

A ABLATION STUDY

Temp	Qwen2.5 VL-32B				Qwen2.5 VL-72B				InternVL3-78B			
	Signal	Perception	Semantic	Generation	Signal	Perception	Semantic	Generation	Signal	Perception	Semantic	Generation
1.0	57.02	48.89	17.03	6.67	78.00	72.49	78.35	13.51	69.18	73.33	71.84	9.53
0.5	63.40	68.63	29.49	36.23	66.49	65.36	58.97	17.39	66.49	65.36	58.97	17.39
0.0	63.92	70.59	28.85	37.68	63.92	60.78	66.03	17.39	63.92	60.78	66.03	17.39

Table 4: Performance of three models under varying temperature settings.(Top- p fixed to 1)

Temperature Table 4 presents the impact of varying temperature settings on three models: Qwen2.5-VL-32B, Qwen2.5-VL-72B, and InternVL3-78B.

For Qwen2.5-VL-32B, lower temperatures ($T = 0.5$ and $T = 0$) yield substantial improvements over $T = 1.0$, particularly on the Perception, Semantic, and Generation levels. A similar trend is observed for InternVL3-78B, where deterministic decoding ($T = 0$ or $T = 0.5$) leads to a more balanced performance profile compared to the stochastic setting. In contrast, Qwen2.5-VL-72B behaves differently: while it achieves the highest Signal and Semantic scores at $T = 1.0$, its Generation accuracy remains relatively low across all settings.

These observations indicate that smaller models tend to benefit from reduced sampling variability, as lower temperatures enhance stability and reliability. Conversely, larger models may require higher temperatures to fully exploit their expressive capacity, though this comes at the cost of weaker generative consistency.

Top- p	Qwen2.5 VL-32B				Qwen2.5 VL-72B				InternVL3-78B			
	Signal	Perception	Semantic	Generation	Signal	Perception	Semantic	Generation	Signal	Perception	Semantic	Generation
1.0	57.02	48.89	17.03	6.67	78.00	72.49	78.35	13.51	69.18	73.33	71.84	9.53
0.5	63.92	69.93	29.49	39.13	62.89	62.09	65.38	13.04	70.62	68.63	72.44	5.80
0.1	63.40	66.67	30.77	44.93	66.49	54.25	64.74	23.19	69.07	69.93	72.44	11.59

Table 5: Performance of three models under varying Top- p settings.(temperature fixed to 1).

Top- p We further investigate the role of nucleus sampling while fixing the temperature to 1.0. The results in Table 5 show heterogeneous effects across models.

For Qwen2.5-VL-32B, reducing Top- p from 1.0 to 0.1 consistently improves Semantic and Generation scores, suggesting that constraining the sampling space mitigates low-quality outputs and enhances reliability. By contrast, Qwen2.5-VL-72B attains its best Signal and Semantic results at $p = 1.0$, but its Generation score is substantially reduced. Interestingly, setting $p = 0.1$ recovers part of this loss, implying a trade-off between precision and diversity.

For InternVL3-78B, performance remains comparatively stable across Top- p values, with minor fluctuations in Generation accuracy. This stability suggests that larger-scale models are less sensitive to sampling truncation, reflecting stronger intrinsic consistency.

Scoring Model The automatic evaluation at the Generation level uses a standardized prompt to guide the scoring models. This prompt, shown in the box below, instructs the evaluator to rate model answers on a scale from 0 to 1 based on specific rules.

Scoring Model	FP	IP	DnG
GPT-4o	23.33	0.00	5.26
InternVL3-78B	23.33	5.26	5.00

Table 6: Comparison of different scoring models (GPT-4o vs. InternVL3-78B) on Generation Level evaluation. FP, IP, and DnG represent different sub-categories.

Prompt Templates for OpenEvaluator

"You are an expert evaluator. Given the following question, reference answer, and model answer, please rate the model answer on a scale of 0 to 1, and explain your reasoning."

Scoring rules:

- If the reference answer is an image but the model output does not contain an image, score 0.
- If the reference answer is text but the model output does not contain text, score 0.
- Otherwise, score based on the similarity and correctness of the model output compared to the reference answer.
- If both text and image are present, consider both in your evaluation.

"Please output your score in the format: `\score{X}`, where X is a number between 0 and 1."

Question: f"Question: {question}"

Since the Generation level relies on automatic evaluation, we further compare different scoring models (Table 6). Both GPT-4o and InternVL3-78B yield consistent FP scores (23.33%). However, discrepancies arise in the IP dimension, where GPT-4o assigns zero while InternVL3-78B detects additional errors (5.26%). For DnG, the results are nearly identical (5.26% vs. 5.00%).

These findings suggest that overall evaluation trends are robust across scoring models, but fine-grained categories may be influenced by the evaluator’s internal biases. Consequently, careful selection of the scoring model is essential to ensure fairness and reliability in automatic evaluation pipelines.

B SPECTRUMBENCH DETAILED INFORMATION

B.1 SIGNAL LEVEL

This layer focuses on the direct processing and understanding of raw, fundamental data formats, much like extracting information from physical signals, as exemplified in Figure 6.

B.2 PERCEPTION LEVEL

This layer associates the features identified at the signal layer with chemical entities (functional groups, fragments, elements, and basic properties), as illustrated in Figure 7.

B.3 SEMANTIC LEVEL

This layer involves higher-level reasoning and comprehensive interpretation, connecting fragmented information to form complete insights or generate novel chemical structures, as depicted in Figure 8.

Examples	
Identifying the type of a spectrum, assessing its data quality, extracting basic features (e.g., peak position, peak intensity), and identifying impurity peaks..	
Sub-Category	Metadata
Spectrum Type Classification	Question: What type of spectrum is this? Choices & Answer: A. Infrared Spectrum (IR). B. Proton Nuclear Magnetic Resonance (H-NMR). C. Heteronuclear Single Quantum Coherence (HSQC). D. Raman Spectrum. Explanation: The spectrum uses ppm as units, which is a chemical shift unit specific to NMR. The chemical shift range typically falls between -2 ppm and 15 ppm, confirming this is a 1H NMR spectrum.
Spectrum Quality Assessment	Question: Does this spectrum show obvious signal quality issues? Choices & Answer: A. Yes. B. No, the signal is very clear. C. Localized noise. D. Very low noise, eligible. Explanation:...
Basic Feature Extraction	Question: Please select the chemical shift range corresponding to the most concentrated signal area in the HSQC spectrum. Choices & Answer: A. δ H 2-4 ppm, δ C 30-60 ppm. B. δ H 6-8 ppm, δ C 120-140 ppm. C. δ H 9-10 ppm, δ C 180-200 ppm. D. δ H 0-1 ppm, δ C 10-20 ppm. Explanation: HSQC spectrum plots ^1H chemical shift on the horizontal axis and ^{13}C on the vertical. Most signals cluster in the 2-4 ppm (^1H) and 30-60 ppm (^{13}C) region.
Impurity Peak Detection	Question: Please observe this spectrum carefully. Besides the signals from the target compound, there is also a distinct additional peak around 1 ppm in the image. What is this peak most likely? Choices & Answer: A. Solvent impurity. B. Target compound. C. Instrument noise. D. Reference standard. Explanation: In NMR spectrum, the peak near 1 ppm is often from impurities introduced during sample processing. Given it's an "extra" signal not part of the target compound, it's likely an impurity.

