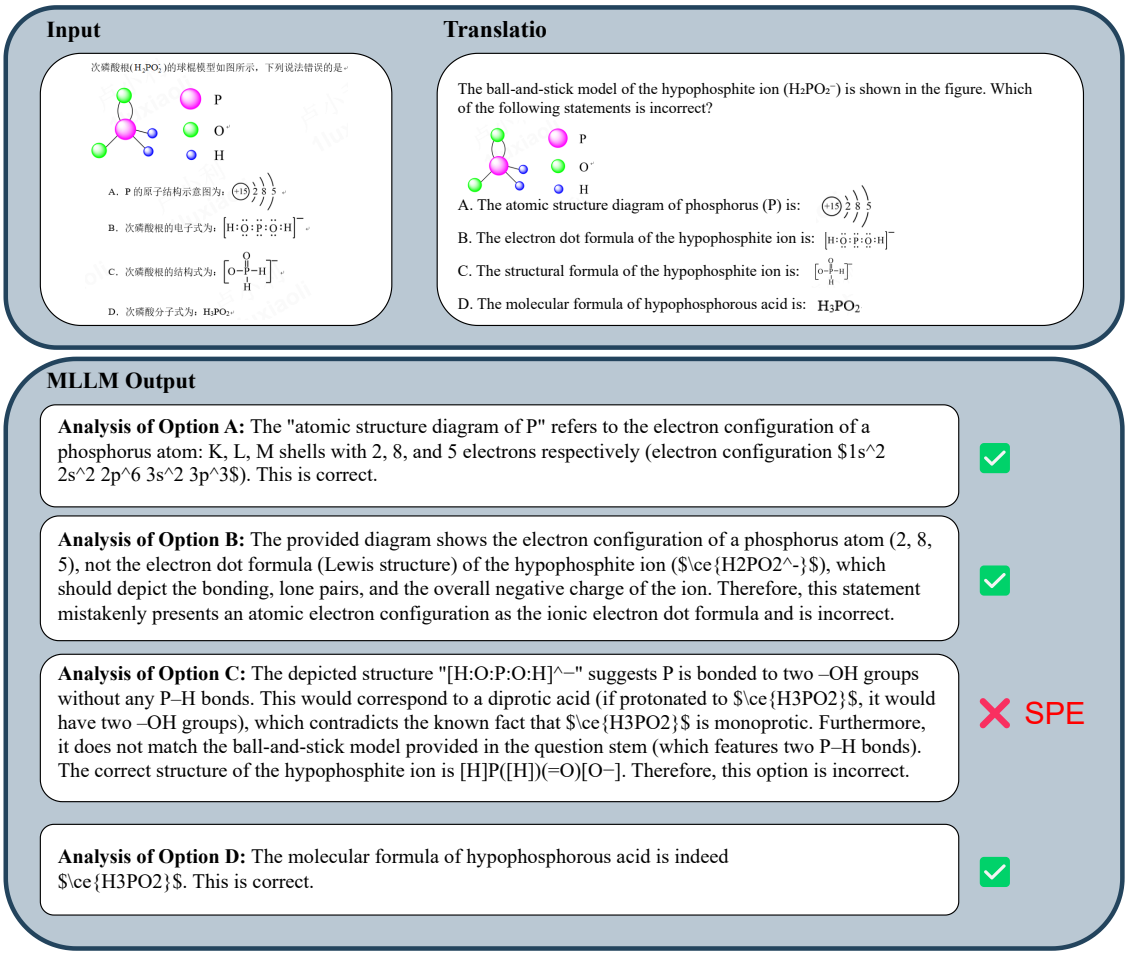

Figure 9. Case Study 2: Root-Cause Analysis The agent diagnosed the MLLM's answer as containing Symbolic Parsing Error (SPE).

