Evaluating Scientific Reasoning in Multi-modal Large Language Models

Anonymous ACL submission

Abstract

Recent advancements in multi-modal large language models (MLLMs) have demonstrated promising capabilities in integrating visual and textual information to solve complex problems. While many of these models exhibit strong performance in mathematics or general vision tasks, it remains unclear whether they possess the scientific reasoning skills necessary to tackle challenges across diverse domains such as physics and chemistry. In this work, we aim to bridge this gap by introducing a new benchmark, VisScience, designed to systematically evaluate MLLMs on multi-disciplinary scientific reasoning. Our benchmark consists of 3,000 carefully curated questions spanning K12 education, with equal representation from mathematics, physics, and chemistry (1,000 problems each). These questions are drawn from 21 subject areas and are categorized into five difficulty levels to reflect a broad range of curricular concepts and reasoning demands. With our VisScience, we analyze MLLMs on scientific reasoning by evaluating 25 representative models, including both open-source and closed-source variants. Our results show that MLLMs's performance varies notably across disciplines—while models generally perform best on mathematics, physics and chemistry questions expose weaknesses in scientific abstraction and visual grounding. Furthermore, we examine model behaviors under multilingual settings, as VisScience is provided in both English and Chinese, enabling a cross-linguistic perspective on scientific reasoning.

1 Introduction

Recently, large language models (LLMs) (OpenAI, 2022; Achiam et al., 2023; GLM et al., 2024; Touvron et al., 2023a,b; Bai et al., 2023a; Brown et al., 2020; Chowdhery et al., 2023; Anil et al., 2023) have demonstrated remarkable capabilities across a wide range of tasks, including natural language understanding, text generation, and complex problem solving. The success of LLMs facilitates the development of multi-modal large language models (MLLMs) (OpenAI, 2023; Team et al., 2023; Anthropic, 2024; Liu et al., 2024b,a; Ye et al., 2023, 2024), which extends these capabilities by integrating the ability to process and analyze both textual and visual information. Evaluation is a significant component in assessing the ability of these MLLMs across various tasks, which has attracted widespread attention and developed rapidly in recent years. For instance, several benchmark datasets are proposed to evaluate the ability of MLLMs in general visual understanding, including MME (Fu et al., 2023), MMMU (Yue et al., 2024), MMBench (Liu et al., 2023), MMStar (Chen et al., 2024a), and SEED-Bench (Li et al., 2023a).

As a primary evaluation domain, mathematical reasoning presents specific challenges, requiring models to handle complex mathematical problems accompanied by visual information. Previous works (Chen et al., 2021, 2022; Cao and Xiao, 2022) focus on geometric problems, resulting in the emergence of various evaluation datasets such as GeoQA (Chen et al., 2021), Geometry3K (Lu et al., 2021), and UniGeo (Chen et al., 2022). Subsequently, several benchmark datasets (Lu et al., 2023; Zhang et al., 2024b; Wang et al., 2024) extend the scope of mathematical reasoning beyond geometry to encompass various branches such as arithmetic, algebraic, statistics, logic, and functions. Notably, MathVista also contains a portion of scientific datasets such as TQA (Kembhavi et al., 2017), SciBench (Wang et al., 2023b), and ScienceQA (Lu et al., 2022). However, despite these advancements, there remains some issues:

- Existing benchmarks often focus narrowly on specific mathematics, neglecting other crucial scientific disciplines like physics and chemistry.
- Existing benchmarks are often collected from

limited sources, resulting in a lack of natural difficulty levels and leading to an incomplete evaluation of models' capabilities.

- Current benchmarks are predominantly available in a single language, limiting the evaluation of MLLMs' multilingual capabilities.

To address the limitations of existing benchmarks and provide a more comprehensive evaluation benchmark, we introduce a more expansive evaluation benchmark, named **VisScience**, integrating both textual and visual information. This benchmark is designed to assess the performance of MLLMs in multi-modal scientific reasoning tasks across disciplines like physics and chemistry alongside mathematics. To construct this benchmark, we gather 3,000 questions from K12 education with each discipline containing 1,000 questions. This benchmark spans a comprehensive range of knowledge points across different chapters, with difficulty levels ranging from 1 to 5, ensuring that models are assessed on both basic and challenging problems.

In order to better understand MLLMs' performance on more detailed subjects within three disciplines, we categorize VisScience into several subjects across each discipline. Specifically, we divide the mathematical part of VisScience into six subjects such as *plane geometry, solid geometry, functions and equations, algebraic operations, probability and statistics, and combinatorial mathematics*. For physics, the dataset is categorized as eight subjects, including *mechanics, thermodynamics, comprehensive experiments and methods, mechanical motion, vibration and waves, optics, electromagnetism, and modern physics*. The chemistry section includes seven topics such as *chemical experiments, organic chemistry, material composition, electrochemistry, chemical reactions, inorganic chemistry, and chemical equilibrium*. In summary, VisScience contains 21 subjects across the three disciplines of mathematics, physics, and chemistry.

We conduct extensive experiments on VisScience to evaluate the scientific reasoning abilities of 25 representative MLLMs. These models include close-source LLMs, close-source and open-source MLLMs, offering a comprehensive analysis of their performance across various disciplines (See Figure 1). Our experimental results reveal several key insights: (1) Closed-source models, such as Claude 3.5-Sonnet and GPT-4o, consistently outperform open-source counterparts, especially on high-difficulty questions that require multi-step reasoning or deep conceptual understanding; (2) Model performance varies notably across disciplines—while models generally perform best on mathematics, physics and chemistry questions expose weaknesses in scientific abstraction and visual grounding; (3) Reasoning errors are the dominant source of failure, often stemming from misinterpretation of diagrams, incorrect assumptions, or incomplete logical chains. We hope that VisScience can serve as a stepping stone for advancing research in multi-modal scientific understanding.

2 VisScience Benchmark

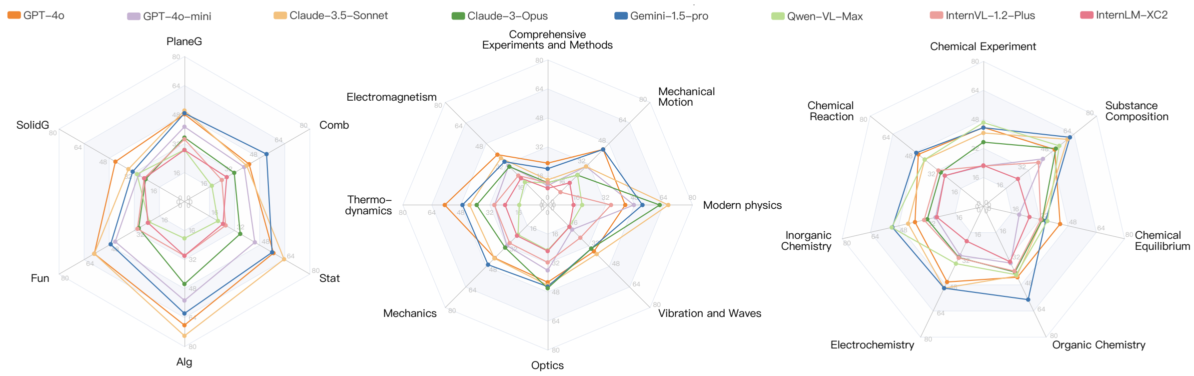
2.1 Overview

We introduce the VisScience benchmark, a meticulously curated collection aimed at evaluating the capabilities of multi-modal large language models (MLLMs) in multi-modal scientific reasoning, with a particular focus on bilingual tasks involving both English and Chinese. This dataset incorporates textual and visual contexts as inputs and spans three scientific disciplines, including mathematics, physics, and chemistry. Each discipline comprises 1,000 questions, meticulously gathered from different chapters to ensure comprehensive coverage of topics and concepts. The core statistics of the VisScience benchmark are presented in Table 1. The distributions of question length in VisScience are provided in Appendix A.1. A detailed introduction of each subjects within the three disciplines is available in Appendix A.2.

2.2 Data Generation

The goal of the VisScience benchmark is to establish a comprehensive, bilingual (Chinese and English) benchmark for evaluating the capabilities of MLLMs in processing and understanding complex, scientifically-oriented tasks across various disciplines. In order to achieve this goal, we present a two-stage data generation pipeline to meticulously construct a benchmark dataset comprising 3,000 questions, evenly distributed with 1,000 questions each in the fields of mathematics, physics, and chemistry. More cases in VisScience are provided in Appendix B.

Data Collection. We gather a total of 450,000 questions from the disciplines of mathematics, physics, and chemistry, each enriched with visual information sourced from K12 education. This collection spans a comprehensive range of knowl-



(a) Mathematics (b) Physics (c) Chemistry

Figure 1: The accuracies of representative MLLMs on VisScience across different subjects

Statistic	Number
Total questions	3000
- multiple-choice questions	2,053 (68.4%)
- Free-form questions	947 (31.6%)
Number of categories of math questions	6
Number of categories of physics questions	8
Number of categories of chemistry questions	7
Number of difficulty levels	5
Unique number of images	3,000
Unique number of questions	3,000
Unique number of answers	1,427
Statistics with Chinese Language	
Maximum question length	1297
Maximum answer length	112
Maximum choice number	5
Average question length	162.85
Average answer length	20.93
Statistics with English Language	
Maximum question length	418
Maximum answer length	92
Maximum choice number	5
Average question length	80.93
Average answer length	12.30

Table 1: Key statistics of VISSCIENCE.

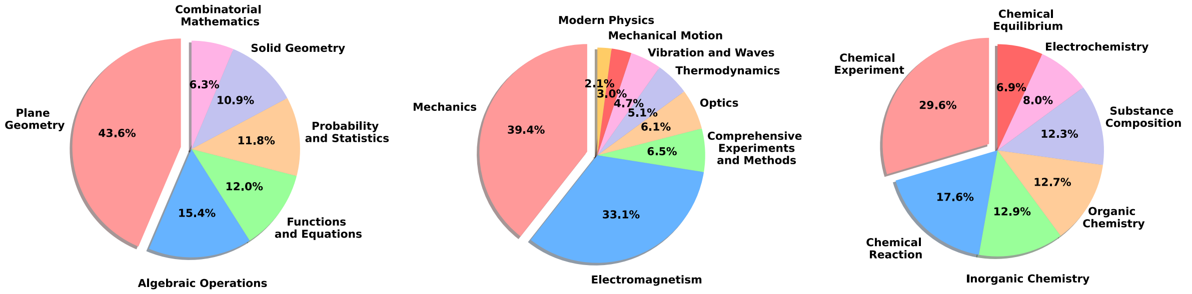
edge points across different chapters, with the difficulty levels scaled based on education grade. Consequently, we cluster 150,000 questions per discipline and carefully select 1,000 representative questions. These questions exemplify a range of difficulty levels and a variety of subjects, guided by the following principles: *(1) Guaranteeing every knowledge point is included in VisScience benchmark. (2) Prioritizing the selection of questions from high-frequency knowledge points. (3) Ensuring a mixture of questions across various difficulty levels.*

In the end, the VisScience benchmark is constructed with 3,000 questions, with each of the three disciplines – mathematics, physics, and chemistry – contributing 1,000 questions. This approach ensures that the benchmark comprehensively covers a wide array of topics within each discipline, reflecting the breadth and depth required for a thorough assessment of MLLMs' capabilities.

Data Annotation. To improve the quality of the VisScience benchmark, we conduct multiple checks using both manual reviews and LLM assessments to confirm the completeness of each question. For textual content, we check for accuracy, coherence and relevance, ensuring that each question aligns with the corresponding scientific discipline and is free of ambiguities. For associated visual content, we rigorously screen out images that are incorrect, unclear, or lacking in detail, retaining only those that are clear and richly informative. To maintain the volume of the VisScience benchmark, we compensate for questions removed due to incomplete information by selecting new questions on identical topics from the original dataset. This approach ensures that the overall number of questions and the breadth of content coverage are consistently maintained. This verification process guarantees that both the textual and visual components of the VisScience benchmark is a reliable and effective tool for evaluating the capabilities of MLLMs in scientific reasoning.

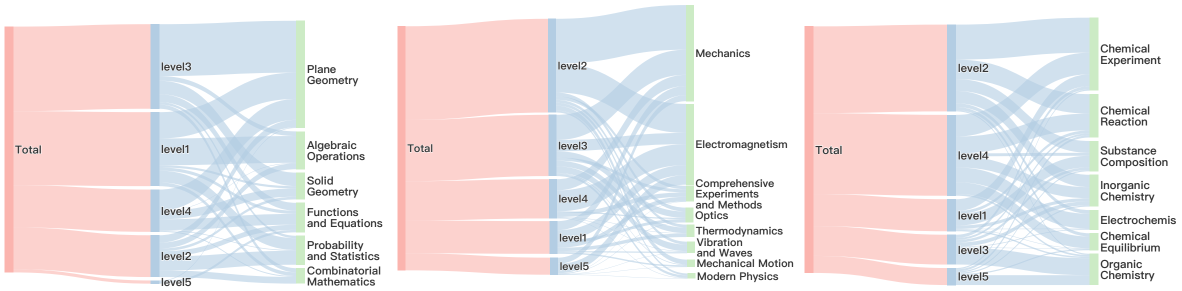
2.3 Data Analysis

We utilize statistical analysis to assess subject distributions and difficulty levels within the VisScience benchmark. Figure 2 presents a visual representation of the categorization, illustrating the distribution of questions across different subjects. Figure 3 depicts the distribution of questions based on difficulty levels, ranging from 1 to 5. Additionally, a comparative analysis between VisScience and other benchmarks is presented in Appendix C.



(a) Mathematics (b) Physics (c) Chemistry

Figure 2: The distribution of detailed subjects in the each discipline within the VisScience benchmark.



(a) Mathematics (b) Physics (c) Chemistry

Figure 3: The distribution of difficulty levels in the each discipline within the VisScience benchmark.

Subject Distributions. To categorize each discipline into more detailed subjects, we first utilize LLM to segment the overall discipline into specific topics based on knowledge points and terminologies presented in the questions. Subsequently, we conduct a manual review of these categories to confirm its rationality and appropriateness, ensuring that each question is accurately categorized. As shown in Figure 2, the mathematical part of the VisScience benchmark is divided into six subjects, i.e., plane geometry (43.6%), algebraic operations (15.4%), functions and equations (12%), probability and statistics (11.8%), solid geometry (10.9%), and combinatorial mathematics (6.3%). Furthermore, the distributions for physics and chemistry disciplines are presented in the figure, providing a comprehensive overview of the scope of the VisScience benchmark within these scientific fields.

Difficulty Levels. To classify the questions into distinct difficulty levels, we first utilize LLM for the initial sorting, and then conduct a manual verification. The questions within each discipline are stratified into five difficulty levels ranging from 1 to 5, defined as follows: Basic, Easy, Intermediate, Advanced, and Expert. Figure 3 shows the distribution of difficulty levels, providing a visual representation of the distribution of questions across different difficulty levels. Each discipline demonstrates a unique profile of topic distribution across the difficulty levels. For instance, in the field of mathematics, *plane geometry* is classified at the intermediate level, *algebraic operations* are positioned at the basic level, and *functions and equations* appears at the highest difficulty level, reflecting their various placement within educational curricula. In physics, *mechanics* dominates the introductory level, which belongs to a fundamental concept in physics education. *Electromagnet* is positioned at the highest difficulty level, demanding the application of various advanced knowledge points. In the discipline of chemistry, *organic chemistry* and *chemical equilibrium* represent the pinnacle of K12 chemical education, requiring deep conceptual understanding and the ability to apply knowledge to complex scenarios.

3 Experiments

3.1 Experimental Setup

Models. We conduct our evaluation across a diverse array of models, including close-source text-only LLMs, close-source MLLMs, and open-source MLLMs. This comprehensive assessment covers 25 models and the sources of models is reported in Appendix D.1.

Evaluation Details. The evaluation process is conducted through two steps: generation and judgment. During the generation phase, the models are tasked with producing responses based on a set of questions. For zero-shot setting, we directly prompt the models with these questions without any examples. For 2-shot Chain-of-Thought (CoT) setting, we provide the models with two relevant examples before they are prompted with the questions. For MLLMs, we supply the models with the textual questions and the corresponding image to obtain their responses. During the judgment phase, we utilize GPT-4o to evaluate the models' responses by comparing them with the standard answers to assess consistency. This phase involves calculating the accuracy across different subjects and levels. The prompts used in two phases is defined in Appendix D.2.