Figure 6: Example tasks and question formats at the Signal Level.

B.4 GENERATION LEVEL

This layer focuses on creating novel data, such as generating a 2D image of a molecule from its SMILES string, predicting the Mass Spectrum for a given chemical structure, or designing a new molecule with specific properties, as illustrated in Figure 10.

B.5 DATA DISTRIBUTION

To provide an overview of the data landscape, Figure 9 presents two pie charts: the left illustrates the distribution of different spectrum types (e.g., NMR, IR), while the right shows the categorization of spectroscopic task types. These distributions reflect the diversity of data and tasks within our study. It should be noted that the spectrum type statistics were generated by having GPT-4o scan and summarize all spectra in the benchmark. However, there are potential limitations: GPT may

Examples	
Identifying functional groups like -OH from a mass spectrum; determining the presence of isotopes like ^{13}C ; assigning a ^1H NMR triplet to a methyl group; predicting molecular weight from a mass spectrum.	
Sub-Category	Metadata
Basic Property Prediction	<p>Question: What Given the mass spectrum image, what is the most likely molecular ion peak (m/z) observed for this compound?</p> <p>Choices & Answer: A. 85. B. 107. C. 120. D. 150.</p> <p>Explanation: The strongest peak at m/z 107.0 is the molecular ion (M^+), with an adjacent m/z 109.0 peak ($\sim 1/3$ intensity) indicating one chlorine atom ($^{35}\text{Cl}/^{37}\text{Cl} \approx 3:1$). Smaller peaks ($m/z$ 93.0, 108.0) are fragments.</p>
Elemental Composition Prediction	<p>Question: Observe the provided mass spectrum image. The significant $M+2$ peak suggests the presence of which element?</p> <p>Choices & Answer: A. Fluorine (F). B. Chlorine (Cl). C. Bromine (Br). D. Iodine (I).</p> <p>Explanation: The intensity ratio of the m/z 51 and 53 peaks ($\sim 3:1$) reflects chlorine's natural isotopes, ^{35}Cl (75.77%) and ^{37}Cl (24.23%), giving an $M+2$ peak about one-third the main peak.</p>
Functional Group Recognition	<p>Question: Based on this infrared spectrum, what functional group is most likely present in the molecule?</p> <p>Choices & Answer: A. Carbonyl group ($\text{C}=\text{O}$). B. Hydroxyl group ($-\text{OH}$). C. Amino group ($-\text{NH}_2$). D. Nitro group ($-\text{NO}_2$).</p> <p>Explanation: In the infrared spectrum, a pair of sharp absorption peaks around 3300 cm^{-1} are typical of the symmetric and asymmetric N-H stretching vibrations in a primary amino group ($-\text{NH}_2$).</p>
Peak Assignment	<p>Question: Given the chemical formula $\text{C}_6\text{H}_5\text{F}$. Observe this H-NMR spectrum. The singlet peak around ~ 7.3 ppm in the image is most likely assigned to which part of the molecule?</p> <p>Choices & Answer: A. Methyl group. B. Fluoro-substituted carbon. C. Aromatic ring protons. D. Alkene protons</p> <p>Explanation: The 7.3 ppm shift is typical for aromatic protons in fluorobenzene ($\text{C}_6\text{H}_5\text{F}$). Though misdescribed as a singlet, it's a complex multiplet from H-H and H-F coupling, with the shift confirming its aromatic nature.</p>

Figure 7: Example tasks and question formats at the perception level.

have recognition errors, and some spectrum-involving benchmarks lack actual image data (e.g., predicting NMR spectrum properties from molecular characteristics in *de novo* generation tasks). Additionally, in tasks like multimodal fusion reasoning and forward generation problems, a single benchmark instance might include multiple spectra. Thus, the number of spectra does not align with the number of benchmarks, and this pie chart is provided only as a general reference.

Examples	
Elucidating a complete molecular structure from one or more spectra; verifying a proposed structure against spectral data; and reasoning across different modalities (e.g., text and spectrum) to answer complex questions.	
Sub-Category	Metadata
Fusing Spectroscopic Modalities	<p>Question: The molecular formula of the compound is C₆H₁₁NO. Use this information together with the provided IR spectrum to infer possible structural features.</p> <p>Choices & Answer: A. Amide. B. Alcohol. C. Ester. D. Alkene.</p> <p>Explanation: Infrared spectroscopy shows a strong 1650 cm⁻¹ peak (C=O) and a 3300–3500 cm⁻¹ peak (N–H). Their coexistence, along with N and O in the formula, clearly indicates an amide group.</p>
Molecular Structure Elucidation	<p>Question: Given the mass spectrum of an unknown compound with a molecular formula C₁₁H₁₆, predict the most likely molecular structure (SMILES) consistent with the observed fragments.</p> <p>Choices & Answer: A. CC(C)=C1C=CC=C1. B. CC(C)CC1=CC=CC2=CC=CC=C12. C. CC(C)(C)CC1=CC=CC=C1. D. CCC(C)C1=CC=CC2=CC=CC=C12.</p> <p>Explanation: The base peak at m/z 91 indicates a benzyl (C₆H₅CH₂–) structure, while m/z 133 represents loss of a methyl group. Only CC(C)(C)CC1=CC=CC=C1 fits both fragmentations.</p>
Multimodal Molecular Reasoning	<p>Question: The Raman spectrum of the molecule OC1CCCC1=O (2-hydroxycyclopentanone) shows a series of strong peaks in the 2800–3000 cm⁻¹ region. These peaks are most likely attributed to which type of molecular vibration?</p> <p>Choices & Answer: A. C–H stretching. B. O–H stretching. C. C=O stretching. D. N–H stretching.</p> <p>Explanation: In Raman spectroscopy, 2800–3000 cm⁻¹ is characteristic of C–H stretching. The strong peak here arises from cycloalkane C–H vibrations, while O–H (3200–3600 cm⁻¹) and C=O (~1700 cm⁻¹) peaks are absent.</p>

Figure 8: Example tasks and question formats at the semantic level.