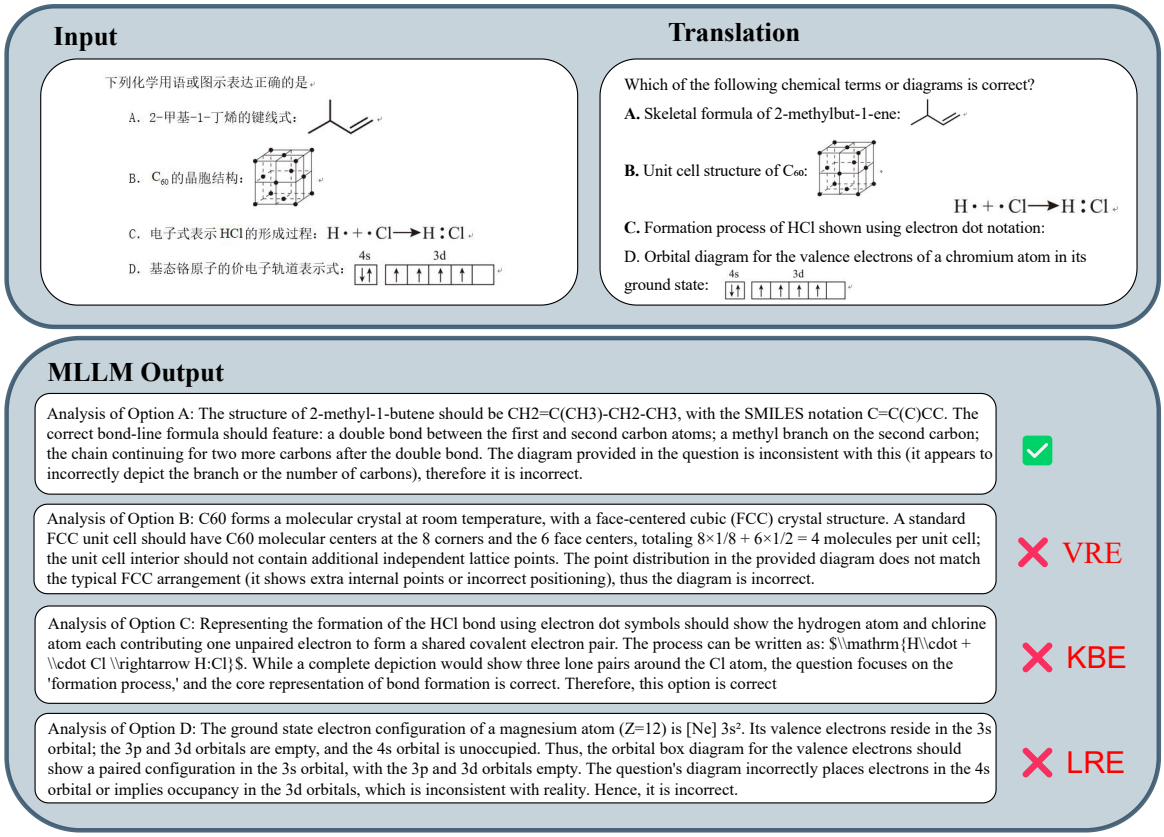

Figure 10. Case Study 3: Root-Cause Analysis The agent diagnosed the MLLM's answer as containing Visual Recognition Error (VRE), Knowledge-based Error (KBE) and Logical Reasoning Error (LRE).

Table S4. Base optimization of 1,2-aminoarylation of ethylene

entry	base	conv. of **1**/%	yield of **2**/%	yield of **3**/%	yield of **S1**/%
1	DIPEA	>99	<2	75	9
2	$NaHCO_3$	>99	<2	74	10
3	K_2CO_3	>99	<2	65	14
4	K_3PO_4	>99	<2	61	16
5	Et_3N	>99	<2	20	62
6	pyridine	>99	17	<2	56
7	DBU	>99	<2	2	78
8	DABCO	>99	<2	5	11(68)[a]
9	w/o base	>99	14	58	<2

[a]the yield in the bracket referred to the substitution product by DABCO

Table S12. Comparative analysis: experimental triplet energies vs. computed values for arenes and biphenyl halides at CPCM(acetone)-B3LYP-D3BJ/def2-TZVPP level of theory.

Compounds	Experimental E_T[18] (kcal•mol^{-1})	DFT calculated E_T (kcal•mol^{-1})	Difference (kcal•mol^{-1})
benzene	84.4	82.4	-2.0
anisole	80.8	77.5	-3.3
biphenyl	65.5	64.5	-1.0
biphenyl chloride	63.6	62.9	-0.7
biphenyl bromide	n.a.	62.9	-
biphenyl iodide	62.9	62.4	-0.5

n.a. not available

Figure 11. Figures in ChemTable. These two table indicate that most figures in ChemTable are made of text instead of structral chemistry.