3.2 Overall Experimental Results

Table 2 demonstrates the performance of several models on VisScience within the version of the Chinese language. Experimental results show that the close-source models achieves best performance on VisScience. Specifically, Claude3.5-Sonnet achieves an accuracy of 53.4% in mathematics, GPT-4o attains a 38.2% accuracy in physics, and Gemini-1.5-Pro accomplishes an accuracy of 47.0% in chemistry. Among open-source models, InternVL-1.2-Plus stands out, demonstrating robust capabilities across various scientific disciplines with accuracies of 30.1% in mathematics, 24.8% in physics, and 31.2% in chemistry. Despite this, there is a notable disparity in performance between close-source and open-source models, with close-source models generally exhibiting superior performance. The performance of InternVL-1.2-Plus, although trailing behind the advanced close-source models such as GPT-4o, Claude3.5-Sonnet, and Gemini-1.5-Pro, showing significant potential for improvement. Notably, the performance in physics underscores unique challenges that necessitate targeted improvements in model training. This discipline often involves the interpretation of conceptual and numerical data, challenging the reasoning and computational abilities of MLLMs. As evidenced in Table 2, even advanced models like GPT-4o achieve relatively lower accuracies in physics compared to other disciplines. Results on VisScience with the version of the English language are provided in Appendix E.1.

3.3 Results on Different Subjects.

The mathematical part of VisScience encompasses a wide range of subjects, including plane geometry, solid geometry, functions and equations, algebraic operations, probability and statistics, and combinatorial mathematics. Table 3 reports the comprehensive results across different mathematical subjects. It is evident that models like Claude3.5-Sonnet and GPT-4o in close-source MLLMs excel across multiple subjects, particularly in *functions and equations*, *probability and statistics*, and *algebraic operations*. Conversely, open-source models show a more varied performance with notable strengths in certain areas but generally lower scores compared to close-source models. For instance, InternVL-1.2-Plus and InternVL-Chat-V1.5 perform relatively well in *plane geometry*, and *functions and equations*. These detailed performance on different subjects provide valuable insights into the specific strengths and weaknesses of various MLLMs. Additionally, results on physics and chemistry across different subjects are presented in Appendix E.2 and Appendix E.3, respectively.

3.4 Analysis of Results

To analyze the causes of errors in model responses, we meticulously review incorrect answers to identify common patterns. We specifically focus on GPT-4o to illustrate specific instances of errors and their distributions across the disciplines of mathematics, physics, and chemistry. Figure 4 demonstrates the distributions of these errors, categorizing them into several types such as reasoning error, knowledge error, calculation error, vision recognition error, and question misunderstood error. Notably, across all disciplines, reasoning errors are the most prevalent, indicating a challenge in model's ability to solve scientific problems that involve visual information. Specifically, reasoning errors account for 56.5% of the total errors in mathematics, 50.1% in physics, and 40.6% in chemistry, respectively. This is followed by knowledge error, which is particularly significant in chemistry, constituting 33.2% of the errors in that discipline. Similarly, knowledge error also represent the second most common error type in physics. However, knowledge error in mathematics is less prevalent, making up only 8.8% of the total errors. This indicates that while the model struggle with conceptual and fundamental principles in chemistry and physics, it demonstrate a better grasp of mathematical con-

Model	LLM	Input	Mathematics	Physics	Chemistry
		Close Source Models (APIs)			
Text-only LLMs					
Zero-shot ChatGPT	-	Q	22.4	22.7	18.6
Zero-shot GPT-4	-	Q	25.9	30.4	33.1
Zero-shot Claude-2	-	Q	27.3	22.0	24.4
Zero-shot Claude3-Opus	-	Q	29.3	30.8	32.5
Zero-shot Claude3.5-Sonnet	-	Q	29.7	35.3	36.9
Zero-shot GPT-4o	-	Q	31.1	38.0	39.6
2-shot CoT Claude2	-	Q	27.8	21.7	23.9
2-shot CoT ChatGPT	-	Q	20.2	18.6	21.3
2-shot CoT GPT-4	-	Q	32.1	31.5	32.4
Multi-modal LLMs					
Gemini-1.0-Pro	-	Q, I	26.6	23.70	27.8
Gemini-1.5-Pro	-	Q, I	49.4	38.1	47.0
GPT-4o	-	Q, I	51.7	38.2	41.6
GPT-4o-mini	-	Q, I	42.6	29.8	28.4
Qwen-VL-Max	-	Q, I	35.5	30.70	42.5
Qwen-VL-Plus	-	Q, I	27.6	26.5	37.7
Claude3.5-Sonnet	-	Q, I	53.4	38.0	43.1
Claude-3 opus	-	Q, I	34.4	31.1	34.1
GLM-4V	-	Q, I	24.2	19.2	25.0
Step-1V	-	Q, I	28.1	23.5	25.0
		Open Source Models			
General Multi-modal LLMs					
mPLUG-Owl	LLaMA-7B	Q, I	7.6	8.3	9.5
LLaMA-Adapter-V2	LLaMA-7B	Q, I	9.6	10.3	10.8
MiniCPM-Llama3-V2.5	LLaMA3-8B	Q, I	15.4	17.9	19.5
LLaVA-1.5	Vicuna-13B	Q, I	15.5	15.2	18.8
LLaVA-1.5	Vicuna-7B	Q, I	13.0	13.5	16.0
DeepSeek-VL	DeepSeek-LLM-7B	Q, I	8.3	16.8	21.0
ShareGPT4V	Vicuna-7B	Q, I	15.7	14.0	19.0
ShareGPT4V	Vicuna-13B	Q, I	16.4	14.9	18.4
SPHINX-Plus	LLaMA2-13B	Q, I	17.0	15.3	20.4
InternLM-XC2	InternLM2-7B	Q, I	24.9	18.3	25.6
InternVL-1.2-Plus	Nous-Hermes-2-Yi-34B	Q, I	30.1	24.8	31.2
InternVL-Chat-V1.5	Mixtral 8*7B	Q, I	26.9	20.8	23.7
CogVLM	Vicuna-7B	Q, I	16.7	14.5	17.0
CogVLM2	LLaMA-3-8B	Q, I	23.2	14.4	21.0
GLM-4V-9B	GLM-4-9B	Q, I	24.7	19.3	22.5
InternVL2-8B	InternLM2-7B	Q, I	28.5	24.4	27.8
Qwen2-VL-7B	Qwen2-7B	Q, I	27.1	6.3	4.5
InternVL2.5-8B	InternLM2-7B	Q, I	29.3	23.8	29.9
Qwen2.5-VL-7B	Qwen2.5-7B	Q, I	38.2	7.1	5.0

Table 2: **Results on VisScience within the version of the Chinese language across the disciplines of mathematics, physics, and chemistry.** For input, Q represents for question, I represents for image. The highest scores among close-source and open-source models are highlighted in red and blue, respectively.

cepts. Vision recognition error is another significant type of error, accounting for 18.8% of the errors in mathematics, making it the second most prevalent error type in this discipline. This error category is also significant in physics and chemistry, where it constitutes 17.8% and 15.3% of the errors, respectively. This type of error highlights the challenges faced by the model in processing and understanding visual information. Furthermore, calculation error accounts for a small portion of the errors, especially in chemistry, indicating that the model excels particularly in handling numerical computations. More detailed examples of these errors can be found in Appendix F.

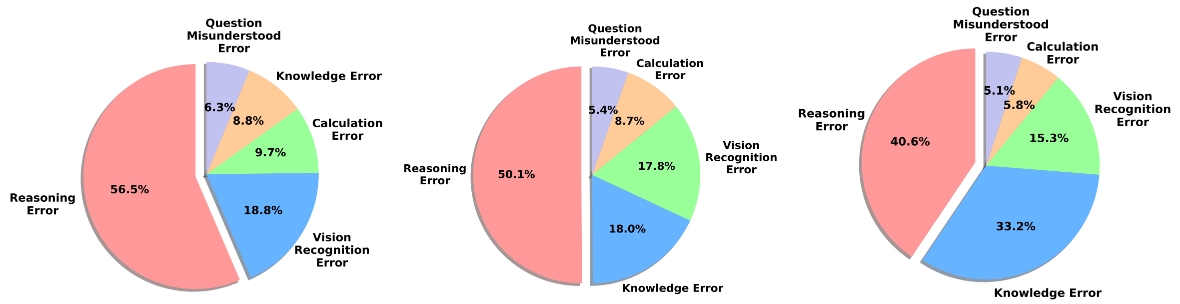
4 Related Works

Multi-modal Reasoning Benchmarks. Recently, the evaluation of multi-modal large language models (MLLMs) (OpenAI, 2023; Team

Model	Mathematics						
	ALL	PlaneG	SolidG	Fun	Alg	Stat	Comb
Close Source Models (APIs)							
Text-only LLMs							
Zero-shot ChatGPT	22.40	20.18	11.93	18.33	13.63	15.25	26.98
Zero-shot GPT-4	25.90	30.73	18.35	28.33	17.53	24.58	33.33
Zero-shot Claude-2	27.30	27.06	25.69	25.83	31.17	31.36	25.40
Zero-shot Claude3-Opus	29.30	30.28	21.10	32.50	27.27	34.75	31.75
Zero-shot Claude3.5-Sonnet	29.70	33.94	15.60	33.33	27.27	27.12	34.92
Zero-shot GPT-4o	31.10	36.24	24.77	35.83	25.32	24.58	31.75
2-shot CoT Claude2	27.80	30.05	26.61	25.00	28.57	27.97	26.98
2-shot CoT ChatGPT	20.20	23.17	20.18	19.17	17.53	22.88	14.29
2-shot CoT GPT-4	32.10	37.16	31.19	28.33	22.08	30.51	38.10
Multi-modal LLMs							
Gemini-1.0-Pro	26.60	24.08	22.02	23.73	35.71	29.66	34.92
Gemini-1.5-Pro	49.40	48.74	33.03	47.06	61.69	55.93	52.38
GPT-4o	51.70	48.17	44.04	57.50	68.18	56.78	41.27
GPT-4o-mini	42.60	41.28	29.36	44.17	54.55	44.92	38.10
Qwen-VL-Max	35.50	34.86	27.52	35.83	50.00	33.05	26.98
Qwen-VL-Plus	27.60	27.98	18.35	29.17	31.17	35.59	20.63
Claude3.5-Sonnet	53.40	50.23	35.78	57.50	74.03	63.56	39.68
Claude3-Opus	34.40	35.31	24.77	29.17	45.45	35.59	31.75
GLM-4V	24.20	28.57	30.28	22.50	20.26	21.37	17.46
Step-1V	28.10	31.68	24.71	23.15	48.85	40.57	22.64
Open Source Models							
General Multi-modal LLMs							
mPLUG-Owl	7.60	6.19	10.09	5.00	12.34	7.63	7.94
LLaMA-Adapter-V2	9.60	10.78	10.09	7.50	9.09	13.56	4.76
MiniCPM-Llama3-V2.5	15.40	23.62	19.27	15.83	26.62	26.27	15.87
LLaVA-1.5-13B	15.50	15.83	15.60	12.50	18.83	14.41	14.29
LLaVA-1.5-7B	13.00	12.84	12.84	15.83	14.29	11.86	11.11
DeepSeek-VL	8.30	13.99	8.26	10.00	11.04	10.17	7.94
ShareGPT4V-7B	15.70	16.06	16.51	13.33	14.29	17.80	17.46
ShareGPT4V-13B	16.40	15.60	11.93	19.17	17.53	22.03	14.29
SPHINX-Plus	17.00	21.79	19.27	15.83	20.13	22.88	7.94
InternLM-XC2	24.90	25.92	22.02	22.50	27.92	27.97	20.63
InternVL-1.2-Plus	30.10	34.40	25.69	30.00	29.87	26.27	23.81
InternVL-Chat-V1.5	26.90	28.44	25.69	23.33	29.87	24.58	26.98
CogVLM	16.70	16.06	23.85	17.50	17.53	13.56	19.05
CogVLM2	23.20	21.56	22.02	29.17	22.73	26.27	20.63
GLM-4V-9B	14.70	25.23	20.18	19.17	27.27	33.05	19.05
InternVL2-8B	29.20	32.85	25.52	29.76	27.55	29.73	24.32
Qwen2-VL-7B	27.33	30.53	25.27	27.54	25.31	28.66	23.09
InternVL2.5-8B	29.89	33.98	25.39	29.95	28.61	31.55	25.72
Qwen2.5-VL-7B	37.25	41.74	22.94	27.50	44.16	39.83	38.10

Table 3: **Results on the mathematical part of VisScience across different subjects.** Subjects: PlaneG: plane geometry, SolidG: solid geometry, Fun: functions and equations, Alg: algebraic operations, Stat: probability and statistics, Comb: combinatorial mathematics. The highest scores among close-source and open-source models are highlighted in red and blue, respectively.

et al., 2023; Anthropic, 2024; Bai et al., 2023b; Wang et al., 2023a; Liu et al., 2024a,b) in various reasoning tasks has become increasingly crucial. So many benchmark datasets for these tasks span several categories are proposed like MME (Fu et al., 2023), MMMU (Yue et al., 2024), MMBench (Liu et al., 2023), MMStar (Chen et al., 2024a), SEED-Bench (Li et al., 2023a), and CMMMU (Zhang et al., 2024a), which evaluate models' capabilities to apply logic and inference; mathematical reasoning; scientific reasoning, and agent-based reasoning. MMMU covers university-level questions from six domains, which is utilized to assess MLLMs' advanced perception and reasoning abilities. CMMMU (Zhang et al., 2024a) evaluates models' reasoning abilities across various disciplines through bilingual multi-modal questions in Chinese and English. Existing bench-



(a) Mathematics (b) Physics (c) Chemistry

Figure 4: Error distributions of GPT-4o on VisScience across the disciplines of mathematics, physics, and chemistry.

mark like ScienceQA (Lu et al., 2022) is a specialized dataset designed to evaluate the capabilities of MLLMs, particularly in the domain of scientific reasoning. Furthermore, several benchmarks such as MathVista (Lu et al., 2023), MathVerse (Zhang et al., 2024b), and MATH-Vision (MATH-V) (Wang et al., 2024) are specially designed to evaluate the mathematical reasoning capabilities of MLLMs. While these benchmarks are valuable, they present limitations such as an overemphasis on mathematics and a broad array of topics that often lack depth in science-related questions and exhibit uneven difficulty levels. Our dataset addresses these shortcomings by providing 3,000 scientific reasoning questions across mathematics, physics, and chemistry, which is collected from K12 education. Additionally, it includes bilingual questions in Chinese and English, enriching the knowledge base and offering a more extensive range of difficulty levels to create a more comprehensive evaluation platform.

Multi-modal Large Language Models. Recently, the success of large language models (LLMs) (Du et al., 2021; Zeng et al., 2022; Achiam et al., 2023; Gao et al., 2023; GLM et al., 2024; Bai et al., 2023a) has spurred the ongoing development of multi-modal large language models (MLLMs). These MLLMs (Liu et al., 2024b; Liu et al.; Wang et al., 2023a; Li et al., 2023b; Dai et al., 2024; Bai et al., 2023a) expand upon traditional LLM capabilities by integrating the ability to process and analyze both text and images. Currently, close-source MLLMs like Gemini (Team et al., 2023), GPT-4v (OpenAI, 2023), Qwen-VL (Bai et al., 2023b), and Claude3 (Anthropic, 2024) demonstrate impressive capabilities in general image understanding and scientific reasoning. Besides, the development of open-source multi-modal large language models (MLLMs) continues to expand, providing an important complement to their closed-source models. These open-source MLLMs, such as mPLUG-Owl (Ye et al., 2023, 2024), LLaMA-Adapter-V2 (Gao et al., 2023), MiniCPM (Hu et al., 2024), LLaVA-1.5 (Liu et al., 2024a), LLaVA-NeXT (Liu et al.), DeepSeek-VL (Lu et al., 2024), ShareGPT4V (Chen et al., 2023a), SPHINX (Gao et al., 2024), InternVL (Chen et al., 2023b), InternVL 1.5 (Chen et al., 2024b), InternLM-XComposer2 (Dong et al., 2024), and CogVLM (Wang et al., 2023a), also achieves advance performance, further enriching the landscape of MLLM domain. Here, we utilize our specially curated benchmark VisScience to evaluate these MLLMs across tasks in mathematics, physics, and chemistry. This comprehensive evaluation aims to assess their capabilities in scientific reasoning.

5 Conclusion

In this paper, we introduce a comprehensive benchmark, VisScience, designed to evaluate the capabilities of multi-modal large language models (MLLMs) in scientific reasoning across mathematics, physics, and chemistry. VisScience consists of 3,000 questions, evenly distributed across these three disciplines, spanning 21 subjects and categorized into five difficulty levels. We conduct evaluations using VisScience on 25 prominent models, including both closed-source and open-source variants. The experimental results reveal that model performance is not uniformly correlated with question complexity. Surprisingly, models often fail simpler problems that require basic scientific intuition or conceptual understanding, especially in lower-grade physics and chemistry topics. By offering a challenging set of questions across varied scientific fields, VisScience provides a robust benchmark for assessing the scientific reasoning abilities of MLLMs.