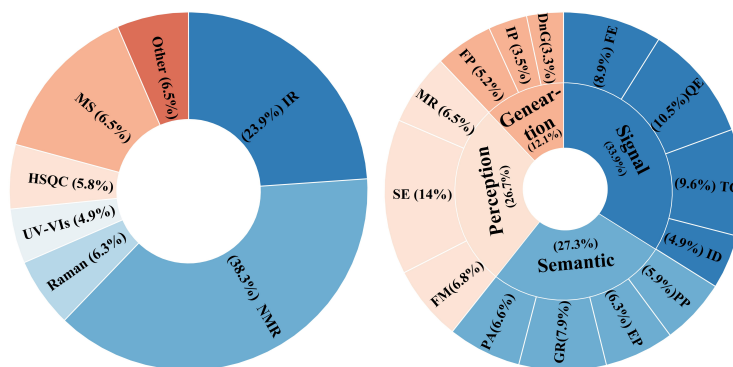


Figure 9: Distribution of spectrum types and spectroscopic task categories.

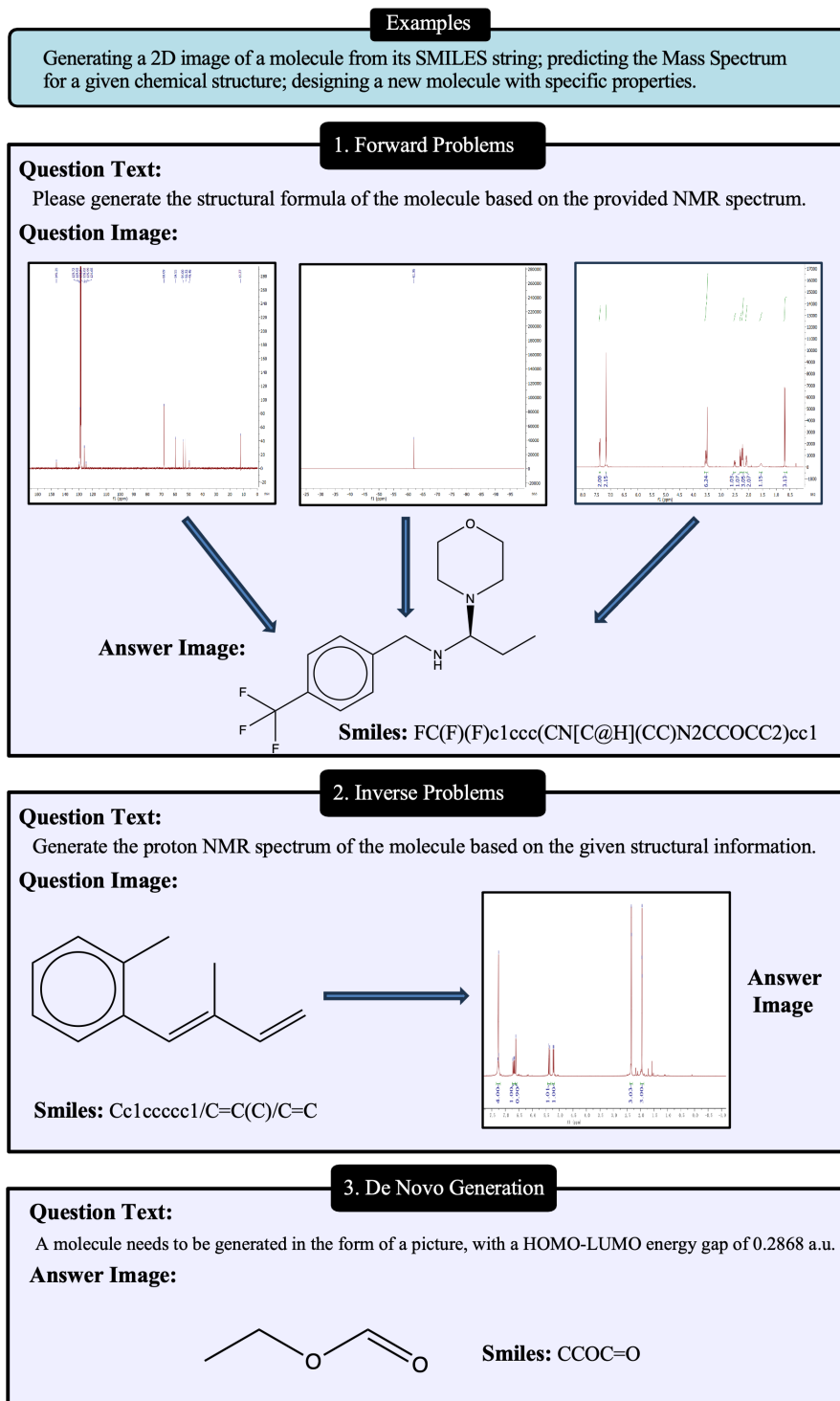


Figure 10: Example tasks and question formats at the Generation Level.

C SPECTRUMANNOTATOR TECHNICAL DETAILS

In the main text, we briefly introduced the function of SpectrumAnnotator. In this section, we will introduce its specific technical details.

MolPuzzle (Guo et al., 2024b) represents the first benchmark specifically designed for LLMs in spectroscopic analysis, employing a three-stage approach to generate question-answer pairs. While this template-based generation method offers efficiency, it suffers from limited coverage of spectroscopic domains and overly simplistic question formats. In the field of spectroscopy, high-quality data and benchmarks are crucial to advance AI research. The design of SpectrumAnnotator originates from two key insights: First, the process of creating benchmarks shares similarities with the supervised data generation methods used in LLM pre-training and post-processing. Just as high-quality training data is essential for model performance, well-designed benchmarks are equally critical for evaluating and advancing the field. Second, we aim to utilize LLMs’ few-shot and zero-shot capabilities to generate diverse benchmarks, enabling batch processing of seed datasets to construct large-scale pre-training and post-processing data. Additionally, we leverage LLMs’ discriminative abilities for preliminary data screening and establish closed-loop mechanisms for continuous improvement.

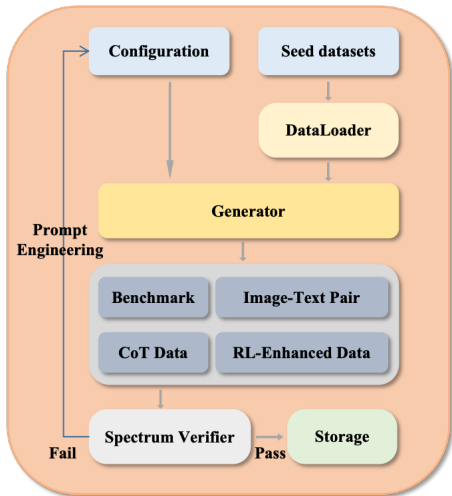


Figure 11: Technical architecture of SpectrumAnnotator, illustrating the data flow from seed datasets through generation to quality verification.

As illustrated in Figure 11, SpectrumAnnotator consists of several key components that work together to generate high-quality spectroscopic benchmarks. **Configuration & Seed Datasets** form the foundation of the system. Seed datasets are extracted from multiple data sources containing essential spectroscopic information, while the configuration is a YAML configuration file that primarily configures prompt templates, instructing the generator on what prompts to use, along with model configurations and other parameters. As shown in Figure 12, taking property prediction as an example, the configuration specifies the seed datasets from MolPuzzle and provides question templates to guide the generator’s output.