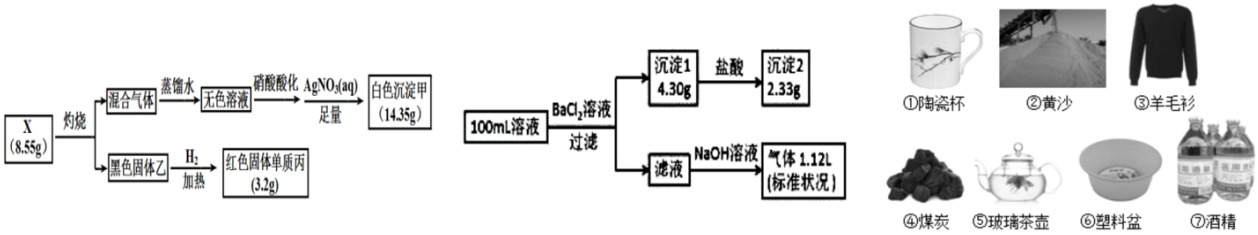

Figure 12. Figures in MMCR-Bench. A lot of figures of MMCR-Bench are too easy to examine MLLMs' abilities. These contains can be easily experssed by words and MLLMs' task center on OCR.

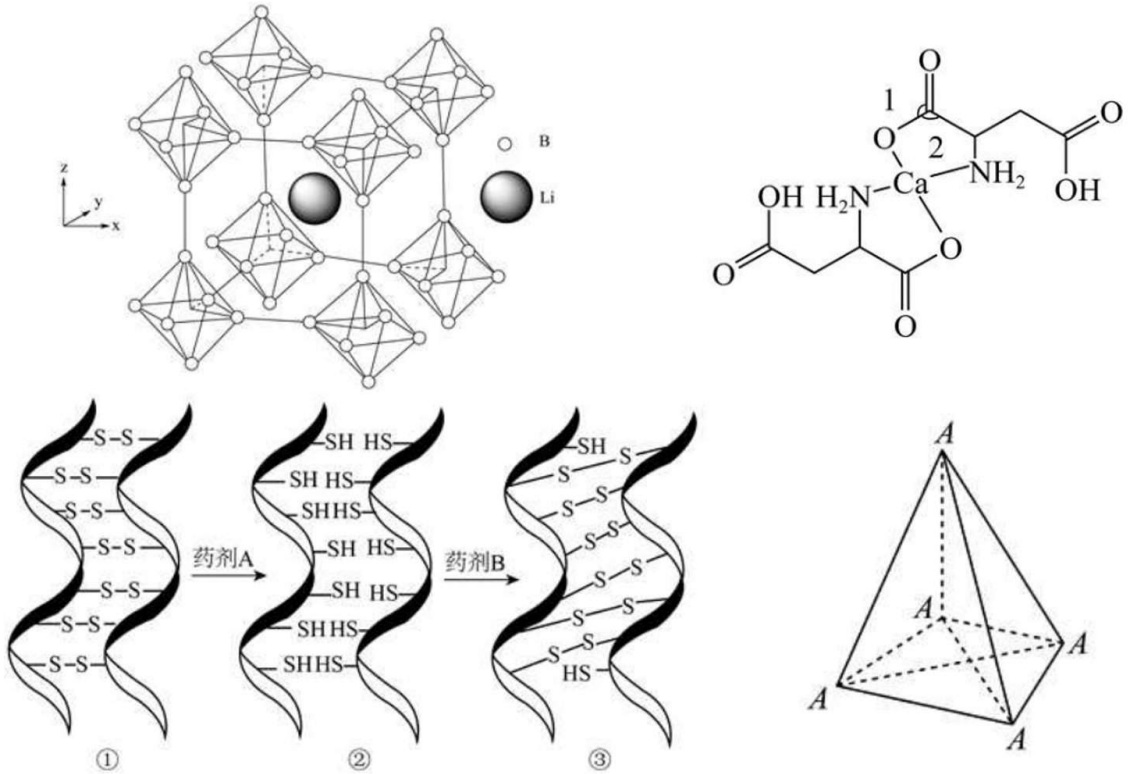

Figure 13. Figures in ChemVTS-Bench. These questions contain complex structure of chemistry that hard to express in words. They checkout models' sense of space.