Limitations

Despite the effectiveness of our proposed benchmark and evaluation framework, several limitations remain. First, the evaluation of open-ended responses depends on a single automatic judge (e.g., GPT-4o), which may introduce stylistic or reasoning biases, especially when assessing models with different output formats or alignment strategies. This automated evaluation, while scalable, lacks the nuanced judgment of human assessors and may not fully capture the correctness or reasoning depth of each response. Second, while our benchmark includes multiple disciplines and difficulty levels, it is still grounded in K12 curricula from specific educational contexts. As such, it may not fully represent the diversity of global science education standards or real-world applications of scientific reasoning. Further efforts are needed to extend coverage to additional domains such as biology, engineering, and environmental science.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, and 1 others. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jie Cao and Jing Xiao. 2022. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1511–1520.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. 2021. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024a. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023a. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, and 1 others. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021.

Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, and 1 others. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, and 1 others. 2024. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, and 1 others. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5376–5384.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge (january 2024). *URL https://llava-vl. github. io/blog/2024-01-30-llava-next*, 1(8).

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2023. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, and 1 others. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

OpenAI. 2022. Chatgpt.

OpenAI. 2023. Gpt-4v(ision) system card. In *technical report*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, and 1 others. 2023a. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023b. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, and 1 others. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051.

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

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, and 1 others. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, and 1 others. 2024a. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.11944*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and 1 others. 2024b. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*.

A Dataset Details

A.1 Question Length Distribution

We provide both Chinese and English versions of the VisScience benchmark. The Chinese version features an average of 162.85 words per question, with the longest question comprising 1,297 words. Answers in this version average 20.93 words, with the longest reaching 112 words. Conversely, the English version shows an average of 80.93 words per question, with the longest question spanning 418 words. Answers here average 12.3 words, with the most detailed answer containing 92 words. Figure 5 depicts the distribution of word counts, highlighting the diversity and complexity of questions.

A.2 Detailed Description of Subjects

VisScience consists of three disciplines: mathematics, physics, and chemistry. The mathematics section includes six subjects: *algebraic operations, combinatorial mathematics, functions and equations, probability and statistics, plane geometry, and solid geometry*. The physics section is composed of eight subjects: *mechanics, optics, modern physics, mechanical motion, electromagnetism, vibrations and waves, comprehensive experiments and methods, and thermodynamics*. The chemistry section includes seven subjects: *chemical experiments, chemical reactions, inorganic chemistry, organic chemistry, electrochemistry, substance composition, and chemical equilibrium*. A more detailed introduction of the above subjects is presented as follows:

A.2.1 Mathematics

Algebraic Operations. Algebraic operations include the manipulation of algebraic expressions, such as addition, subtraction, multiplication, and division. They are fundamental for solving algebraic equations and inequalities and are widely applied across various fields of mathematics.

Combinatorial Mathematics. Combinatorial mathematics studies the counting, arrangement, and combination of discrete structures, involving graph theory, number theory, and coding theory. It has significant applications in computer science, optimization, and probability theory.

Functions and Equations. Functions and equations are core parts of mathematics, dealing with relationships between variables and their representations. Functions are mappings between inputs and outputs, while equations are equalities concerning these mappings. Mastering knowledge of functions and equations is fundamental for solving many practical problems and is widely applied in engineering, physics, and economics.

Probability and Statistics. Probability and statistics study the laws of random events and methods of data analysis, including probability distributions, statistical inference, and data analysis techniques. They have broad applications in scientific research, engineering, and economics.

Plane Geometry. Plane geometry studies the shapes and figures in two-dimensional space, including points, lines, angles, and polygons. It is a fundamental part of mathematics education.

Solid Geometry. Solid geometry involves the study of geometric shapes in three-dimensional space, including points, lines, surfaces, and polyhedra. It examines the properties, volumes, and surface areas of these geometric bodies and is foundational for architecture, physics, and engineering.

A.2.2 Physics

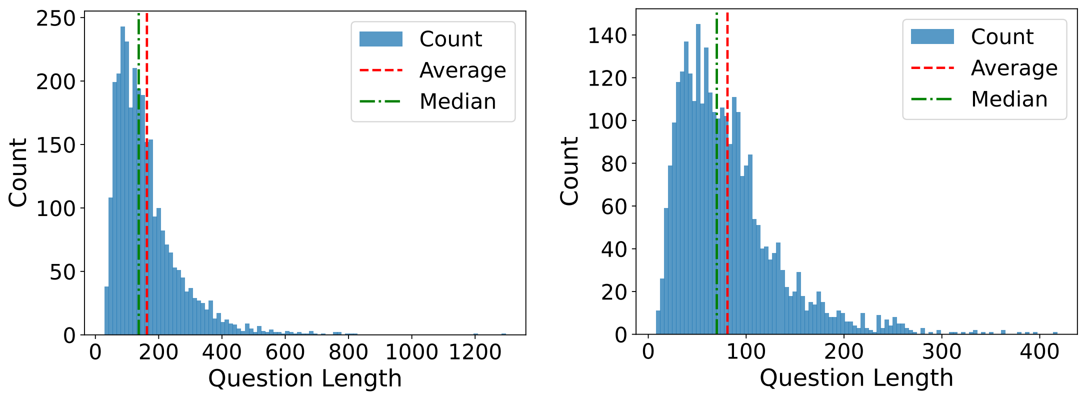
Mechanics. Mechanics studies the motion of objects and the forces acting upon them, including classical mechanics, quantum mechanics, and relativistic mechanics. It is the foundation of physics and is widely applied in engineering, astronomy, and materials science.

Optics. Optics studies the properties of light and its interactions with matter, including reflection, refraction, interference, and diffraction. Optical technologies have broad applications in imaging, communication, and laser technology.

Modern Physics. Modern physics includes theories developed since the 20th century, such as quantum mechanics, relativity, and particle physics. These theories have expanded our understanding of the fundamental laws of nature.

Mechanical Motion. Mechanical motion studies the movement of objects under the influence of forces, including linear motion, rotational motion, and vibration. Understanding mechanical motion is fundamental for the design and analysis of mechanical systems.

Electromagnetism. Electromagnetism studies the interactions between electric and magnetic fields, including electrostatics, magnetic fields, and electromagnetic waves. It is the basis of modern



(a) Chinese Version

(b) English Version

Figure 5: The distributions of word counts per question in the Chinese and English versions of VisScience.

physics and electrical engineering.

Vibration and Waves. Vibration and waves study vibrating systems and wave phenomena, including sound waves, light waves, and electromagnetic waves. They have broad applications in communication, acoustics, and optical technologies.

Comprehensive Experiments and Methods. Comprehensive experiments and methods involve using various experimental techniques and methods in physics teaching and research. They include designing and conducting experiments to observe and analyze the effects of specific variables on outcomes. Through comprehensive experiments, students can grasp the complexities of scientific research, cultivate scientific reasoning abilities, and understand the meticulousness and uncertainties of experimental work.

Thermodynamics. Thermodynamics studies the processes of energy transformation and transfer, including the laws of thermodynamics, thermodynamic systems, phase transitions, and heat engines. Thermodynamics is a fundamental aspect of both physics and engineering, with broad applications in energy, environmental science, and materials science. By investigating the relationship between internal and external energy of objects, thermodynamics reveals the basic principles of energy conversion and transfer in nature, providing theoretical support for the development of modern industrial technology.

A.2.3 Chemistry

Chemical Experiment. Chemical experiments involve studying the properties and changes of substances through experimental methods. Students learn to design experiments, observe chemical reactions, collect and analyze data, and draw conclusions in chemical experiments. Chemical experiments play a crucial role in understanding chemical theories and applying chemical knowledge.

Chemical Reaction. Chemical reactions study the chemical changes between substances, including reaction types, mechanisms, and rates. Understanding chemical reactions is essential for predicting and controlling chemical processes, which have wide applications in pharmaceutical manufacturing, materials science, and environmental engineering.

Inorganic Chemistry. Inorganic chemistry studies the properties and reactions of non-carbon elements and their compounds. It covers a wide range of topics from metals and non-metals to transition metals and coordination compounds and is key to understanding the periodic table of elements and chemical reaction mechanisms.

Organic Chemistry. Organic chemistry studies the structure, properties, and reactions of carbon-containing compounds. It has significant applications in pharmaceutical chemistry, materials science, and biochemistry.

Electrochemistry. Electrochemistry studies the interconversion between electrical and chemical energy, including processes such as batteries, electrolysis, and electroplating. Electrochemistry has important applications in energy storage, corrosion control, and electrochemical sensors.

Substance Composition. Substance composition studies the chemical composition and structure of substances, including the arrangement of molecules, atoms, and ions. It has important applications in chemistry, materials science, and biology.

Chemical Equilibrium. Chemical equilibrium studies the behavior of chemical reactions when they reach a dynamic equilibrium state, including equilibrium constants, Le Chatelier's principle, and solubility equilibrium. Understanding chemical equilibrium is essential for predicting reaction directions and optimizing chemical processes.

B Dataset Case

The VisScience dataset consists of 3,000 carefully selected high-quality questions, evenly distributed across three disciplines: mathematics, physics, and chemistry, with each comprising 1,000 questions. Each discipline within VisScience encompasses several subjects: mathematics includes six subjects, physics contains eight subjects, and chemistry comprises seven subjects. To illustrate the diversity and depth of VisScience, we provide more examples sampled from each discipline. In mathematics, six subjects include algebraic operations, combinatorial mathematics, functions and equations, probability and statistics, plane geometry, and solid geometry are illustrated in Figure 6 to Figure 11. Figure 12 to Figure 18 demonstrate eight subjects within the physics section of VisScience, comprising mechanics, optics, modern physics, mechanical motion, electromagnetism, vibrations and waves, comprehensive experiments and methods, and thermodynamics. The chemistry section includes seven subjects: chemical experiments, chemical reactions, inorganic chemistry, organic chemistry, electrochemistry, substance composition, and chemical equilibrium, which are illustrated in Figure 19 to Figure 25.

C Comparison with Other Benchmarks

We compare the VisScience benchmark with 5 existing benchmarks, including MathVista (Lu et al., 2023), Math-Vision (Wang et al., 2024), CMMMU (Zhang et al., 2024a), ScienceQA (Lu et al., 2022), and SciBench (Wang et al., 2023b).

VisScience vs MathVista. MathVista is a comprehensive multi-modal benchmark for mathematical reasoning, comprising data from 28 existing datasets and 3 newly collected datasets. In MathVista, the majority of questions are annotated after collecting images, which results in a certain homogeneity within the data. In contrast, VisScience directly collects its questions from K12 education, featuring an average question length of 80.93 words. Such questions provide more contextual information, which facilitate a more thorough evaluation of the models' reasoning capabilities. Unlike MathVista that encompasses only seven subjects within mathematics, VisScience offers a far broader scope, including 22 distinct subjects across mathematics, physics, and chemistry. Furthermore, VisScience distinguishes itself by being a bilingual benchmark, including both Chinese and English versions of questions. This feature is particularly advantageous as it assesses MLLMs' capabilities in scientific reasoning across different languages.

VisScience vs Math-Vision. Math-Vision is a mathematics benchmark derived from 19 competitions, covering 16 topics across 5 levels of difficulty. Different from Math-Vision that collected from competitions, VisScience spans a broader educational spectrum, incorporating a natural gradient of difficulty from elementary school to high school. Furthermore, VisScience extends beyond mathematics to include questions from physics and chemistry, significantly broadening its scope and applicability. While Math-Vision primarily focuses on the unique challenges of competitive environments, VisScience is grounded in real-world educational settings.

VisScience vs CMMMU. CMMMU comprises 12,000 manually collected multi-modal questions from university exams, quizzes, and textbooks, which covers 6 core subjects and 30 specific fields. Similar to VisScience, CMMMU is a bilingual benchmark, offering questions in both Chinese and English. Within this dataset, only 1,601 questions are dedicated to the disciplines of mathematics, physics, and chemistry, accounting for only 13.34% of the total dataset. VisScience features a total of 3,000 questions, significantly outnumbering those in CMMMU dedicated to the same subjects. The questions in CMMMU are set at the university level, characterized by high difficulty, demanding that the model possesses substantial professional domain knowledge and expert-level reasoning abilities. In contrast, VisScience comes from K12 education, with a broader range of difficulty. This range allows VisScience to more comprehensively evaluate MLLMs' capabilities across different educational stages.

VisScience vs ScienceQA. ScienceQA is a newly developed benchmark featuring approximately 21,000 multimodal multiple-choice questions across a variety of science topics. In the

ScienceQA dataset, 30.8% of questions incorporate both image and text contexts, providing a multimodal benchmark to test MLLMs in scientific reasoning. The questions in ScienceQA have an average length of only 12.11 words. In contrast, VisScience also serves as a benchmark for evaluating the scientific reasoning abilities of MLLMs, but it typically features longer and more textually detailed questions. Specifically, the Chinese version of VisScience has an average question length of 162.85 words, providing a more comprehensive and intricate testing ground for evaluating the depth of detailed reasoning in MLLMs. Additionally, VisScience contains mathematical problems, further enriching the benchmark's scope by testing MLLMs on their mathematical problem solving capabilities alongside their scientific reasoning.

VisScience vs SciBench. SciBench is a benchmark developed to evaluate the reasoning capabilities of LLMs in solving collegiate-level scientific problems within the domains of mathematics, chemistry, and physics. The majority of the data in SciBench focuses on assessing the scientific reasoning of LLMs, it only includes 177 problems that incorporate visual elements to evaluate the performance of MLLMs. In contrast, VisScience is primarily focused on multimodal scientific reasoning, covering similar subjects such as mathematics, chemistry, and physics. VisScience differentiates itself by offering a more comprehensive range of difficulty levels and subjects, making it a broader benchmark for assessing the capabilities of MLLMs in scientific reasoning.

D Evaluation Details

D.1 The Sources of Models

In Table 4, we present the sources of the models tested on VisScience.

D.2 Prompts

We introduce the prompts used to guide models in generating responses in Chain-of-Thought (CoT) settings and judging the LLMs' answers. The specific prompts can be found in Table 5.

E More Experimental Results

E.1 Results on VisScience in English Version

Table 6 reports a comprehensive comparison of various models on the VisScience benchmark in the English version. The benchmark evaluates performance across three disciplines: mathematics, physics, and chemistry. Among close-source models, GPT-4o demonstrates the highest performance across two disciplines, achieving an accuracy of 53.6% in mathematics and 42.7% in physics. However, Claude3.5-Sonnet surpasses GPT-4o in chemistry with a higher accuracy of 43.6%. Open-source models generally show lower performance compared to close-source counterparts. Notably, InternVL-1.2-Plus displays competitive performance, reaching up to 26.0% in mathematics, 23.6% in physics, and 27.8% in chemistry. The English version of VisScience is designed to facilitate the evaluation of MLLMs that specialize in English, assessing their capabilities in scientific reasoning.

E.2 Results on Physics Across Different Subjects

Table 7 presents a detailed analysis of various models on VisScience across different subjects within the physics section, which includes mechanics, electromagnetism, thermodynamics, comprehensive experiments and methods, optics, vibration and waves, modern physics, and mechanical motion. The table highlights that while GPT-4o exhibits the top performance on the entire physics discipline, the best performance in individual subjects varies notably. For instances, Claude3.5-Sonnet excels specifically in modern physics with an accuracy of 66.67%, significantly surpassing other close-source models in this area. This variation in performance by subject underscores the specialized capabilities of different models. Moreover, this detailed analysis provides more insights, emphasizing the need for targeted improvements to achieve balanced performance across all physics subjects.