DataLoader addresses the challenge of integrating diverse data sources. Ideally, we would like to standardize all seed datasets into a uniform format. However, in practice, this proves challenging as original data may possess complex nested file structures and diverse storage formats. To reduce adaptation complexity, we allow customized DataLoader designs. This design is inspired by PyTorch’s DataLoader, which can properly load, batch, and post-process raw data. Our DataLoader aims to integrate various “seed datasets” into formats that can be processed by generators. The foundation consists of two base classes: *DataSample*, which represents the minimal granular information unit in SpectrumAnnotator and serves as reference information for the Generator to generate individual samples; and *Dataset*, a collection of *DataSample* objects that provides standardized access methods. As demonstrated in Figure 13, the DataLoader adopts a plugin-based architecture with an abstract registry. For different seed datasets, researchers only need to register their custom loaders using simple registration code, enabling seamless integration of diverse data sources.

Generator operates through a three-stage workflow: First, it receives question templates from Configuration (including few-shot examples). Second, for each sample in the seed dataset, the generator uses question templates combined with sample metadata (such as molecular formulas, spectrum

Prompt Template

```

dataset_path: seed_datasets/molpuzzle/meta_data.json
generator_type: benchmark
model_type: internvl
num_samples: 30
output_path: output/benchmark.json
property_pred:
  template: |
    This is a property prediction task. You are given the following spectrum(s) for a molecule: {spectra_list}
    You are provided with one or more spectrum images (such as MS, IR, NMR, UV-Vis, etc.) for a molecule.
    Your task is to create a question and answer that require the examinee to infer a chemical or physical property
    of the compound based on the features visible in the provided spectrum image(s).
    Typical properties include: molecular weight (from mass spectrum), acidity (from IR), color (from UV-Vis),
    solubility (from NMR), and so on. Please focus on designing questions that require the examinee to deduce such
    properties from the actual spectral features shown in the image(s).
    Here are some example questions for your reference. Please output your result in the following JSON format
    (making sure image_path points to the actual spectrum image and is not empty):
    Example 1:
    {{
      "selected_spectrum_type": "MASS",
      "question": "Given the mass spectrum image, what is the most likely molecular ion peak (m/z) observed for
      this compound?",
      "choices": ["85", "90", "120", "150"],
      "answer": "120",
    }}
    Example 2:
    {{
      "selected_spectrum_type": "IR",
      "question": "Given the infrared spectrum image, a very broad absorption band in the 2500–3300 cm-1 range
      and a strong peak around 1700 cm-1 are observed. What property does this most likely suggest for the compound?",
      "choices": ["Strong acidity", "High basicity", "Aromaticity", "Aliphatic character"],
      "answer": "Strong acidity",
    }}

```

Figure 12: Example configuration for property prediction tasks, demonstrating how prompt templates and model parameters are specified.

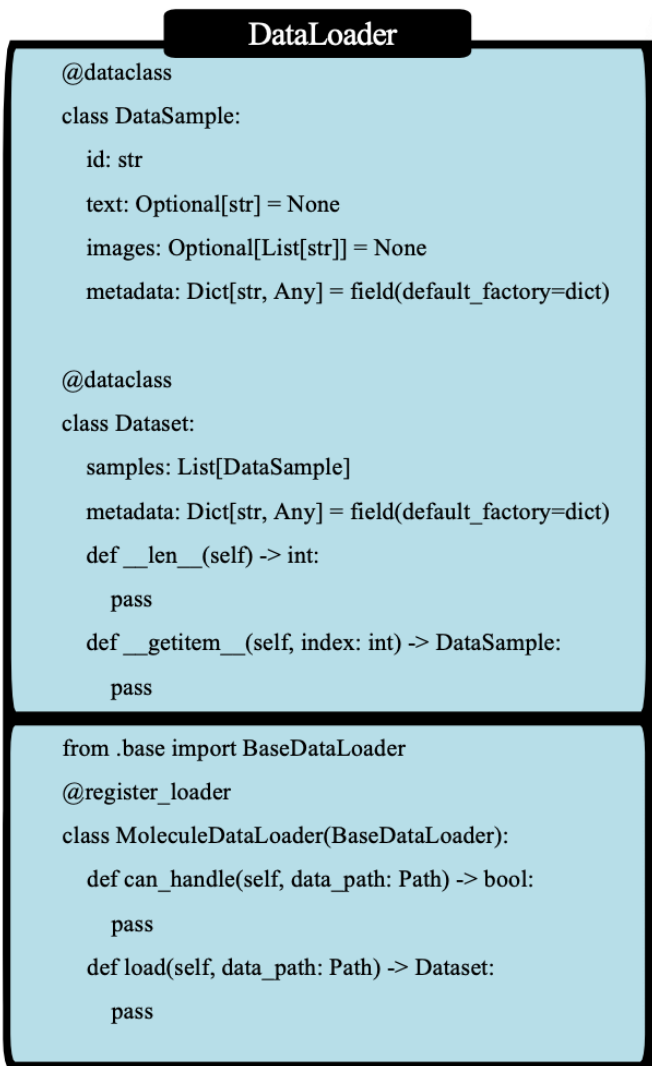


Figure 13: Plugin-based DataLoader architecture showing the registration mechanism for custom data loaders.

paths, SMILES strings, etc.) to render a prompt, which is then passed to the large language model. Third, the model’s output is parsed into standard formats (e.g., question/choices/answer).

Quality Assurance Pipeline ensures the reliability of generated benchmarks. After data generation, the system employs a multi-stage quality assurance process: Initial screening using rule-based methods to check data format and remove non-compliant samples, followed by SpectrumVerifier, a large model-based verification system that identifies suspicious samples requiring manual annotation. This closed-loop mechanism ensures that only high-quality, scientifically valid benchmarks are included in the final dataset. SpectrumAnnotator will be open-sourced to collaborate with the research community in building a robust ecosystem and collectively addressing challenges in spectroscopic data generation and curation.

D BENCHMARKING CANDIDATES

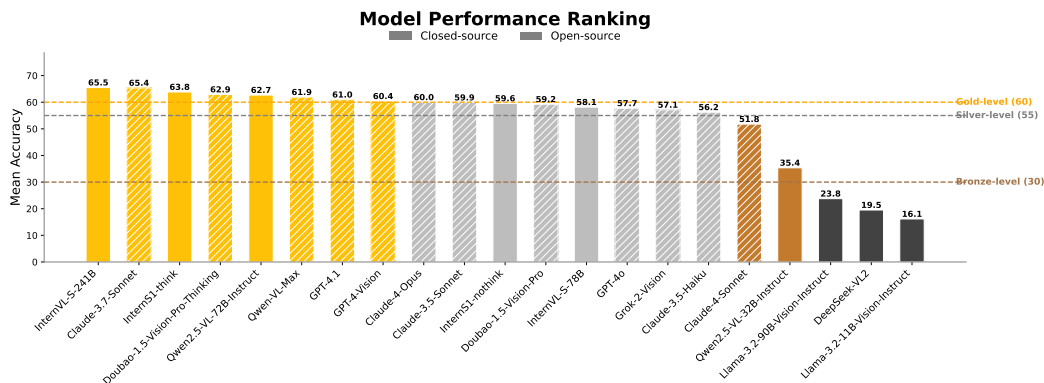


Figure 14: Performance ranking of various LLMs.