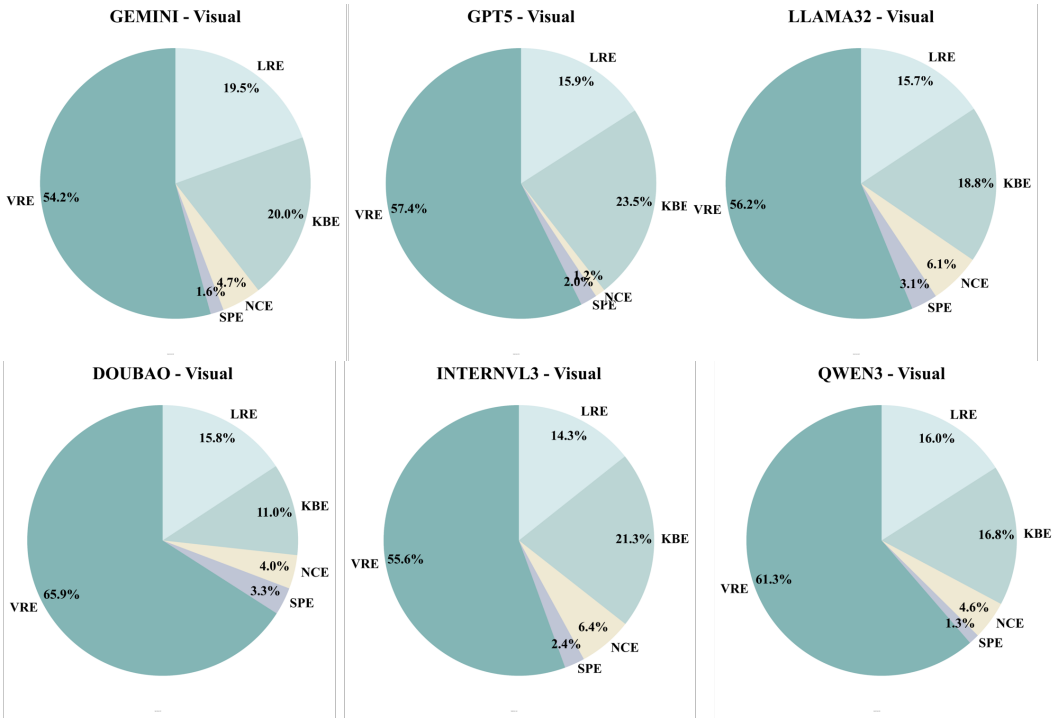

Figure 14. Error analysis of different MLLMs in mode visual

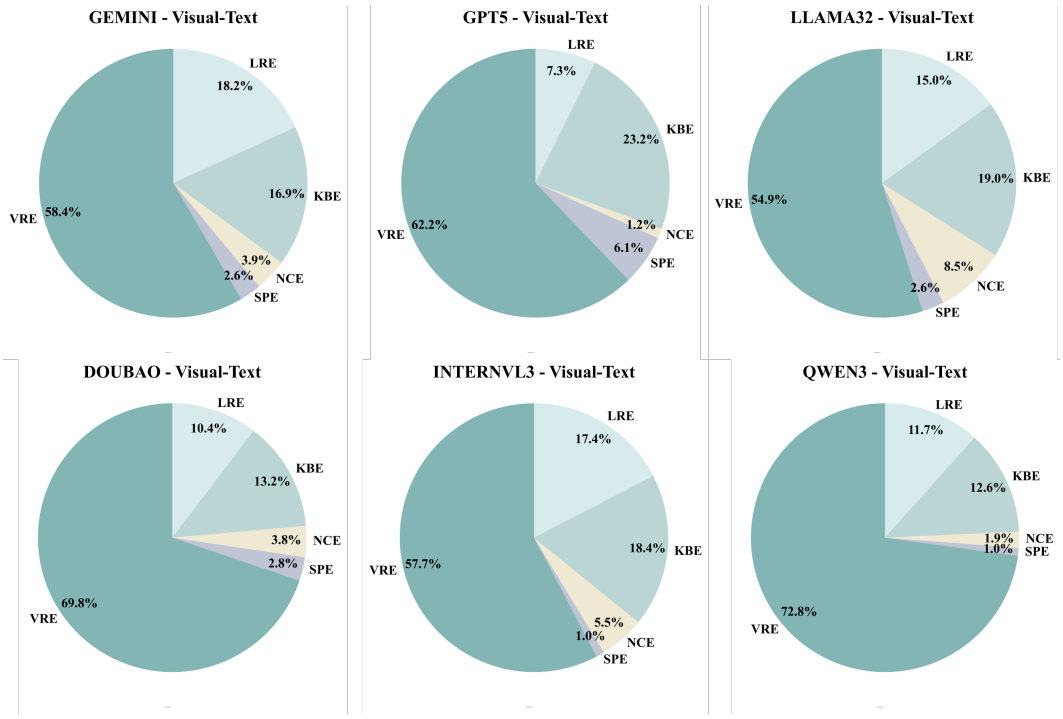

Figure 15. Error analysis of different MLLMs in mode visual-text