E.3 Results on Chemistry Across Different Subjects

Table 8 presents a nuanced view of the performance of various models across different subjects within the chemistry discipline of the VisScience benchmark. The chemistry discipline includes chemical experiment, chemical reaction, inorganic chemistry, electrochemistry, organic chemistry, chemical equilibrium, and substance composition. Notably, Gemini-1.5-Pro stands out among close-source models, excelling across the entire chemistry discipline. It demonstrates particular prowess in organic chemistry and substance composition, achiev-

Model	Input	LLM Size	Source
		Closed Source Models	
Text-only LLMs			
ChatGPT	Q	-	gpt-3.5-turbo
GPT-4	Q	-	gpt-4
Claude-2	Q	-	claude-2
Multi-modal LLMs			
Gemini-1.0-Pro	Q, I	-	gemini-pro
Gemini-1.5-Pro	Q, I	-	gemini-1.5-pro
GPT-4o	Q, I	-	gpt-4o
Claude3-Opus	Q, I	-	claude-3-opus-20240229
Claude3.5-Sonnet	Q, I	-	claude-3-5-sonnet-2024620
Qwen-VL-Plus	Q, I	-	qwen-vl-plus
Qwen-VL-Max	Q, I	-	qwen-vl-max
GLM-4V	Q, I	-	glm-4v
Step-1V	Q, I	-	step-1v
		Open Source Models	
General Multi-modal LLMs			
mPLUG-Owl	Q, I	7B	mPLUG-Owl
DeepSeek-VL	Q, I	7B	deepseek-vl-7b-base
LLaMA-Adapter-V2	Q, I	7B	LLaMA-Adapter V2
LLaVA-1.5	Q, I	7B	LLaVA-v1.5-7B
LLaVA-1.5	Q, I	13B	LLaVA-v1.5-13B
ShareGPT-4V	Q, I	7B	ShareGPT4V-7B
ShareGPT-4V	Q, I	13B	ShareGPT4V-13B
GLM-4v-9B	Q, I	7B	GLM-4v-9B
SPHINX-Plus	Q, I	13B	SPHINX-Plus
InternVL-Chat-V1.5	Q, I	20B	InternVL 1.5
InternVL-1.2-Plus	Q, I	34B	InternVL-Chat-V1-2-Plus
InternLM-XC2	Q, I	7B	InternLM-XComposer2-VL-7B
CogVLM	Q, I	17B	CogVLM-17B
CogVLM2	Q, I	19B	cogvlm2-llama3-chat-19B
MiniCPM-Llama3-V-2_5	Q, I	19B	MiniCPM-Llama3-V 2.5
InternVL2-8B	Q, I	8B	InternVL2-8B
Qwen2-VL-7B	Q, I	7B	Qwen2-VL-7B
InternVL2.5-8B	Q, I	8B	InternVL2.5-8B
Qwen2.5-VL-7B	Q, I	7B	Qwen2.5-VL-7B

Table 4: The source of the models used in the evaluation.

Task	Prompt
Response Generation	You are an exceptionally talented mathematics (physics/chemistry) instructor. Kindly furnish an elaborate, step-by-step solution to the question.
Answer Judgment	You are a highly skilled mathematics (physics/chemistry) teacher. I will provide you with a mathematics (physics/chemistry) problem, along with its ground answer and the model response from the model. Please determine whether the ground answer and the model response are consistent. Note that you do not need to judge the correctness of either answer, only whether they are consistent. If it is a multiple-choice question, both answers must choose the exact same option to be considered consistent. If it is a calculation problem, the relative error between the model response and the ground answer must be less than 0.05 to be considered consistent. If the problem has multiple sub-questions, each sub-question's answer must be identical for consistency. If you find them consistent, please add [Consistent] at the end of your response. If you find them inconsistent, please add [Inconsistent] at the end of your response.

Table 5: Prompts for response generation and answer judgment.

ing impressive accuracies of 57.02% and 61.16%, respectively. Additionally, Qwen-VL-Max leads in chemical experiment and inorganic chemistry, achieving the highest accuracies of 46.28% and 51.94%, respectively. Open-source models demonstrate a range of performances, with InternVL-1.2-Plus leading this group. It achieves the highest open-source accuracy in nearly all subjects. This comprehensive review of model performances within the chemistry section of the VisScience benchmark highlights the need to enhance MLLMs' capabilities in scientific domains, ensuring models are both accurate and adaptable across various disciplines.

F Error Case

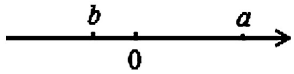
We conduct rigorous tests on a series of open-source and close-source models on VisScience and perform a detailed analysis of the models' responses. These errors in the models' answers can be classified into five categories: reasoning error, vision recognition error, knowledge error, calculation error, and question misunderstood error. We present examples of these five error types across the disciplines of mathematics, physics, and chemistry, with a specific focus on errors made by GPT-4o. Additionally, we demonstrate error examples from other representative close-source models such as GLM-4V, Qwen-VL-max, and Claude 3.5, as well as open-source models like LLAVA-1.5 , GLM-4V-9B and InternVL-Chat-1.5. Notably, it should be noted that the types of errors made by these models in response to the same questions can differ from those made by GPT-4o. This analysis helps to underline the varied challenges faced by different models in processing complex scientific questions, providing insight into their respective strengths and limitations. Figure 26 to Figure 27 demonstrate cases of errors from representative models in the mathematical part of VisScience. Figure 28 to Figure 29 show the incorrect answers in the physics section. Figure 30 to Figure 32 demonstrate the errors in the chemistry section.

This is an appendix.

Model	LLM	Input	Mathematics	Physics	Chemistry
		Close Source Models (APIs)			
Text-only LLMs					
Zero-shot ChatGPT	-	Q	17.4	20.7	25.2
Zero-shot GPT-4	-	Q	29.9	37.7	38.7
Zero-shot Claude-2	-	Q	24.6	22.7	25.6
Zero-shot Claude3-Opus	-	Q	21.7	15.8	29.4
Zero-shot Claude3.5-Sonnet	-	Q	27.2	35.7	35.2
Zero-shot GPT-4o	-	Q	35.2	40.3	42.5
2-shot CoT Claude2	-	Q	25.7	21.9	24.1
2-shot CoT ChatGPT	-	Q	24.4	20.1	22.1
2-shot CoT GPT-4	-	Q	36.5	39.0	38.1
Multi-modal LLMs					
Gemini-1.0-Pro	-	Q, I	26.4	39.1	27.9
Gemini-1.5-Pro	-	Q, I	47.8	35.1	39.1
GPT-4o	-	Q, I	53.6	42.7	43.3
GPT-4o-mini	-	Q, I	43.2	33.7	34.9
Qwen-VL-Max	-	Q, I	30.7	26.4	36.3
Qwen-VL-Plus	-	Q, I	21.9	20.9	29.7
Claude3.5-Sonnet	-	Q, I	50.8	36.6	43.6
Claude3-Opus	-	Q, I	34.4	29.4	34.7
GLM-4V	-	Q, I	23.1	18.5	23.4
Step-1V	7B	Q, I	32.0	19.5	27.6
		Open Source Models			
General Multi-modal LLMs					
mPLUG-Owl	LLaMA-7B	Q, I	7.4	12.3	12.3
LLaMA-Adapter-V2	LLaMA-7B	Q, I	12.6	11.4	16.2
MiniCPM-Llama3-V2.5	LLaMA3-8B	Q, I	24.4	20.6	24.4
LLaVA-1.5	Vicuna-13B	Q, I	15.0	17.4	21.1
LLaVA-1.5	Vicuna-7B	Q, I	17.4	16.6	18.9
DeepSeek-VL	DeepSeek-LLM-7B	Q, I	16.0	16.9	17.8
ShareGPT-4V	Vicuna-7B	Q, I	14.7	17.7	21.3
ShareGPT-4V	Vicuna-13B	Q, I	14.5	16.0	20.2
SPHINX-Plus	LLaMA2-13B	Q, I	17.9	15.7	22.4
InternLM-XC2	InternLM2-7B	Q, I	20.7	20.5	25.0
InternVL-1.2-Plus	Nous-Hermes-2-Yi-34B	Q, I	26.0	23.6	27.8
InternVL-Chat-V1.5	Mixtral 8*7B	Q, I	24.9	23.0	25.9
CogVLM	Vicuna-7B	Q, I	18.5	15.9	23.1
CogVLM2	LLaMA-3-8B	Q, I	24.2	16.6	24.9
GLM-4V-9B	GLM-4-9B	Q, I	24.7	19.2	23.9
InternVL2-8B	InternLM2-8B	Q, I	25.8	22.1	26.5
Qwen2-VL-7B	Qwen2-7B	Q, I	24.3	6.5	4.3
InternVL2.5-8B	InternLM2-7B	Q, I	25.9	23.5	27.7
Qwen2.5-VL-7B	Qwen2.5-7B	Q, I	37.3	6.8	4.9

Table 6: **Results on VisScience within the version of the English language across the disciplines of mathematics, physics, and chemistry.** The highest scores among close-source and open-source models are highlighted in red and blue, respectively.

Question



The positions of numbers a and b on the number line are shown in Figure (1). Then ab 0. (Fill in with ">" or "<")

Answer

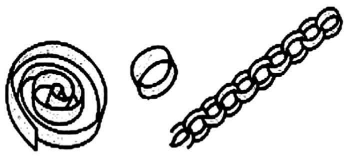
<

Question

The students cut a 75.5 cm paper strip into segments of 7.8 cm each to make circular garlands, as shown in the picture.

(1) One such paper strip can make ________ circular garlands.

(2) If 15 circular garlands are used to make one garland chain, 12 such paper strips can make ________ garland chains.



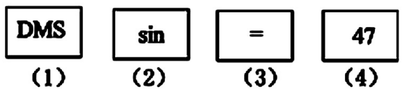
Answer

(1)9
(2)7

Question

If the sequence numbers of the four keys on the calculator are shown in the figure, in the angle measurement unit "degrees," the correct key sequence to calculate sin47° is ( )

A. (1)(2)(3)(4)
B. (2)(4)(1)(3)
C. (1)(4)(2)(3)
D. (2)(1)(4)(3)



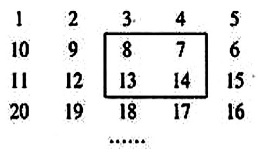
Answer

D

Question

As shown in the figure, arrange all positive integers in a "snake-like" manner as follows. Moving the frame in the figure, the sum of the four numbers in the frame could be

A. 2016
B. 2018
C. 2020
D. 2022

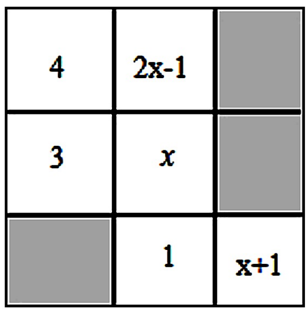


Answer

D

Question

As shown in the figure, in a 3×3 magic square, fill in 9 numbers in the nine blank spaces so that the sum of the three numbers in the same row, the same column, or the same diagonal is equal. The following figure is the magic square that Xiao Lan's classmate is going to fill in, in which three squares have been blacked out. According to the rules of filling the magic square, find the value of x.



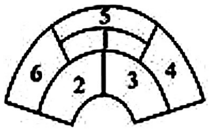
Answer

X=5

Figure 6: Cases of *algebraic operations* in mathematical part of VisScience.

Question

A certain city is building a flowerbed in the central square. The flowerbed is divided into 6 parts (as shown in the figure). Now, 4 different colors of flowers need to be planted, one type in each part, and no adjacent parts can be planted with the same color. How many different planting methods are there (answer with a number).

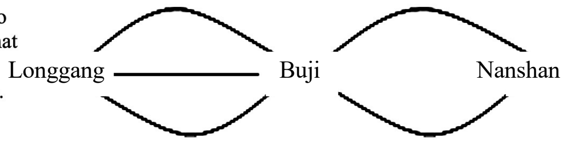


Answer

120

Question

As shown in the picture, A drives from Longgang to Nanshan and must pass through Buji. It is known that there are three roads to choose from Longgang to Buji and two roads to choose from Buji to Nanshan. A has a total of ( ) routes to choose from.
A. 5 B. 6 C. 4 D. 9

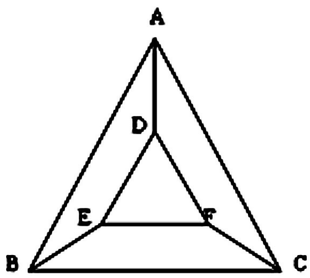


Answer

B

Question

As shown in the figure: Color the six points ABCDEF in the figure with four different colors, requiring each point to be colored with one color, and the two endpoints of each line segment in the figure to be colored differently. The number of different coloring methods is . (Answer with a number)

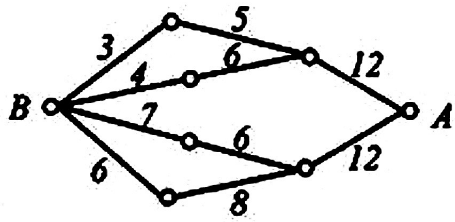


Answer

264

Question

As shown in the figure, small circles represent network nodes, the lines between the nodes indicate they have network connections, and the numbers on the lines show the maximum information flow through that segment of the network per unit time. If information is transmitted from node A to node B and can be split to travel along different routes simultaneously, the maximum amount of information that can be transmitted per unit time is ( )
A. 26
B. 24
C. 20
D. 19



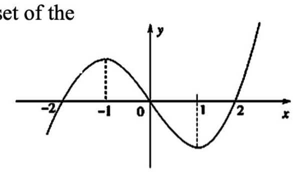
Answer

D

Figure 7: Cases of *combinatorial mathematics* in mathematical part of VisScience.

Question

The graph of the differentiable function f(x) on R is shown in . Then the solution set of the inequality $(x^2 - 2x\ 3)f'(x) > 0$ is _______.
$((-\infty, -1) \cup (-1, 1) \cup - (3, +\infty))$

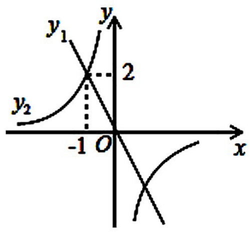


Answer

D

Question

As shown in the figure, the direct proportionality function (y_{1}) and the inverse proportionality function (y_{2}) intersect at point E (-1, 2). If (y_{1} > y_{2} > 0), then the range of values for (x) is

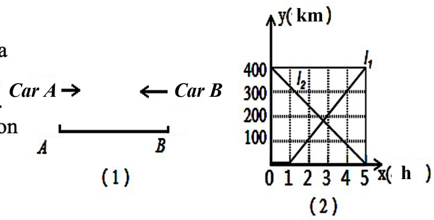


Answer

X<1

Question

As shown in Figure (1), location B is directly east of location A. At a certain moment, car B starts traveling from location B to location A, and 1 hour later, car A starts traveling from location A to location B. When car A reaches location B, car B simultaneously reaches location A.
As shown in Figure (2), the horizontal axis x (hours) represents the travel time of both cars (calculated from the moment car B starts traveling), and the vertical axis y (kilometers) represents the distance of both cars from location A. How many kilometers are there between locations A and B?

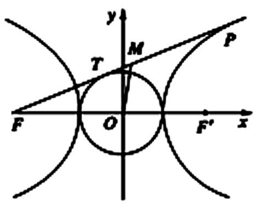


Answer

400 kilometers

Question

As shown in the figure, from the left focus F of the hyperbola (frac{x^2}{a^2} - frac{y^2}{b^2} = 1) (a > 0, b > 0), a tangent to the circle $(x^2 + y^2 = a^2)$ is drawn at point T. Extending FT intersects the right branch of the hyperbola at point P. If M is the midpoint of the line segment FP and O is the origin of the coordinates, then the relationship between (|MO| - |MT|) and (b - a) is ( )
A. (|MO| - |MT| > b - a)
B. (|MO| - |MT| = b - a)
C. (|MO| - |MT| < b - a)
D. Uncertain



Answer

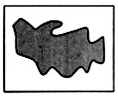
B

Figure 8: Cases of *functions and equations* in mathematical part of VisScience.

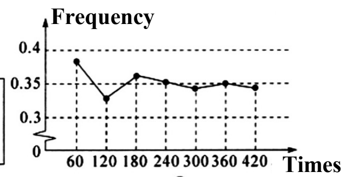
Question

As shown in Figure ①, there is an irregular pattern (the shaded part in the figure) on a flat ground. Xiaoming wants to know the area of this pattern, so he took the following approach: he surrounded the irregular pattern with a rectangle of length 5m and width 4m, and then randomly threw small balls at the rectangular area and recorded the number of times the balls landed on the irregular pattern (balls landing on the boundary line or outside the rectangular area were not counted). He plotted the results of several valid trials into a line graph as shown in Figure ②, from which he estimated the area of the irregular pattern to be approximately (   )

A. 6m²
B. 7m²
C. 8m²
D. 9m²



①



②

Answer

X<1

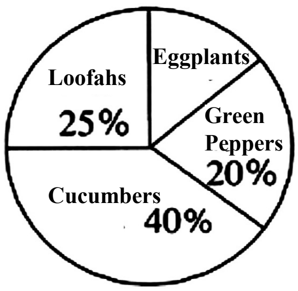
Question

In a vegetable plot, green peppers, cucumbers, loofahs, and eggplants are planted. As shown in the image, it represents the planting area percentage of each vegetable relative to the total area.

(1) The planting area of eggplants accounts for (_____) % of the total vegetable plot area.

(2) If the planting area of loofahs is 30m², the total area of this vegetable plot is (_____) m².

(3) (_____) has the largest planting area, which is greater than the planting area of loofahs by (_____) % of the total vegetable plot area.