D.1 OPEN-SOURCE MODELS

Qwen2.5-VL-32B-Instruct(Bai et al., 2025). Alibaba’s open-source Vision-Language multimodal large model that handles reasoning and generation for images, text and video. It employs a hierarchical tagging architecture, supports multi-turn conversations and complex reasoning, and both the model weights and code are publicly available.

Qwen2.5-VL-72B-Instruct(Bai et al., 2025). Qwen2.5’s larger-scale model enhances cross-modal reasoning and instruction-following capabilities, delivering superior performance on benchmarks such as MMMU and M3Exam while supporting multitasking and multilingual inputs - and is completely open-source.

InternVL3-78B (Chen et al., 2024b). Shanghai AI Lab releases the multimodal model, combining native multimodal pre-training, variable visual position encoding (V2PE), MPO, and test-time scaling to approach GPT-4o performance.

Llama-3.2-11b-Vision-Instruct(Meta AI, 2024). Meta’s 11 B lightweight multimodal model locks Llama-3.1 8 B text and pairs it with a ViT encoder. Two-stage training: image-text alignment then SFT+DPO, using RoPE-2D. Open-source.

Llama-3.2-90b-Vision-Instruct(Meta AI, 2024). The 90B features a more advanced vision adapter with cross-attention layers to inject image features into the LLM core. It is tuned with SFT and RLHF for enhanced performance on complex visual reasoning tasks.

DeepSeek-VL-2(Wu et al., 2024). An open-source model from DeepSeek-AI featuring a Mixture-of-Experts (MoE) backbone and a dynamic tiling vision encoder for high-resolution images. It achieves or exceeds the state-of-the-art performance at the time on benchmarks like MMMU and DocVQA, with its code and weights fully available on GitHub.

Doubao-1.5-Vision-Pro (Doubao Team, 2025). It features a dynamic resolution visual encoder and MoE architecture, supporting visual QA, text-image matching, and image description. With billions of parameters, it shows strong generalization across scenarios and is available for self-hosting and fine-tuning.

Doubao-1.5-Vision-Pro-Thinking (Doubao Team, 2025). It integrates a “Deep Thinking Mode” and is trained with multi-round Reward Learning and reasoning style training. It excels in scientific, mathematical, and chain-of-thought reasoning. Supports open-source calling and API integration.

GLM-4.5V(VTeam, 2025). An open-source vision-language model from Zhipu AI and Tsinghua University that introduces a versatile “thinking paradigm” for enhanced reasoning. It leverages scalable reinforcement learning and supports full-spectrum vision reasoning, including GUI agent operations and code generation from screenshots.

InternS1(Intern-S1 Team, 2025). A vision-language model developed by Shanghai AI Laboratory that features a specialized “Thinking” mode for enhanced multi-step reasoning. This mode allows the model to perform a series of self-guided logical steps to solve complex problems, particularly in scientific, mathematical, and logical domains.

D.2 CLOSED-SOURCE MODELS

GPT-4o (OpenAI, 2024). OpenAI’s flagship “omni” model natively supports text, audio, and image modalities. Delivers GPT-4-level intelligence with significantly faster response times and enhanced multimodal capabilities.

GPT-4.1(OpenAI, 2025). A reinforced version of GPT-4 deployed through the OpenAI API, offering improved handling of complex instructions and logical reasoning; accepts multimodal inputs but is primarily geared toward text-centric tasks.

GPT-4-Vision(OpenAI, 2023). A version of GPT-4 equipped with image input capabilities, optimized for understanding images and text and for the generation of conversational content, widely used for image-based Q&A.

Claude-3.5-Haiku. Anthropic’s fastest and most cost-effective model in the Claude3.5 family—offers very low latency, strong coding and reasoning ability, and often exceeds Claude Opus on intelligence benchmarks despite being lightweight.

Claude-3.5-Sonnet (Anthropic, 2024). Anthropic’s multimodal large language model has mixed inference capabilities and powerful visual understanding functions. It supports a context of 200K tokens and is skilled in natural writing and code generation.

Claude-3.7-Sonnet (Anthropic, 2025a). An evolution of Claude3.5 Sonnet that introduces hybrid reasoning—users can choose between fast modes or step-by-step logical chains; offers strong task flexibility, extended context windows, and deep instruction-following in multimodal settings.

Claude-4-Opus (Anthropic, 2025b). Anthropic’s flagship model, designed for complex tasks. It boasts a powerful memory architecture and parallel tool invocation capabilities, and integrates with Claude Code, performing exceptionally in coding and reasoning benchmark tests.

Claude-4-Sonnet (Anthropic, 2025b). Claude-3.7-Sonnet’s successor, balancing performance and speed, with low latency and high resource efficiency, excels in code generation.

Grok-2-Vision(xAI, 2024). The multi-modal model of xAI combines language and visual processing capabilities to handle various images and documents, and supports multilingual recognition and style analysis.

Qwen-VL-Max. The closed-source flagship model of Alibaba’s Qwen series has been optimized for deployment in enterprise-level multimodal tasks, supporting joint input of images, text, videos, and others, with ultra-large parameter volume and high inference capability.

Gemini-2.5-Pro(Gemini Team). A multimodal model from Google DeepMind that achieves state-of-the-art performance on frontier reasoning and coding benchmarks. It excels at multimodal understanding, including the ability to process up to 3 hours of video content and convert it into interactive code. Its combination of long context, multimodality, and enhanced reasoning capabilities unlocks new agentic workflows and complex problem-solving.

E ERROR CASES STUDY

E.1 SIGNAL LEVEL

We observe that the model struggles to distinguish localized noise from clean signals in the spectrum quality assessment task. For example, given the question “Does this spectrum show obvious signal quality issues?”, the ground-truth label was “Localized noise” or “Very low noise, eligible”, indicating minor but noticeable signal interference. However, the model incorrectly predicted “No, the signal is very clear”, resulting in a failed case. This misclassification reveals a key limitation: the model tends to overestimate the clarity of the spectrum when the noise is not global or strongly pronounced. In visual inspection, localized artifacts—though subtle—can be clearly identified by human annotators, whereas the model often dismisses them as negligible. It lacks sufficient sensitivity to weak or local signal distortions, or has overfit to globally noisy or clean examples during training, causing it to ignore partial imperfections. This insight aligns with our general observation: the model often fails to distinguish noise from true signal, especially when the noise is spatially sparse or located at the margins of the image. Such behavior may stem from the fact that the model treats the entire spectrum as a holistic input, and lacks mechanisms to perform fine-grained regional quality assessment. Additionally, for models not inherently multi-modal, spectra are often encoded as image representations and then passed through vision encoders or captioning modules, potentially discarding low-level noise patterns. As a result, noise may not be retained in the model’s internal representation, leading to overly optimistic predictions.

E.2 PERCEPTION LEVEL

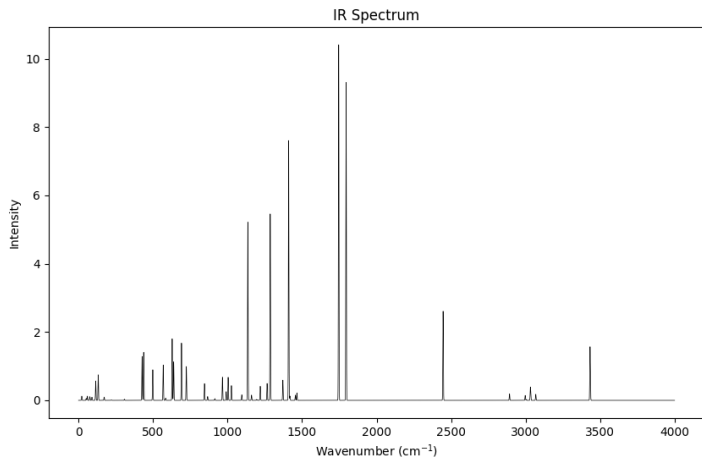


Figure 15: A Case of Functional Group Recognition

We found that for functional group recognition and peak assignment tasks, large language models such as Doubao-1.5-pro-thinking often fail to produce chemically accurate predictions, even when the visual features in the spectra are clear to human experts. For instance, in the functional group recognition task (Figure 15), the infrared (IR) spectrum exhibits a strong absorption band characteristic of a **carbonyl group (C=O)**, typically near 1700 cm^{-1} . However, the model incorrectly predicted **hydroxyl group (-OH)**. This suggests that the model likely over-relied on the presence of a broad peak or baseline shift, possibly mistaking low-intensity or overlapping signals for OH-stretching vibrations. In the peak assignment task (Figure 16), given the molecular formula $\text{C}_{10}\text{H}_7\text{Cl}$ and a clear singlet near 6.8 ppm in the ^1H -NMR spectrum, the expected answer was **aromatic CH next to a double bond**, i.e., a non-substituted position in the naphthalene ring. Yet the model responded with **aromatic CH adjacent to Cl**, a chemically invalid assignment considering the splitting pattern and electronic environment. This indicates a lack of fine-grained chemical reasoning and possibly an overemphasis on token-level keyword association rather than structural context. These cases expose the model’s semantic-level misunderstanding, which goes beyond visual misinterpretation.

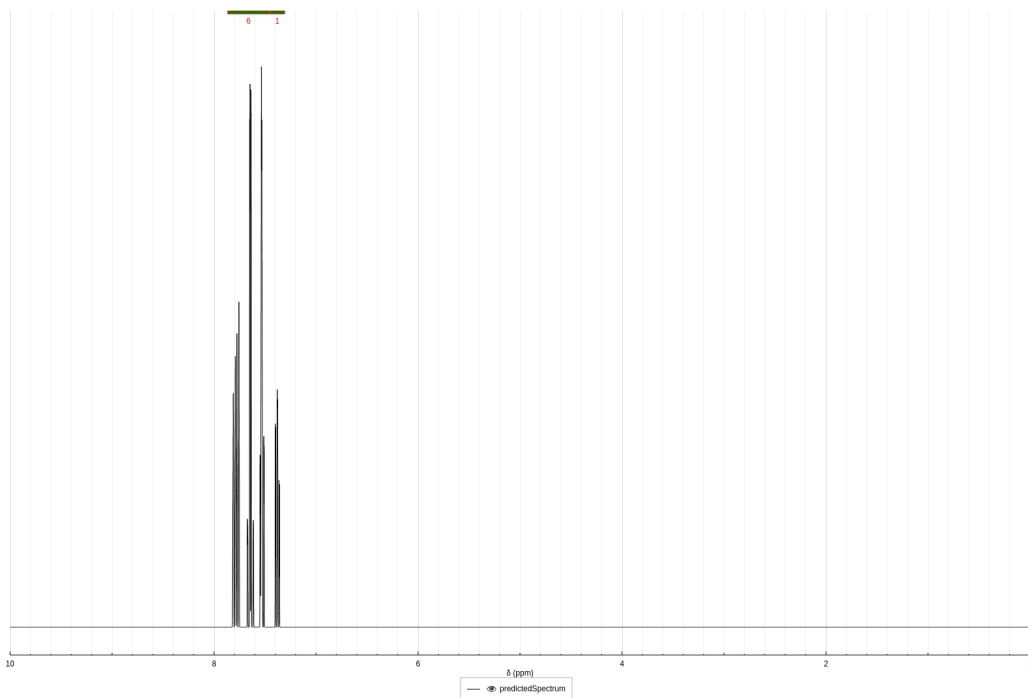


Figure 16: A Case of Peak Assignment

tion and highlights a deficiency in chemically grounded reasoning. We hypothesize two contributing factors. Firstly, the model may rely heavily on language priors, rather than truly integrating spectral visual features with molecular structure. Secondly, it lacks domain-specific supervision. Pretraining on generic data may not sufficiently expose the model to physical rules of spectroscopy, such as electron-withdrawing effects, chemical shift theory, or group frequency ranges.

E.3 SEMANTIC LEVEL

At the semantic level, tasks involving **molecular structure elucidation** and **multi-modal reasoning** remain particularly challenging. Consider the example below:

In this case, the model is asked: “The molecular formula of the compound is $C_4H_8O_2$. Use this information together with the provided IR spectrum image to infer possible structural features.” The correct answer should be **Ether**, based on the absence of a strong carbonyl absorption near 1700 cm^{-1} and the elemental composition. However, the model incorrectly predicts **Carboxylic acid**, likely due to over-reliance on superficial signal patterns that resemble O–H stretching or C=O bands.

Even when the molecular formula is omitted (pure spectrum-based reasoning), the model continues to produce incorrect predictions, revealing a deficiency in cross-modal semantic alignment. This suggests that while LLMs may perform well on shallow text-image associations, they struggle with integrating spectral data and chemical constraints in a chemically meaningful way.

E.4 GENERATION LEVEL

Not surprisingly, the performance on generation tasks—especially structure generation—is significantly worse. This suggests that while models like **Claude-3.7-Sonnet** perform well on earlier levels such as perception, syntactic understanding, and basic semantic reasoning, they still struggle with more complex **forward problems** that require inferring new molecular structures from spectral data. **De novo generation** and **inverse problems** (e.g., predicting spectra from structure) pose even greater challenges, as they demand deeper chemical understanding and cross-modal generalization.