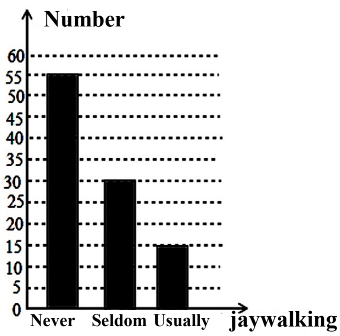


Answer

15 120 cucumber 15

Question

The Municipal Traffic Police Brigade conducted traffic safety knowledge publicity for the students of a certain school. They randomly surveyed the instances of jaywalking among these students anonymously and created a statistical chart as shown in the figure. Please answer the following questions based on the information in the chart:
(1) How many students were surveyed in this survey?
(2) If the total number of students is 1500, estimate how many of them frequently jaywalk.
(3) Share your insights based on the information reflected in the chart. (No more than 30 words).



Answer

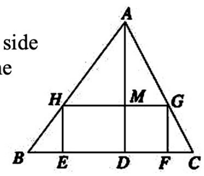
(1) 100
(2) 225
(3) You can simply share your own understanding based on the actual situation; there isn't a single correct answer.

Figure 9: Cases of *probability and statistics* in mathematical part of VisScience.

Question

As shown in the figure, rectangle ABCD, R is the midpoint of CD, point M moves along the side BC, E and F are the midpoints of AM and MR respectively. Then the length of EF with the movement of point M( )

A.Becomes shorterB. Becomes longerC. Remains the sameD.Cannot be determined

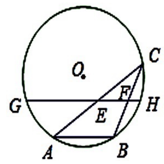


Answer

C

Question

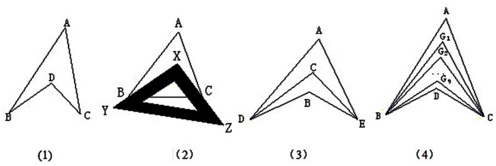
As shown in the figure, AB is a chord of the circle O. Point C is a moving point on the circle O, and ∠ACB=30°. Points E and F are the midpoints of AC and BC, respectively. The line EF intersects the circle O at points G and H. If the radius of the circle O is 7, then the maximum value of GE+FH is



Answer

10.5

Question



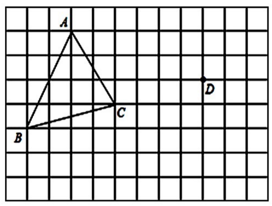
Exploration and Discovery:
The shape shown in Figure 1 resembles a common learning tool—a compass. Let us call this shape a "compass diagram." What kind of mathematical knowledge is hidden in this simple diagram? Now, use your intelligence and solve the following problems:
(1) Observe the "compass diagram" and explore the relationship among ∠BDC and ∠A, ∠B, ∠C, and explain the reasoning;
(2) Directly use the above conclusion to solve the following three problems:
① As shown in Figure 2, place a set square XYZ on △ABC such that the two right-angle edges XY, XZ just pass through points B and C. If ∠A=50°, then ∠ABX+∠ACX= °;
② As shown in Figure 3, DC bisects ∠ADB, and EC bisects ∠AEB. If ∠DAE=50° and ∠DBE=130°, find the measure of ∠DCE;
③ As shown in Figure 4, the 10 division lines of ∠ABD and ∠ACD intersect at points G_{1}, G_{2}…G_{9}. If ∠BDC=140° and ∠BG_{1}C=77°, find the measure of ∠A.

Answer

(1) ∠BDC=∠A+∠B+∠C; (2) ①40°; ②90°; ③70°

Question

In a square grid where the side length of each small square is 1 unit, the positions of the three vertices of △ABC are as shown in the figure. △ABC is now translated so that point A is moved to point D, and points E and F correspond to points B and C, respectively.
(1) Please draw the translated △DEF and find the area of △DEF;
(2) If AD and CF are connected, the relationship between these two segments is .



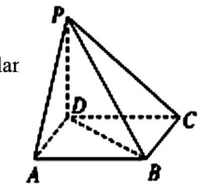
Answer

(1) The area of △DEF = 7;
(2) Parallel and equal.

Figure 10: Cases of *plane geometry* in mathematical part of VisScience.

Question

As shown in the figure, point P is outside the plane of square ABCD. PD is perpendicular to plar ABCD, and PD equals AD. The angle between PA and BD is

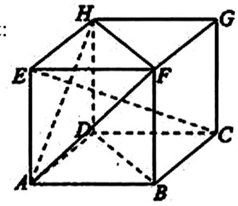


Answer

60°

Question

As shown in the figure, in the cube ABCD-EFGH, the following statements are incorrect:
A. BD // Face FHA
B. EC ⊥ BD
C. EC ⊥ Face FHA
D. The angle between the skew lines BC and AH is 60°

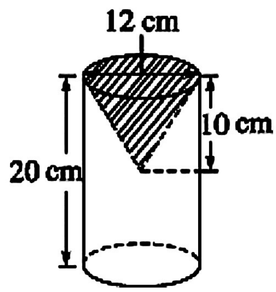


Answer

D

Question

As shown in the figure, the remaining part after a cone is removed from a cylinder is to be calculated. Please calculate its volume.

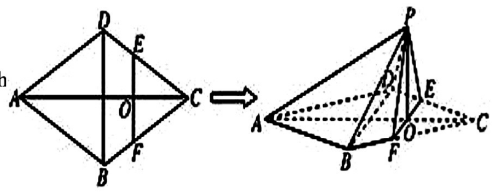


Answer

1884 cm3

Question

As shown in the figure, in the rhombus ABCD with side length 4 and ∠DAB = 60°, points E and F are on sides CD and CB respectively, and point E does not coincide with points C and D. EF is perpendicular to AC, and EF intersects AC at O. By folding △CEF along EF to the position of △PEF, the plane PEF is perpendicular to the plane ABFED.Let the volume of the triangular pyramid P-ABD be V_{1}, and the volume of the quadrilateral pyramid P-BDEF be V_{2}. Find the value of V_{1}:V_{2} when PB is minimized.



Answer

4:3

Figure 11: Cases of *solid geometry* in mathematical part of VisScience.

Question

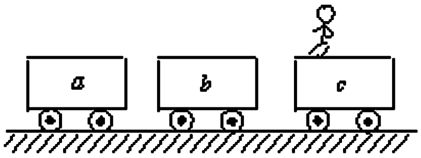
As shown in the figure, three identical flat-topped cars a, b, and c are aligned in a straight line and rest on a smooth horizontal surface. A child on car c jumps to car b, and then immediately jumps from car b to car a with the same horizontal speed relative to the ground. After jumping onto car a, the child remains stationary relative to car a. Thereafter ( )

A:a and c have equal speeds

B:a and b have equal speeds

C:The speed relationship of the three cars is v_c > v_a > v_b

D:a and c move in opposite directions



Answer

CD

Question

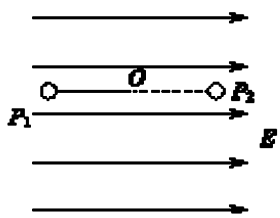
As shown in the figure, in a vertical plane, a positively charged small ball is tied to one end of an inextensible light string of length L, with the other end of the string fixed at point O. They are in a uniform electric field, with the direction of the field being horizontal to the right and the field strength being E. It is known that the force exerted by the electric field on the small ball equals the gravitational force on the small ball. Now, the small ball is first pulled to the position P_{1} in the figure, making the light string straight and parallel to the field direction, and then the ball is released from rest. It is known that when the small ball passes the lowest point, its vertical velocity component becomes zero instantaneously due to the tension in the string, while the horizontal component remains unchanged (ignoring air resistance). The tension T in the string when the small ball reaches point P_{2} at the same height as point P_{1} is ( )

A:2mg

B:3mg

C:4mg

D:5mg



Answer

B

Question

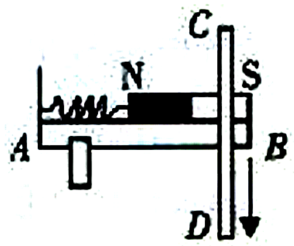
As shown in the figure, a bar magnet is placed parallel to the smooth horizontal edge AB of the table, with a vertical block at end A. A horizontal spring connects the block and the magnet. Currently, a straight conductor CD, with a downward current, is placed vertically in front of the magnet and offset towards the South pole with a small gap between them. Within a short time after placing CD, the following judgments are correct:

A: The length of the spring will increase

B: The length of the spring will decrease

C: The magnet will move outward on the table, closer to the straight conductor CD

D: The magnet will move inward on the table, away from the straight conductor CD



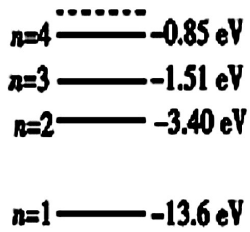
Answer

BD

Figure 12: Cases of *mechanics* in physics part of VisScience.

Question

According to the hydrogen atom energy level diagram (as shown):



A: The energy of the photons emitted when the hydrogen atom transitions is continuous

B: The smaller the orbital radius of the electron, the greater the energy of the hydrogen atom

C: The hydrogen atom in its ground state is the most stable

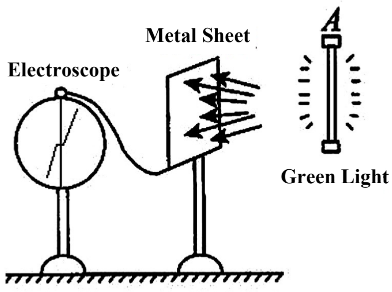
D: To excite the hydrogen atom in the ground state, 11eV photons can be used

Answer

C

Question

As shown in the figure, an electroscope is connected to a metal plate. When the metal plate is illuminated with a green light at point A, the electroscope's needle deflects at a certain angle. After turning off the light, the needle maintains its position at a certain angle.



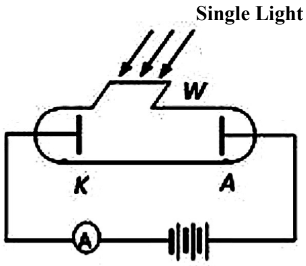
(1) When a small metal sphere carrying a slight negative charge touches the metal plate, the electroscope's needle deflection angle will ______ (fill in "increase", "decrease", or "remain unchanged").
(2) If the electroscope needle is reset to zero and the metal plate is illuminated with yellow light from a sodium lamp of the same intensity, the electroscope needle does not deflect. If a more intense infrared lamp is used to illuminate the metal plate, the observation will show ______ (fill in "deflection" or "no deflection") of the electroscope needle. If a very faint purple light is used instead, the observation will show ______ (fill in "deflection" or "no deflection") of the electroscope needle.

Answer

(1) Reduce (2) No Yes

Question

As shown in the figure, the circuit is used to study the photoelectric effect. The cathode K and anode A are two electrodes sealed in a vacuum glass tube. When K is illuminated, it can emit photoelectrons. The anode A absorbs the photoelectrons emitted by the cathode K, forming a photocurrent in the circuit. When the cathode K is illuminated with monochromatic light a, the pointer of the ammeter deflects; when the photoelectric tube cathode K is illuminated with monochromatic light b, the pointer of the ammeter does not deflect. Which of the following statements is correct? (Fill in the correct answer letter)



A. The wavelength of light a must be less than the wavelength of light b
B. Increasing the intensity of light a alone may increase the current through the ammeter
C. Increasing the intensity of light a alone can increase the maximum initial kinetic energy of the ejected electrons
D. The work function of the cathode material is related to the frequency of the incident light

E. When the cathode K is illuminated with monochromatic light a, if the polarity of the power supply is reversed, the reading of the ammeter may drop to zero

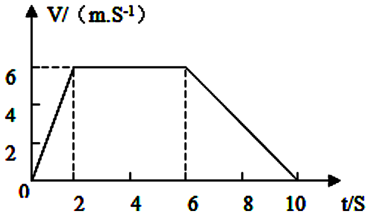
Answer

ABE

Figure 13: Cases of *modern physics* in physics part of VisScience.

Question

As shown in the figure is the v-t graph of an elevator moving upward, the maximum speed of the object's motion equals m/s, and the height of uniform ascent equals ( ).

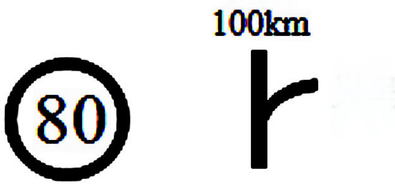


Answer

From the graph, it can be seen that the elevator first accelerates uniformly upwards, then moves upwards at a constant speed, and finally decelerates uniformly upwards, with a maximum speed of 6m/s.
The height of the constant speed upward motion h = 6×4m = 24m.
Therefore, the answer is: 6, 24m.

Question

To ensure orderly and safe highway traffic, many traffic signs are placed along the roadside.
As shown in the picture, image A is a speed limit sign indicating the maximum allowed speed is 80 km/h; image B is a route indicator sign, showing the distance of 100 km to Hangzhou. The physical meanings of the two data mentioned are ( )
A: 80 km/h is average speed
B: 80 km/h is instantaneous speed
C: 100 km is displacement
D: 100 km is distance

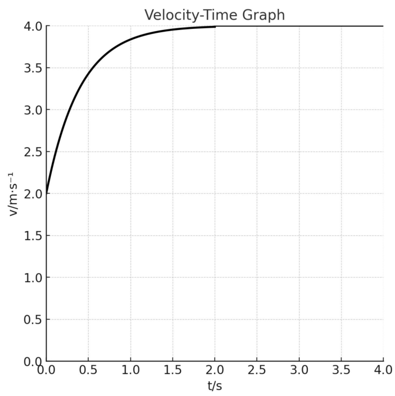


Answer

Solution: The maximum allowable speed represents the speed at a certain position, which is the instantaneous speed, so 80 km/h refers to the instantaneous speed;
There are still 100 km to Hangzhou, and 100 km is the length of the trajectory, which is the distance. Therefore, BD is correct, and AC is incorrect.
The answer is: BD

Question

The value of the area enclosed by the velocity-time graph of a particle's straight-line motion and the coordinate axes is equal to the displacement of the particle during that period. This holds true not only for uniform linear motion and uniformly accelerated linear motion but also for any variable motion. The figure shows the velocity-time graph of a certain particle during the 0-4s interval of straight-line motion. According to the graph, the displacement of the particle during this period is closest to ( )
A:15m
B:13m
C:11m
D:7m



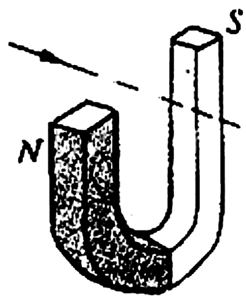
Answer

Solution: The area enclosed by the graph line and the time axis is approximately 15 squares, so the displacement is: x=15×1×1m=15m.
Thus, the answer is: A.

Figure 14: Cases of *mechanical motion* in physics part of VisScience.

Question

As shown in the figure, when an electron beam passes through the two magnetic poles in the direction indicated by the arrow in the figure, its deflection direction is ( )
A: Upward
B: Downward
C: Toward the N pole
D: Toward the S pole

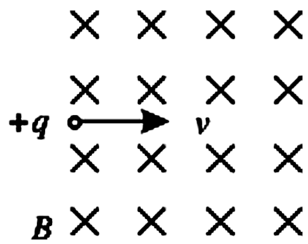


Answer

B

Question

As shown in the figure, a positively charged particle enters a uniform magnetic field horizontally to the right, neglecting gravity. Regarding the motion of this charged particle after it enters the magnetic field, the following judgments are correct: ( )
A: The particle deflects upwards
B: The particle deflects downwards
C: The particle does not deflect
D: The particle quickly stops moving

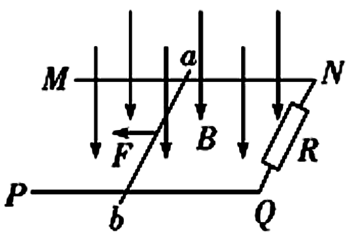


Answer

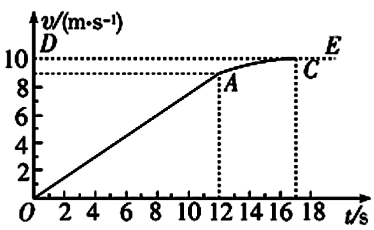
A

Question

As shown in Figure A, a uniform magnetic field B=0.5T exists in space, directed vertically downward. MN and PQ are rough, parallel, long straight rails in the same horizontal plane, with a distance L=0.2m between them; R is a resistor connected at one end of the rails. The ab is a conducting rod with a mass of m=0.1kg bridged across the rails. From time zero, a small motor exerts a horizontal leftward traction force F on rod ab, causing it to start moving from rest and accelerating along the rails. During this process, the rod always remains perpendicular to and in good contact with the rails. Figure B shows the v-t graph of the rod, where segment OA is a straight line, and AC is a curve. The small motor reaches its rated power P=4.5W at the end of 12s, after which it maintains constant power. The resistance of other parts except R is negligible, and g=10m/s^{2}.
(1) Find the acceleration of the conductor rod ab during the 0-12s interval;
(2) Find the coefficient of kinetic friction between rod ab and the rails and the value of the resistor R;
(3) Draw the graph of the traction force's power versus time (P-t) on the answer sheet.