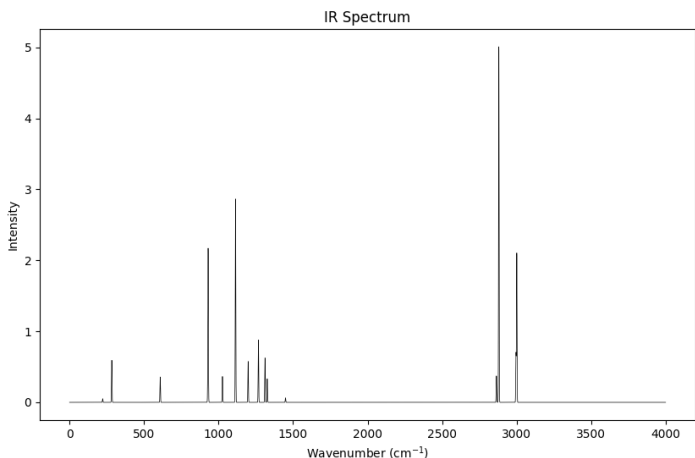


Figure 17: A Case of Fusing Multi-Modalities

In these settings, most models exhibit clear signs of overfitting or default to high-frequency patterns seen in training data.

Surprisingly, **Doubao-1.5-Vision-Pro-Thinking** demonstrates promising performance on forward problems, aligning well with its strong results in earlier semantic-level tasks such as functional group recognition, peak assignment, and molecular structure elucidation. This consistency suggests that the model may have a better internal representation of cross-modal chemical semantics, though its capability still falls short in full generation settings.

F MODEL ACCURACY VS. TOKEN ASSUMPTIONS

We conduct a comparative analysis of several Multimodal Large Language Models (MLLMs) from both semantic and generative levels, focusing on three representative tasks: Molecule Elucidation (ME), Fusing Spectroscopic Modalities (FM), and Forward Problems (FP). As shown in Figure 18, the performance gap among models is significant. Notably, models with lower average token assumptions, such as *DeepSeek-VL2*, tend to exhibit lower accuracy. In contrast, models with higher token assumptions, such as *Doubao-1.5-Vision-Pro-Thinking*, achieve superior performance, especially on complex *de novo* generation tasks like FP. This suggests that a longer reasoning chain, reflected in higher token usage, benefits complex problem-solving. However, the trade-off is increased computational cost and significantly longer inference time. These results highlight the efficiency-performance dilemma in MLLMs.

G DETAILED DATA STRUCTURE

This section details the comprehensive seed datasets curation pipeline and the three primary data structures that underpin our framework: the foundational **seed datasets**, the structured **benchmark data**, and the standardized **evaluation results**.

G.1 SEED DATA CURATION DETAILS

The seed datasets are curated from three primary sources to ensure both diversity and scientific rigor:

1. **Proprietary collections and in-house experimental data:** These include unpublished spectroscopic measurements and curated datasets from our collaborating laboratories. This source comprises approximately 238,869 molecular data points covering 8 types of spectra, offering higher authenticity and usability compared to most computationally generated spectra.

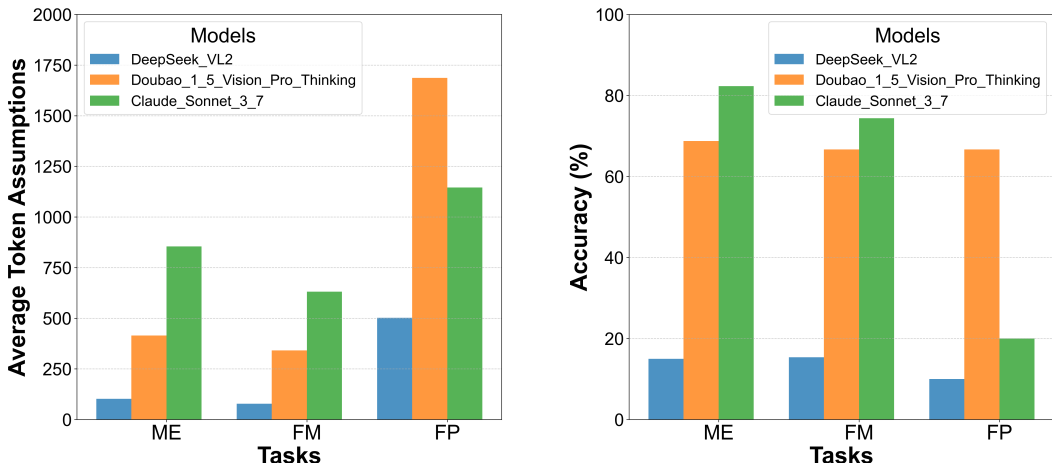


Figure 18: Model accuracy aligns with the model size.

- Public repositories and benchmark datasets:** We integrate data from a range of widely recognized and authoritative sources, including SDBS (of Advanced Industrial Science & , AIST), QM9S (Zou et al., 2023), NovoBench (Zhou et al., 2024a), and MolPuzzle (Guo et al., 2024b), among others. In total, seven distinct repositories and public datasets are used, collectively encompassing over 1.01 million unique chemical compounds.
- Literature mining:** Spectral data are systematically extracted from the *Supporting Information* sections of peer-reviewed publications, with a focus on articles from leading journals such as the *Journal of the American Chemical Society (JACS)* and *ACS Catalysis*.

All collected datasets undergo a unified processing pipeline that systematically maps each entry into three core chemical spaces: SMILES, molecular formula, and spectra. The resulting seed datasets are organized at the level of individual chemical substances, with each record containing the compound’s SMILES, molecular formula, and a structured set of associated spectra, all stored in a standardized JSON format. This robust foundation facilitates downstream annotation and interoperability.

G.2 SEED DATASETS STRUCTURE

The seed dataset is constructed by extracting essential information from raw experimental data, serving as the foundation for benchmark generation. Each entry contains a molecular index, SMILES string, molecular formula, and a list of associated spectra. An illustrative structure is provided in Listing 1. The `path` field is a list that may contain multiple files for a given spectrum type, accommodating cases such as multiple mass spectra for a single molecule.