(A)



(B)

Answer

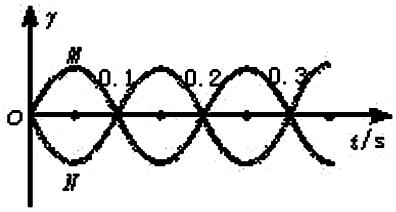
0.75m/s^{2} Dynamic friction coefficient is 0.2 Resistance value is 0.4 ohms\nThe graph is a parabola passing through the origin from 0 to 12 seconds (but the vertex of the parabola is not the origin). After 12 seconds, the power is constant.

Figure 15: Cases of *electromagnetism* in physics part of VisScience.

Question

M and N are two points in the direction of wave propagation in the medium, with a distance s =\"1.5\" m, and their vibration images are shown in the figure. The possible values for the wave speed are( ):

A:15 m/s
B:7.5 m/s
C:5 m/s
D:3 m/s



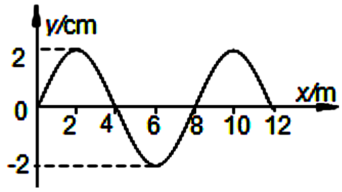
Answer

ACD

Question

As shown in the figure, the image represents a mechanical wave propagating along the x-axis at a certain moment. From the figure, it can be seen that the amplitude A and wavelength λ of this wave are respectively ( )

A:A=2cm,λ=4m
B:A=2cm,λ=8m
C:A=4cm,λ=4m
D:A=4cm,λ=8m



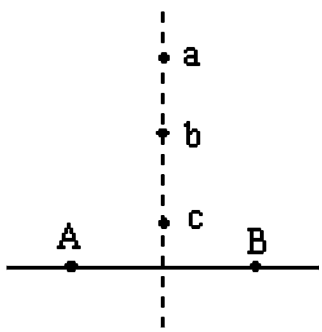
Answer

According to the definition of amplitude: Amplitude A is the maximum distance the vibrating object moves away from the equilibrium position, read the maximum value of y as A=2cm. Based on the fact that wavelength equals the distance between two adjacent wave crests or troughs, read the wavelength λ=4m.
So choose B.

Question

There are two vibrating sources A and B on the water surface, vibrating in exactly the same way. On the perpendicular bisector of the line connecting A and B, there are three points a, b, and c. At a certain moment, point a is the convergence point of the crests of the two waves, and point c is the nearest convergence point of the troughs of the two waves from point a. Point b is located on the line connecting points a and c, as shown in the figure. Which of the following statements is correct? ( )

A: Both a and c are points of constructive interference, and b is a point of destructive interference.
B: a is a point of constructive interference, and c is a point of destructive interference.
C: Both a and c are currently points of constructive interference, but after some time, they will become points of destructive interference.
D: a, b, and c are all points of constructive interference.



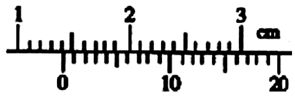
Answer

Solution: From the question, A and B are two sources of vibration with identical conditions. Points a, b, and c lie on the perpendicular bisector of the line AB, meaning the distances from a, b, and c to the two wave sources are equal, resulting in a path difference of zero. The vibrations at these three points all reinforce each other.\nTherefore, the answer is: D

Figure 16: Cases of *vibration and waves* in physics part of VisScience.

Question

A vernier caliper with 20 divisions was used to measure the diameter of a small ball during an experiment. The reading is shown in the figure, and the diameter of the small ball is d = ( ) mm.



Answer

Solution: The main scale reading is 14mm, and the vernier scale reading is 0.05×3=0.15mm, so the final reading is 14.15mm.

Question

In the experiment to measure the rated power of a 3.8V bulb, it is known that the rated power of the bulb is approximately 1.5W. Besides the bulb, ammeter, voltmeter, switch, and wires, the teacher also provides:
A. 6V power supply;
B. 4V power supply;
C. A variable resistor of "5Ω 2A";
D. A variable resistor of "100Ω 0.1A".
(1) To accurately measure the rated power of the bulb, the variable resistor you choose is , and the power supply you choose is (fill in the letter).
(2) When the bulb is glowing normally, the ammeter pointer is shown as in Figure. Please calculate the actual rated power W and the resistance of the bulb Ω.

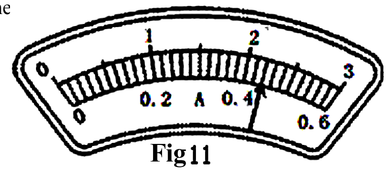


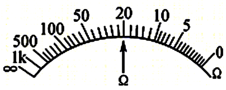
Fig11

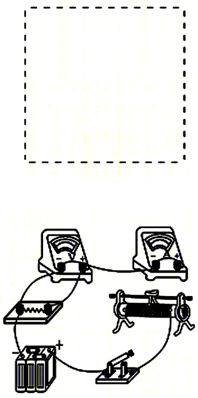
Answer

(1) C_ , B__;
(2) 1.67 , 8.6 .

Question

Measure the resistance value of an unknown resistor.
(1) A student first roughly measures the resistance using a multimeter. The multimeter's switch is set to the ×10Ω range. After zero adjustment, the red and black probes are connected to the resistor's terminals, and the pointer reading is as shown in the figure. The measured resistance value is ________Ω.
(2) The student then plans to accurately measure the resistance using the VA method. The provided experimental equipment includes: 8V DC power supply; voltmeter (0-10V, internal resistance about 20kΩ); ammeter (0-50mA, internal resistance about 10Ω); sliding rheostat (0-20Ω, 1A); switch, and wires. Based on the experimental requirements and the provided equipment, refer to the partially incomplete physical circuit below and draw the experimental circuit diagram in the dashed box below, completing the unconnected wires.
(3) In the experiment mentioned in (2), after connecting the circuit and closing the switch, the student found that both the ammeter and voltmeter had no readings. The student used a multimeter to check for circuit faults. The operations were as follows: The multimeter was set to the DC voltage range, and the red and black probes were connected to: the positive and negative terminals of the power supply; the two ends of the rheostat's resistance wire; between the "-" terminal of the ammeter and the "+" terminal of the voltmeter. The multimeter's pointer deflected in all cases, indicating that the wire connected between ____________ might have broken.
(4) In the experiment, the student moves the rheostat's sliding head, records multiple sets of ammeter and voltmeter readings (U, I), and plots the U-I curve on graph paper. In this experiment, the measured value is ________ the true value. (fill in ">", "=" or "<")





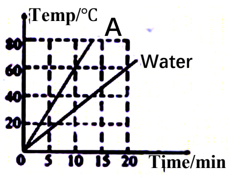
Answer

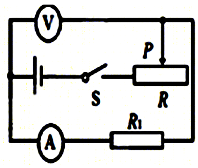
(1) 200
(2) As shown in the figure
(3) Connect the ammeter and voltmeter
(4) <

Figure 17: Cases of *comprehensive experiments and methods* in physics part of VisScience.

Question

Using two identical electric heaters to heat substances A and water, both with a mass of 2kg, the relationship between their temperature and time is shown in the diagram. Based on this, determine the amount of heat absorbed by substance A in 10 minutes ( ).
A:5.04×10^5 J
B:4.2×10^5 J
C:2.52×10^5 J
D:Insufficient conditions, cannot be calculated



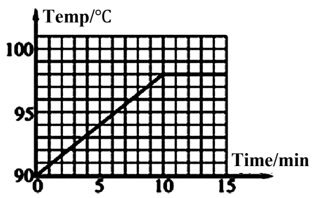


Answer

C

Question

In the experiment of exploring the "boiling law of water," Xiaoming recorded the temperature every 30s after the water temperature rose to 90°C, and then plotted the temperature versus time graph, as shown in the figure. If the thermometer used by Xiaoming is accurate, it can be concluded from the graph that the boiling temperature of water is °C, which indicates that the atmospheric pressure at that time is 1 standard atmosphere (choose "greater than," "less than," or "equal to").

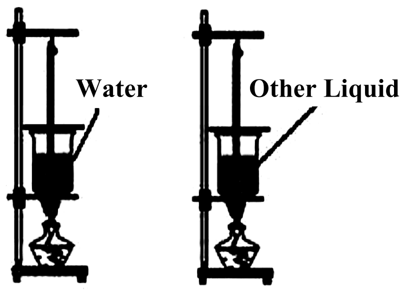


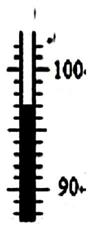
Answer

From the figure, it can be seen that the boiling point of water is 98°C. The boiling point of water under standard atmospheric pressure is 100°C. The boiling point of a liquid decreases as the pressure decreases, therefore the atmospheric pressure at that time is lower than 1 standard atmospheric pressure.
So the answer is: 98; lower.

Question

Some students in a group conducted an experiment to "compare the heat absorption capacity of different substances" using the setup shown in the figure.
(1) When designing the experimental plan, the following controlled variables need to be determined. Which one do you think is unnecessary?
A. Use the exact same heating method
B. Use the same amount of alcohol in the alcohol lamps
C. Use the same mass of water and another liquid
D. Use the same container to hold water and another liquid
(2) When heated to a certain point, the water starts boiling. The temperature at this moment is shown in Figure C. Therefore, the boiling point of water is °C, which indicates that the atmospheric pressure during the experiment is (choose "greater than", "less than", or "equal to") one standard atmosphere.
(3) The other liquid does not boil at the corresponding moment, but the reading of the thermometer is much higher than the water temperature. Based on this phenomenon, the preliminary conclusion of this experiment is: the heat absorption capacity of different substances is (choose "the same" or "different").





Answer

(1) B (2) 97, less than (3) different

Figure 18: Cases of *thermodynamics* in physics part of VisScience.

Question

The apparatus designed by an extracurricular group for synthesizing ethyl acetate in the laboratory is shown in the figure. Concentrated sulfuric acid is placed in A, ethanol and anhydrous sodium acetate are placed in B, and a saturated sodium carbonate solution is placed in D.

Known information:

① Anhydrous calcium chloride can form insoluble $CaCl2·6C2H5OH$ with ethanol.

② Boiling points of related organic compounds:

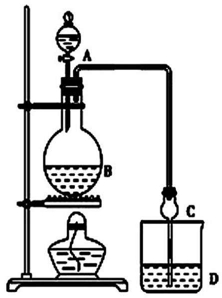
Reagent Ether Ethanol Acetic acid Ethyl acetate

Boiling point/°C 34.7 78.5 118 77.1

Please answer:

(1) The role of concentrated sulfuric acid is ________; if isotope 18O tracing method is used to determine the provider of oxygen atoms in the water molecules produced, write the chemical equation indicating the position of 18O: __________.

(2) The role of spherical drying tube C is ______________________. If a few drops of phenolphthalein are added to D before the reaction, the solution appears red; the reason for this phenomenon (expressed in ionic equation) is ________; the phenomenon in D after the reaction is __________________.



Answer

(1) Catalyst, dehydrating agent

CH3COOH + C2H518OH CH3CO18OC2H5 + H2O

(2) Prevent backflow, condensation CO + H2O⇌HCO + OH− The solution layers, with the upper layer being a colorless oily liquid and the lower layer solution becoming lighter in color

Question

To purify solid potassium nitrate containing a small amount of potassium chloride, a student conducts the experiment shown in the figure. Answer the following questions:

(1) Place the sample in a beaker and add an appropriate amount of water to dissolve it, while stirring with a glass rod. The purpose of stirring is
____.

(2) Add an appropriate amount of ________ solution to the solution to convert potassium chloride into a precipitate.

(3) Filter the mixture using the filtration apparatus and operations shown in the figure, and identify two errors in the figure:

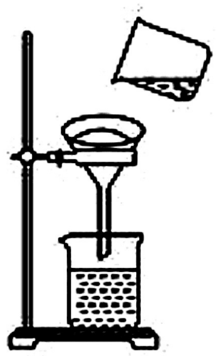
① ____;

② ____.

(4) To obtain potassium nitrate crystals from the filtrate, two methods that can be selected are:

① ____;

② ____.



Answer

(1) Speed up the dissolution of a solid

(2) Silver nitrate

(3) ① The liquid was not guided by a glass rod

② The lower end of the funnel was not close to the inner wall of the beaker

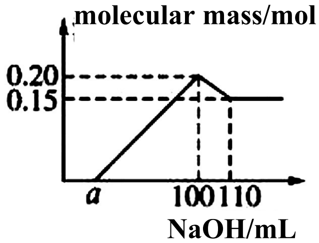
(4) ① Cool the hot saturated solution

② Evaporate the solvent

Figure 19: Cases of *chemical experiment* in chemistry part of VisScience.

Question

A certain experimental group conducted the following analysis on an unknown solution containing Al^{3+}:
(1) Added a small amount of sodium hydroxide, no obvious change;
(2) Continued adding NaOH solution, a white precipitate appeared;
(3) Added an excess of sodium hydroxide, the white precipitate noticeably decreased.
Through quantitative analysis, the group determined the relationship between the precipitate and the volume of sodium hydroxide added, as shown in the figure.
The following statements are incorrect ( )



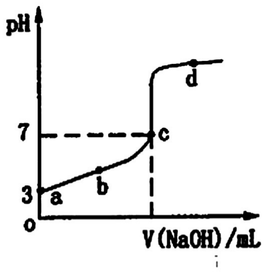
A: The unknown solution contains at least 3 cations
B: The concentration of the added NaOH solution is 5 mol·L⁻¹
C: If another ion is a divalent cation, then a = 10
D: If the final precipitate is filtered, washed, and ignited, its mass is certainly 6 g

Answer

D

Question

At room temperature, gradually add 0.1 $mol \cdot L^{-1}$ NaOH solution to 20 mL N $mol \cdot L^{-1}$ CH_3COOH solution; the pH curve is shown in the figure. Which of the following statements is correct?



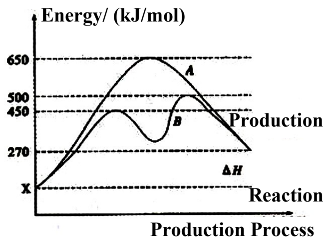
A. At point a, $c(CH_3COOH) = 10^{-3} mol \cdot L^{-1}$
B. At point b, the ion concentrations satisfy the relationship: $c(Na^{+})>c(CH_3COO^{-})>c(H^{+})>c(OH^{-})$
C. At point c, the ion concentrations satisfy the relationship: $c(Na^{+})+c(OH^{-})=c(H^{+})+c(CH_3COO^{-})$
D. During the titration process, it is impossible to have: $c(CH_3COOH)>c(CH_3COO^{-})>c(H^{+})>c(Na^{+})>c(OH^{-})$

Answer

C

Question

(1) Given: ①$Fe(s)+1/2O_2(g)=FeO(s)$ $\triangle H_1=-272.0 KJ \cdot mol^{-1}$
②$2Al(s)+3/2(g)=Al_2O_3(s)$ $\triangle H_2=-1675.7 KJ \cdot mol^{-1}$
The thermochemical equation for the thermite reaction between Al and FeO is ____.
A student believes that the thermite reaction can be used for industrial iron smelting. Your judgment is_ (fill in "can" or "cannot"), your reason is



(2) For a certain reversible reaction where both reactants and products are in the gaseous state, the reaction pathways under different conditions are respectively A and B, as shown in the figure. ① According to the figure, determine if the reaction is (fill in "endothermic" or "exothermic"). When the reaction reaches equilibrium, keeping other conditions unchanged, increasing the temperature will (fill in "increase", "decrease" or "remain unchanged") the conversion rate of reactants.
② The pathway B indicates that the condition used for this reaction is (fill in the letter).
A. Increasing temperature B. Increasing the concentration of reactants
C. Lowering temperature D. Using a catalyst

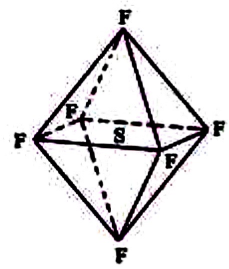
Answer

(1) $2Al(s) + 3FeO(s) = Al2O3(s) + 3Fe(s)$ $\triangle H$ = "-859.7" $kJ \cdot mol^{-1}$. No, this reaction needs to be initiated and requires a large amount of energy.
(2) ① Absorb; Decrease; ② D.

Figure 20: Cases of *chemical reaction* in chemistry part of VisScience.