Listing 1: Example structure of a seed dataset entry.

```
{
  "molecule_index": "MOL_0001",
  "smiles": "CCCCC1=CC=CC=C1",
  "formula": "C10H14",
  "spectra": [
    {"spectrum_type": "IR", "path": ["IR/MOL_0001.png"]},
    {"spectrum_type": "MASS", "path": ["MASS/MOL_0001.jpg",
    "MASS/MOL_0001_2.jpg"]},
    {"spectrum_type": "C-NMR", "path": ["C-NMR/MOL_0001.png"]},
    {"spectrum_type": "H-NMR", "path": ["H-NMR/MOL_0001.png"]}
  ]
}
```

G.3 SPECTRUMBENCH DATA STRUCTURE

The benchmark data structure is designed to support a diverse range of tasks, including signal interpretation, perception, and semantic understanding. Each entry includes a unique identifier, image path(s), question, answer choices, ground truth answer, category, sub-category, data source, and timestamp. A representative example is shown in Listing 2. After processing by SpectrumLab, three additional fields are appended: `model_response` (the model’s reasoning and output), `model_prediction` (the answer extracted from the model response), and `pass` (a boolean indicating whether the model’s prediction matches the ground truth).

Listing 2: Example of a benchmark data entry.

```
{
  "id": "Perception_a9cf_250723_235951_318294",
  "image_path": [
    "data/Perception/Basic Property Prediction/Perception_a9cf_q.png"
  ],
  "question": "Given the mass spectrum image, what is the most likely
molecular ion peak (m/z) observed for this compound?",
  "choices": ["85", "90", "120", "133"],
  "answer": "133",
  "category": "Perception",
  "sub_category": "Basic Property Prediction",
  "source": "",
  "timestamp": "2025-07-23 23:59:51"
}
```

G.4 EVALUATION RESULTS STRUCTURE

The evaluation results structure records the model’s predictions and performance for each benchmark instance. Listing 3 illustrates the format. For all data structures, the `image_path` field is specified relative to the data directory to ensure clarity and reproducibility. This standardized design facilitates systematic benchmarking and transparent evaluation across a wide range of spectroscopic machine learning tasks.

Listing 3: Example of an evaluation results entry.

```
{
  "id": "Signal_9131_250723_110552_245529_2",
  "image_path": [
    "data/Signal/Spectrum Type Classification/Signal_9131_2_q.png"
  ],
  "question": "What type of spectrum is shown in the image?",
  "choices": [
    "Infrared Spectrum (IR)",
    "Proton Nuclear Magnetic Resonance (H-NMR)",
    "Mass Spectrometry (MS)",
    "Carbon Nuclear Magnetic Resonance (C-NMR)"
  ],
  "answer": "Mass Spectrometry (MS)",
  "category": "Signal",
  "sub_category": "Spectrum Type Classification",
  "source": "",
  "timestamp": "2025-07-23 11:05:52",
  "model_prediction": "Mass Spectrometry (MS)",
  "model_response": "\\answer{Mass Spectrometry (MS)}",
  "pass": true
}
```

H COST ANALYSIS

To ensure consistency and fairness across all experiments, SpectrumLab employs a unified model interface and conducts all inference via API services, regardless of whether the underlying models

are open-source or proprietary. This standardized evaluation pipeline enables direct and equitable comparison of model performance. With the exception of the generation-level scoring model, each benchmark run requires an average of 572 model invocations. The use of remote APIs introduces network latency, resulting in variability in inference times. Depending on the model architecture and complexity, the total time required to complete the full SpectrumBench benchmark ranges from approximately 40 minutes to 2 hours. For each model, we systematically record the overall inference time and the estimated monetary cost associated with completing the benchmark.

Given the current benchmark prompts and SpectrumLab’s prompt engineering design, a complete run of the benchmark requires approximately 1,219,083 input tokens and 41,522 output tokens (as measured on InternVL3-78B, this figure is provided for reference only). Models with more elaborate reasoning or “thinking” capabilities may incur even higher token consumption.

Table 7 summarizes the key statistics for representative models evaluated in this study.

Table 7: Resource consumption and cost for representative models on the full SpectrumBench benchmark.

Model	Inference Time (min)	Cost (USD)
Claude-3.5-Haiku	99	\$0.94
Claude-3.5-Sonnet	70	\$7.47
Claude-4-Opus	123	\$24.00
Claude-4-Sonnet	90	\$11.66
GPT-4o	103	\$4.23
GPT-4-Vision-Preview	113	\$8.08
GPT-4.1-2025-04-14	103	\$1.54
Grok-2-Vision	62	\$2.12
InternVL3-78B	120	N/A

I USAGE OF LARGE LANGUAGE MODELS IN THIS MANUSCRIPT

In preparing this manuscript, we used a large language model (LLM) solely for editorial purposes. Its functions were limited to proofreading for typographical errors, correcting grammatical mistakes, and enhancing the clarity and readability of the text.

J LIMITATIONS

While this work introduces the concept of SpectrumWorld , it is important to acknowledge that the field of AI for Spectroscopy remains in its nascent stages, we recognize several limitations within our primary contributions, SpectrumBench and SpectrumLab .

Limitations of SpectrumBench First, regarding Task Format, SpectrumBench currently supports only multiple-choice and a limited number of open-ended questions. While this design is suitable for Large Language Models (LLMs), it is insufficient for evaluating a broader range of machine learning models, such as Convolutional Neural Networks (CNNs) and Graph Neural Networks (GNNs), as discussed in our introduction. Second, concerning Spectrum Type, although we have incorporated a wide array of spectrum types compared to previous works (Lu et al., 2025; Xu et al., 2025; Bushuiev et al., 2024b; Zhou et al., 2024a), several crucial spectroscopic modalities remain uncovered. Notable examples include X-ray Diffraction (XRD) (Guo et al., 2024a; Salgado et al., 2023) and fluorescence spectra (Parker & Rees, 1962), which are vital for comprehensive material characterization. Finally, addressing Spectroscopic Task Type, spectroscopy techniques are fundamental across diverse scientific disciplines, including physics, astronomy, chemistry, and biology, primarily for characterizing substances like molecules, proteins, peptides, and SMILES sequences. From the perspective of LLMs, a generic categorization of modalities into “text” and “images” is inadequate for representing the complexity of data. The inherent diversity of spectroscopic modalities complicates the immediate definition of all possible tasks. Consequently, SpectrumBench presently lacks important benchmarks in several areas, such as spectrum-spectrum retrieval (Curry et al., 1969;

Wang et al., 2022; Lu et al., 2025) and peptide sequence analysis (Zhou et al., 2024a). We acknowledge that it will be challenging for SpectrumBench to encompass all relevant tasks in the near future, and we aim to foster collaborative efforts with the community and various laboratories to collectively advance the development of AI in spectroscopy.

Limitations of SpectrumLab Our second main contribution, SpectrumLab, also presents certain limitations. Firstly, regarding its data functionality, while SpectrumLab successfully unifies seed datasets and provides data curation tools-SpectrumAnnotator, it currently lacks tools for the preprocessing and segmentation of raw data across multiple spectroscopic modalities. Secondly, concerning metrics, the current evaluation framework within SpectrumLab is relatively simplistic, relying primarily on accuracy and a lenient, LLM-based scoring method for open-ended questions. In future iterations, we plan to define and incorporate a broader array of task-specific metrics to enable more nuanced and robust model evaluation.