Question

The sulfur hexafluoride molecule has an octahedral configuration (molecular structure as shown in the figure), is poorly soluble in water, has good insulation and flame retardant properties, and is widely used in the electrical industry. The correct conjecture is ( )



A: Each atom in SF₆ achieves a stable structure with 8 electrons

B: SF₆ dichloro derivatives have 3 kinds

C: The SF₆ molecule is a non-polar molecule with polar bonds

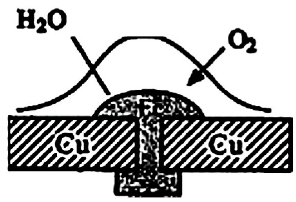
D: SF₆ is an atomic crystal

Answer

C

Question

Iron rivets nailed to the copper plate (as shown in the figure) are in a weakly acidic water film. The following statements are correct ( )



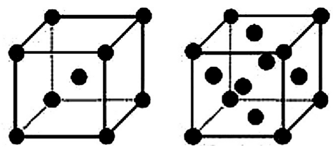
A: Iron rivets are not easily corroded in this environment
B: The chemical equation for the galvanic cell reaction: 2Fe+O<sub>2</ sub>+2H₂O═2Fe(OH)₂
C: The reaction occurring in the water film in contact with iron: O<sub>2</ sub>+4Fe[−]+2H₂O═4OH[−]
D: If in an acid rain environment, the following will occur: Cu − 2e<sup> − </ sup>═Cu²⁺

Answer

B

Question

(Total 7 points) Mn and Fe are both fourth-period transition elements. Answer the following questions:
(1) The electron configuration of the valence electrons of Mn element is __________
(2) Fe atoms or ions have more empty orbitals of similar energy in their outer layers and can form complexes with some molecules or ions.
① The structural characteristic that molecules or ions forming complexes with Fe atoms or ions should possess is __________
② The type of hybrid orbital of the C atom in the ligand CN^{-} in the hexacyanoferrate ion [Fe(CN)] is __________, write the structural formula of a simple molecule that is isoelectronic with CN^{-} __________
(3) Ferric chloride is a solid at room temperature, with a melting point of 282 °C and a boiling point of 315 °C, and sublimates easily above 300 °C. It is easily soluble in water as well as in organic solvents such as ether and acetone. Based on this, determine the crystal type of ferric chloride __________
(4) The crystal of metallic iron has two packing methods at different temperatures, and the unit cells are shown in the figure. The ratio of the actual number of Fe atoms in face-centered cubic unit cells to body-centered cubic unit cells is __________.



Body-centered cubic **Face-centered cubic**

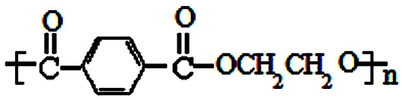
Answer

(1) 3d^{5}4s^{2}
(2) ① Contains a lone pair of electrons ② sp :N≡N:(3) Molecular crystal (4) 2:1

Figure 21: Cases of *inorganic chemistry* in chemistry part of VisScience.

Question

The mascot for the 2008 Olympics, Fuwa, has an outer material made of pure wool and is filled with non-toxic polyester fiber (as shown in the figure). Which of the following statements is correct? ( )
A. The chemical composition of wool and polyester fiber is the same.
B. Polyester fiber and wool can both hydrolyze under certain conditions.
C. The monomers of this polyester fiber are terephthalic acid and ethanol.
D. Polyester fiber is a pure substance.

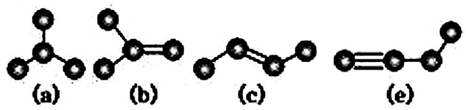


Answer

Question

The picture shows 4 organic compounds formed by 4 carbon atoms (hydrogen atoms are not drawn)

(1) Write the systematic name of organic compound (a)____________.
(2) Organic compound (a) has one isomer, write its structural formula ________________.
(3) Among the above organic compounds, the one that is an isomer of (c) is _______ (fill in the letter).
(4) Write the structural formula of any organic compound that is a homologue of (e)___________________.

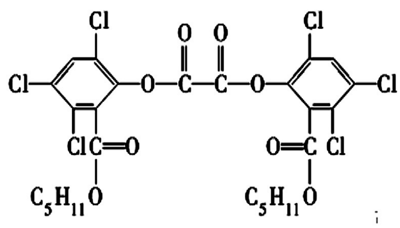


Answer

(1) 2-methylpropane
(2) CH_{3}CH_{2}CH_{2}CH_{3}
(3) (b)
(4) CH≡CH (or other reasonable answer)

Question

On summer nights, children are often seen playing with glowing "magic wands" in the square. The glowing principle of the "magic wand" is based on the oxidation of oxalyl ester by hydrogen peroxide, which generates energy. This energy is then transferred to fluorescent substances to emit fluorescence. The structural formula of oxalyl ester (CPPO) is shown in the figure. The correct statement regarding this is ( )
A: The molecular formula of oxalyl ester is C26H24Cl6O8
B: 1 mol of oxalyl ester reacts with dilute sodium hydroxide solution (halogen atoms on the benzene ring do not hydrolyze), consuming at most 4 mol of NaOH
C: The hydrolysis of oxalyl ester can yield two kinds of organic substances
D: 1 mol of oxalyl ester fully reacts with hydrogen gas, requiring 6 mol of hydrogen gas



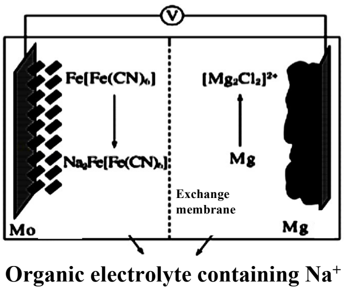
Answer

AD

Figure 22: Cases of *organic chemistry* in chemistry part of VisScience.

Question

The working principle of the new rechargeable sodium-ion battery represented by Prussian blue$${\rm Fe[Fe(CN)_{6}]}$$ during discharge is shown in the figure. The following statement is incorrect( )



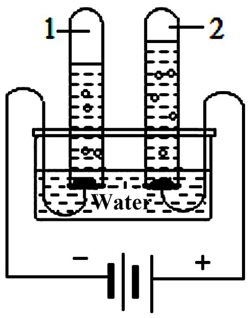
A: During discharge, the positive electrode reaction is$${\rm Fe[Fe(CN)_{6}] +2Na^{+}+2e^{-}=Na_{2}Fe[Fe(CN)_{6}]}$$
B: During charging, $${\rm Mo(}$$molybdenum$${\rm)}$$ foil is connected to the negative terminal of the power supply
C: During charging, $${\rm Na^{+}}$$ moves from the left chamber to the right chamber through the exchange membrane
D: When an amount of $${\rm 0.2mol}$$ electrons passes through the external circuit, the mass change of the negative electrode is $${\rm 2.4g}$$

Answer

B

Question

The electrolysis of water experiment reveals the composition of water. In the experiment shown in the figure, the substance obtained in test tube 1 is (fill in the chemical formula).

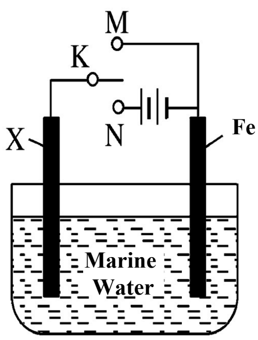


Answer

H_{2}

Question

Utilizing the apparatus shown in the figure, the electrochemical protection of iron can be simulated.
To reduce iron corrosion:
① If the switch $${\rm K}$$ is positioned at $${\rm N}$$, then $${\rm X}$$ should be ________________, and this electrochemical protection method is ________.
② If the switch $${\rm K}$$ is positioned at $${\rm M}$$, then $${\rm X}$$ should be ________________, and this electrochemical protection method is ________.



Answer

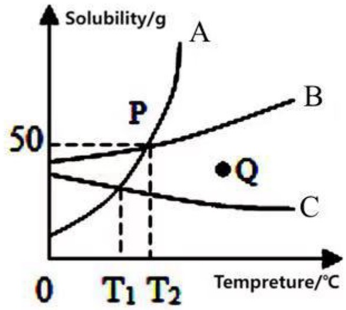
1. Cathodic protection method with external current using inert electrodes such as carbon rods or Pt.
2. Cathodic protection method with sacrificial anodes using metals more active than iron, such as Zn.

Figure 23: Cases of *electrochemistry* in chemistry part of VisScience.

Question

As shown in the figure, these are the solubility curves of three solid substances A, B, and C. The following statements are incorrect ( ).

A: The solubility of substance C decreases with the increase in temperature.
B: At T₁°C, the saturated solution of substance B has the highest mass fraction of solute.
C: Point P indicates that at T₁°C, the solubility of substances A and B are equal, both being 50%.
D: To keep the mass fraction of solute unchanged and make the solution of substance A at point Q reach saturation, the method of cooling should be adopted.



Answer

C

Question

The mineral water bottle labels the mineral content, as shown in the picture. The "potassium, magnesium, sodium, and calcium" mentioned here refer to ( )

A:elements B:molecules C:substances D:atoms

Mineral Content: mg/L	
Potassium 1-10	Magnesium 1-10
Sodium 1-16	Metasilicic Acid 35-70
Calcium 5-65	TDS 70-330

Answer

A

Question

The atomic structure of a certain element is shown in the figure. Xiaohong's mistake in understanding it is ( )

A. There are 12 protons in the nucleus
B. There are 3 electron shells around the nucleus
C. This element is a metal element
D. This atom carries a negative charge after losing electrons

Answer

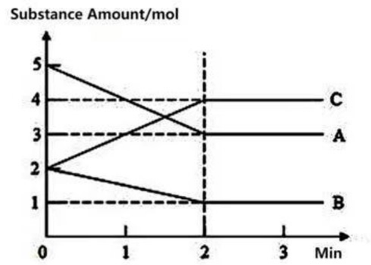
D

Figure 24: Cases of *substance composition* in chemistry part of VisScience.

Question

During the process of a certain reversible reaction from 0 to 2 minutes, the changes in the amounts of various substances at different reaction times are shown in the figure.

Thus, the reactants of the reaction are____, the products are_____, the chemical equation is_______; from the beginning of the reaction to 2 minutes, can the reaction rate be represented by C? _________ (If yes, calculate the reaction rate; if no, explain the reason) _______________. After 2 minutes, the amounts of substances A, B, and C no longer change with time, indicating that the reaction has reached a ___________ state under this condition.



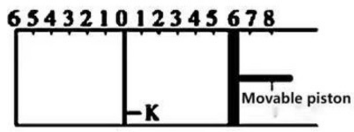
Answer

AB; C; not allowed; no volume; balance

Question

As shown in the figure, the partition K can move left and right. 2 mol of A and 1 mol of B are added to container A, and 2 mol of C and 1 mol of He are added to container B. At this time, K stops at position 0. The reaction occurs:

$2A(g) + B(g) \rightleftharpoons 2C(g)$, after reaching equilibrium and returning to temperature. The following statements are incorrect ( )



A. After reaching equilibrium, the partition K finally stops between the left scale 0 and 2.
B. If K stops at position 1 on the left at equilibrium, then the piston stops at position 6 on the right.
C. When equilibrium is reached, the amount of substance B in container A is less than the amount of substance B in container B.
D. Whether the partition K slides or not can determine whether the reactions on both sides have reached equilibrium.

Answer

B

Question

At 20 °C, two test tubes containing 10 g of water each were added with equal masses of solid substances A and B respectively, and both were fully dissolved, resulting in the phenomena as shown in the figure.

(1) At 20 °C, the solution in test tube (fill in "A" or "B") must be a saturated solution;

(2) The solubility curve of substance A is represented by (fill in "a" or "b"), and the method to make the remaining solid in test tube A continue to dissolve is .

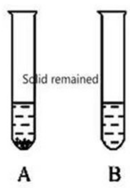


Figure 1

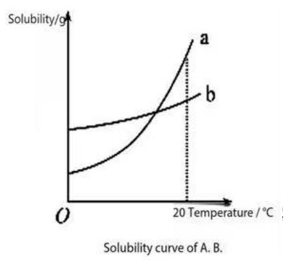


Figure 2

Answer

(1)A (2)b Increase the temperature or add more solvent

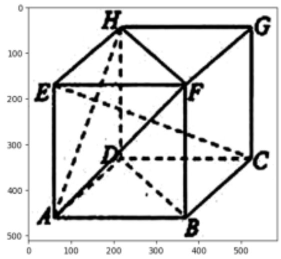
Figure 25: Cases of *chemical equilibrium* in chemistry part of VisScience.

Question

As shown in the figure, in the cube ABCD-EFGH, the following statements are incorrect:

A. BD // Face FHA
B. EC ⊥ BD
C. EC ⊥ Face FHA
D. The angle between the skew lines BC and AH is 60°

Ground Truth*: *D



GPT-4o ***Knowledge Error***

To determine which of the statements are incorrect based on the given cube ABCD-EFGH, let's analyze each statement:
Statement A: BD // Face FHA
This statement is incorrect. The line BD is a diagonal on the base ABCD of the cube, while the face FHA is a vertical face. BD does not lie parallel to face FHA and instead intersects it.
Statement B: EC ⊥ BD
Both EC and BD are diagonals drawn from one vertex to another in different planes but intersect at the point D. They are diagonal cross-sections of the cube, so this statement seems plausible. However, under cube geometry, two non-parallel lines do intersect, but they are not orthogonal.
Statement C: EC ⊥ Face FHA
This statement is correct. EC is perpendicular to face FHA because it runs horizontally from E to C, and the face FHA is vertical (runs from the top face to the bottom face).
Statement D: The angle between the skew lines BC and AH is 60°
Skew lines BC and AH do not intersect. The angle between skew lines is not simply determined without specific geometric conditions. Stating a specific angle like 60° is generally incorrect without calculation support.
So the correct identification of incorrect statements are:
A. Incorrect BD // Face FHA.
D. Incorrect angle specification between skew lines BC and AH at 60°.

Claude3.5 ***Vision Recognition Error***

Let's analyze each statement one by one:

A. BD // Face FHA
This statement is incorrect. BD is a diagonal of the cube that passes through the center, while face FHA is a plane. A line cannot be parallel to a plane unless it's parallel to a line in that plane. In this case, BD intersects face FHA at point D, so it's not parallel.
B. EC ⊥ BD
This statement is correct. EC and BD are both diagonals of the cube that pass through its center. In a cube, these diagonals are perpendicular to each other.
C. EC ⊥ Face FHA
This statement is incorrect. EC is not perpendicular to face FHA. For a line to be perpendicular to a plane, it must be perpendicular to all lines in that plane passing through their point of intersection. EC intersects face FHA at point E, but it's not perpendicular to all lines in FHA passing through E.
D. The angle between the skew lines BC and AH is 60°
This statement is correct. In a cube, the angle between two skew edges (edges that don't intersect and aren't parallel) is indeed 60°.

Therefore, the incorrect statements are A and C.

Qwen-VL-Max ***Calculation Error***

In the given image, we can see a cube with labeled vertices. Let's analyze each statement:

A. BD // Face FHA - This statement is correct because line segment BD is parallel to face FHA. Both are perpendicular to the common edge BF.
B. EC ⊥ BD - This statement is also correct because line segment EC is perpendicular to line segment BD. They form a right angle at vertex E.
C. EC ⊥ Face FHA - This statement is not accurate. While EC is perpendicular to some edges of face FHA (such as EF), it is not perpendicular to the entire face FHA.
D. The angle between the skew lines BC and AH is 60° - This statement is incorrect. In a cube, the angle between skew lines BC and AH is actually 90 degrees, not 60 degrees.

Therefore, the incorrect statement is option D. The angle between the skew lines BC and AH is 60°.

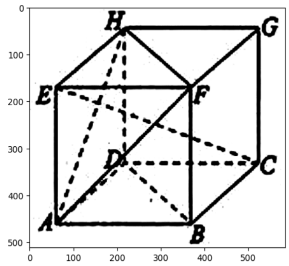
Figure 26: Cases of errors in the mathematical of VisScience for several classic close-source models.

Question

As shown in the figure, in the cube ABCD-EFGH, the following statements are incorrect:

A. BD // Face FHA
B. EC ⊥ BD
C. EC ⊥ Face FHA
D. The angle between the skew lines BC and AH is 60°

Ground Truth： D



LLaVA-1.5_13B ***Reasoning Error***

The incorrect statement in the figure is:

A. BD // Face FHA

In the cube ABCD-EFGH, the statement "BD // Face FHA" is incorrect. The correct statement would be "BD // Face FGH."

GLM-4V-9B ***Knowledge Error***

The given information states that the angle between the skew lines BC and AH is 60°. However, this angle cannot be determined based on the given information about the cube ABCD-EFGH. The correct answer is therefore D.

InternVL 1.5 ***Vision Recognition Error***

internvl_answer:
The image shows a cube with vertices labeled A, B, C, D, E, F, G, and H. The cube is a three-dimensional figure where all faces are squares, and all edges are of equal length. The diagonals of each face are also of equal length and are perpendicular to each other.

Let's analyze each statement:
A. BD // Face FHA
This statement is incorrect. In a cube, the diagonals of each face are perpendicular to each other. Since BD is a diagonal of the face FHA, it cannot be parallel to the face itself.
B. EC ⊥ BD
This statement is incorrect. In a cube, all diagonals are perpendicular to each other. Since EC is a diagonal of the face FHA, it is perpendicular to BD.
C. EC ⊥ Face FHA
This statement is incorrect. In a cube, the diagonals of each face are perpendicular to the face itself. Since EC is a diagonal of the face FHA, it is perpendicular to the face.
D. The angle between the skew lines BC and AH is 60°
This statement is incorrect. In a cube, all edges are of equal length, and the diagonals of each face are perpendicular to each other. The angle between the skew lines BC and AH is not 60°; it is 90°, as they are perpendicular to each other.

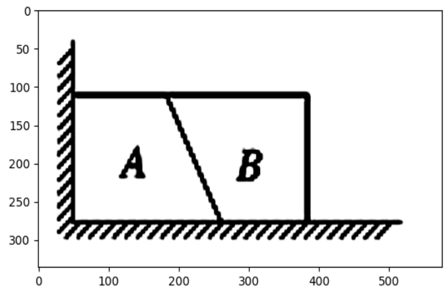
Therefore, all the statements are incorrect.

Figure 27: Cases of errors in the mathematical of VisScience for several classic open-source models.

Question

As shown in the figure, object A is in close contact with the horizontal surface, the vertical wall, and object B. All contact surfaces are smooth, and both A and B are stationary. Which of the following statements are correct?

A: Object A is subjected to three normal forces.
B: Object A is subjected to two normal forces.
C: Object B is subjected to two normal forces.
D: Both object A and object B are subjected to only one normal force.



Ground Truth*: *D

GPT-4o ***Reasoning Error***

The problem involves analyzing the normal forces acting on objects A and B that are stationary and in contact with smooth surfaces. Since the surfaces are smooth, there will be no frictional forces, only normal forces.
Let's analyze each object:
Object A:
- A is in contact with the horizontal surface. Therefore, there will be a normal force exerted by the horizontal surface on A.
- A is in contact with a vertical wall. Hence, there will be a normal force exerted by the wall on A.
- A is in contact with object B. Thus, there will be a normal force exerted by B on A.
So, object A experiences three normal forces.
Object B:
- B is in contact with the horizontal surface. Therefore, there will be a normal force exerted by the horizontal surface on B.
- B is in contact with object A. Hence, there will be a normal force exerted by A on B.
So, object B experiences two normal forces.
Thus, the correct options are:
- A: Object A experiences three normal forces.
- C: Object B experiences two normal forces.

Claude3.5 ***Vision Recognition Error***

Let's analyze this problem step by step:
1. Object A:
   - It's in contact with the horizontal surface (floor)
   - It's in contact with the vertical wall
   - It's in contact with object B
   Each of these contacts will result in a normal force acting on object A.
2. Object B:
   - It's in contact with the horizontal surface (floor)
   - It's in contact with object A
   Each of these contacts will result in a normal force acting on object B.
3. Remember, a normal force is always perpendicular to the surface of contact.
Given this analysis, we can conclude that:
- Object A experiences three normal forces: from the floor, from the wall, and from object B.
- Object B experiences two normal forces: from the floor and from object A.
Therefore, the correct answer is option A: Object A experiences three normal forces.

Qwen-VL-Max ***Reasoning Error***

Object A is in close contact with the horizontal surface, the vertical wall, and object B. According to the definition of the normal force, when an object is in close contact with another object or surface, it will experience a normal force from that other object or surface. Therefore, object A experiences three normal forces.

So the correct answer is A.

GLM-4V ***Knowledge Error***

Solution: Since all contact surfaces between objects A and B are smooth, and both objects are stationary, we have FN1=FN2 and FN3=0. Therefore, object A experiences two normal forces, and object B also experiences two normal forces. Hence, option C is selected.
Answer: C

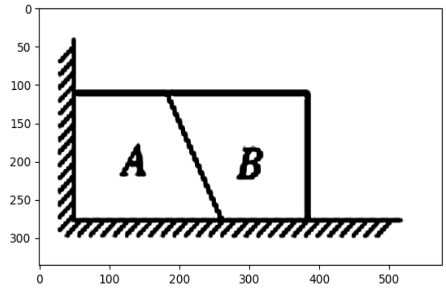
Figure 28: Cases of errors in the physics part of VisScience for several classic close-source models.

Question

As shown in the figure, object A is in close contact with the horizontal surface, the vertical wall, and object B. All contact surfaces are smooth, and both A and B are stationary. Which of the following statements are correct?

A: Object A is subjected to three normal forces.
B: Object A is subjected to two normal forces.
C: Object B is subjected to two normal forces.
D: Both object A and object B are subjected to only one normal force.

Ground Truth：D



LLaVA-1.5_13B ***Reasoning Error***

A: Object A experiences three normal forces.

GLM-4V ***Vision Recognition Error***

The diagram shows object A in contact with the horizontal surface, the vertical wall, and object B. Since all contact surfaces are smooth and both objects are stationary, we can apply the principles of physics to determine the normal forces acting on them.
Step 1: Analyze the forces acting on object A.
Object A is in contact with the horizontal surface and the vertical wall. The normal force acting on object A from the horizontal surface is perpendicular to the surface and directed upwards. The normal force acting on object A from the vertical wall is perpendicular to the wall and directed away from the wall. Therefore, object A experiences two normal forces.
Step 2: Analyze the forces acting on object B.
Object B is in contact with the horizontal surface and object A. The normal force acting on object B from the horizontal surface is perpendicular to the surface and directed upwards. The normal force acting on object B from object A is perpendicular to the contact surface between them and directed away from object A. Therefore, object B experiences two normal forces.
Step 3: Compare the normal forces acting on object A and object B.
Both object A and object B experience two normal forces each. The normal force acting on object A from the vertical wall is equal in magnitude and opposite in direction to the normal force acting on object B from object A. However, the normal forces acting on object A and object B from their respective surfaces are independent of each other.
Therefore, the answer is B: Object A experiences two normal forces

InternVL 1.5 ***Reasoning Error***

A: Object A experiences three normal forces.

Figure 29: Cases of errors in the physics part of VisScience for several classic open-source models.

Question

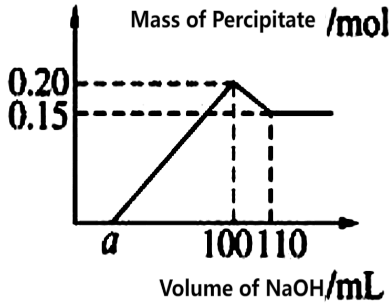
A certain experimental group conducted the following analysis on an unknown solution containing Al^{3+}:

(1) A small amount of sodium hydroxide was added, with no obvious change observed;
(2) More sodium hydroxide solution was added, resulting in a white precipitate;
(3) An excess amount of sodium hydroxide was added, and the white precipitate significantly decreased.

The experimental group performed a quantitative analysis and obtained the relationship between the amount of precipitate and the volume of sodium hydroxide added, as shown in the figure.

Which of the following statements is incorrect?

A: The unknown solution contains at least three types of cations.
B: The molarity of the added sodium hydroxide solution is 5 mol·L⁻¹.
C: If the other ion is a divalent cation, then a = 10.
D: If the final precipitate is filtered, washed, and calcined, its mass must be 6 g.



Ground Truth*：*D

GPT-4o ***Reasoning Error***

To find the incorrect statements, we need to delve into the chemical principles demonstrated and analyze the data provided in the graph systematically.
Here are the detailed steps and reasoning regarding each statement:

1. Unknown Solution Analysis:
The unknown solution is known to contain Al^{3+} ions which precipitate as $Al(OH)_3$ when NaOH is added.
Initial addition of NaOH: No obvious change, indicating that the OH^- ions are not sufficient to exceed the solubility product (Ksp) of $Al(OH)_3$.
Further addition of NaOH:A white precipitate of $Al(OH)_3$ appears, indicating the concentration of OH^- has now exceeded the Ksp of $Al(OH)_3$.
Excess NaOH: The white precipitate dissolves due to the formation of soluble complex ion $[Al(OH)_4]^-$.
2. Quantitative Relationship from the Graph:
The graph shows the relationship between the precipitate amount (in moles) and the NaOH added volume (in mL):
- Maximum precipitate mol: 0.20 mol
- Corresponding volume of NaOH: 100 mL
3. Evaluation of Statements:
(A) The unknown solution contains at least 3 cations.
This is not necessarily true based on the given information. The data provided only indicates the presence of Al^{3+}; no information about the presence of other cations can be conclusively determined. Therefore, this statement can be incorrect.
(B) The concentration of the added NaOH solution is 5 mol·L⁻¹.
Given that 0.20 mol of $Al(OH)_3$ precipitate appears when 100 mL of NaOH is added, and knowing the stoichiometry of the reaction: $Al^{3+} + 3OH^- \rightarrow Al(OH)_3$
0.2 mol Al^{3+} corresponds with 3 * 0.2 mol = 0.6 mol OH^-. For 100 mL NaOH to provide 0.6 mol OH^-, the concentration of NaOH must indeed be:
Concentration = 0.6 mol / 0.1 L = 6 mol·L⁻¹
This conflicts with 5 mol·L⁻¹. Thus, this statement is indeed incorrect.
(C) If another ion is a divalent cation, then a = 10.
For divalent cation, such as M^{2+}, precipitation occurs as $M(OH)_2$. Considering the graph where maximum precipitate occurs around 0.15 mol at some volume 'a', assuming it's from $M(OH)_2$: $M^{2+} + 2OH^- \rightarrow M(OH)_2$
At point 'a', V_a * C_NaOH ≥ (0.15 * 2 = 0.3 mol OH) / 5 mol·L⁻¹ = 0.06 L NaOH = 60 mL
Thus a ≈ 60 mL, not 10 mL, making this statement incorrect.
(D) If the final precipitate is filtered, washed, and ignited, its mass is certainly 6 g.
Upon ignition, $Al(OH)_3$ decomposes into Al_2O_3: $2Al(OH)_3 \rightarrow Al_2O_3 + 3H_2O$
The molar mass of Al_2O_3 is 102 g/mol. With 0.20 mol of $Al(OH)_3$ yielding 0.10 mol Al_2O_3: 0.10 mol * 102 g/mol = 10.2 g

Thus, this treatment yields about 10.2 g of Al_2O_3 and not 6 g, making this statement incorrect.
Conclusion: The following statements are all incorrect: A, B, C, and D.
E: None of the statements are correct

Claude3.5 ***Knowledge Error***

Let's analyze each statement based on the information provided and the graph:

A. The unknown solution contains at least 3 cations:
This statement is likely correct. The behavior described (initial precipitation followed by dissolution in excess base) is characteristic of amphoteric hydroxides like Al(OH)3. The graph shows a single precipitate forming and then partially dissolving, which is consistent with the presence of Al3+. There's no clear evidence of other cations.
B. The concentration of the added NaOH solution is 5 mol·L-1:
We can't determine this from the given information. The graph shows volume of NaOH added, not its concentration.
C. If another ion is a divalent cation, then a = 10:
We don't have enough information to confirm or deny this statement. The value of 'a' isn't directly related to the presence of a divalent cation based on the given data.
D. If the final precipitate is filtered, washed, and ignited, its mass is certainly 6 g:
This statement is incorrect. The graph shows the amount of precipitate in moles, not grams. At the end point, there's 0.15 mol of precipitate. Without knowing the exact composition of the precipitate after ignition, we can't determine its mass.

Given the question asks which statements are incorrect, the answer should include B and D. Statement A appears to be correct based on the information given, and C cannot be confirmed or denied with the available data.

Figure 30: Cases of errors in the chemistry part of VisScience for several classic close-source models.

Question

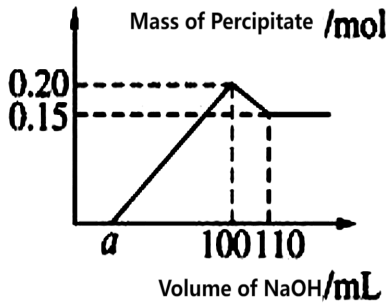
A certain experimental group conducted the following analysis on an unknown solution containing Al^{3+}:

(1) A small amount of sodium hydroxide was added, with no obvious change observed;
(2) More sodium hydroxide solution was added, resulting in a white precipitate;
(3) An excess amount of sodium hydroxide was added, and the white precipitate significantly decreased.

The experimental group performed a quantitative analysis and obtained the relationship between the amount of precipitate and the volume of sodium hydroxide added, as shown in the figure.

Which of the following statements is incorrect?

A: The unknown solution contains at least three types of cations.
B: The molarity of the added sodium hydroxide solution is 5 $mol \cdot L^{-1}$.
C: If the other ion is a divalent cation, then a = 10.
D: If the final precipitate is filtered, washed, and calcined, its mass must be 6 g.



Ground Truth*：*D

Qwen-VL-Max ***Knowledge Error***

The image shows a graph that plots the amount of precipitation against the volume of sodium hydroxide (NaOH) added to an unknown solution. The x-axis represents the volume of NaOH added in mL, while the y-axis indicates the amount of precipitation in moles.

From the graph, we can see that when a small amount of NaOH is added, there is no visible change in the solution. As more NaOH is added, a white precipitate appears. When an excess of NaOH is added, the white precipitate decreases significantly.

Based on this information, let's analyze each statement:

A: The unknown solution contains at least 3 cations.
This statement is correct because the addition of a small amount of NaOH does not cause any reaction, indicating that some ions do not react with NaOH. Then, a white precipitate forms upon further addition of NaOH, suggesting the presence of at least one cation that reacts with NaOH. Finally, the decrease in the precipitate after adding an excess of NaOH implies the existence of another cation that also reacts with NaOH but has a different behavior than the first one. Therefore, it is reasonable to conclude that the solution contains at least three cations.
B: The concentration of the added NaOH solution is 5 mol·L^-1.
This statement is incorrect. Although the graph provides data points for the amount of precipitation and the volume of NaOH added, it does not provide enough information to determine the concentration of the NaOH solution. To calculate the concentration, we would need additional details such as the initial volume or mass of the NaOH solution used.
C: If another ion is a divalent cation, then a = 10.
This statement is correct. According to the graph, the volume of NaOH required to form the maximum amount of precipitation is around 100 mL. Since the graph shows that the amount of precipitation increases linearly with the volume of NaOH added until reaching a maximum value, we can assume that the reaction follows a simple stoichiometry. If the other ion is a divalent cation, it will require twice as much NaOH to form the same amount of precipitation compared to a monovalent cation. Thus, if the volume of NaOH needed to reach the maximum precipitation was 100 mL for a monovalent cation, it would be 200 mL for a divalent cation. However, since the graph shows that the maximum precipitation occurs at around 100 mL, we can infer that the other ion must be a monovalent cation. In this case, a = 10, which corresponds to the point where the slope changes from positive to negative.
D: If the final precipitate is filtered, washed, and ignited, its mass is certainly 6 g.
This statement is incorrect. While the graph shows that the maximum amount of precipitation formed is approximately 0.15 mol, we cannot directly convert this into grams without knowing the molar mass of the precipitated compound. Furthermore, filtering, washing, and igniting the precipitate may result in losses due to incomplete recovery or decomposition during heating. Therefore, we cannot confidently predict the exact mass of the final product based solely on the given information.

In conclusion, the correct answer is B: The concentration of the added NaOH solution is 5 mol·L^-1. This statement is incorrect because there is insufficient information provided to determine the concentration of the NaOH solution.

GLM-4V ***Reasoning Error***

Solution: According to the graph, when 10 mL of NaOH is added, the precipitation does not dissolve, indicating that the concentration of aluminum ions is less than 0.20 mol/L. When 11 mL of NaOH is added, the precipitation begins to dissolve, which means that the concentration of aluminum ions is greater than 0.20 mol/L but less than 0.25 mol/L. Since the initial volume of the solution is 100 mL and the final volume after adding NaOH is 110 mL, the total volume of the solution is 210 mL. Therefore, the concentration of sodium hydroxide is 0.20 mol/L. Thus, option B is incorrect.

When 11 mL of NaOH is added, the precipitation begins to dissolve, which means that the concentration of aluminum ions is greater than 0.20 mol/L but less than 0.25 mol/L. Therefore, the concentration of sodium hydroxide is 0.20 mol/L. Thus, option B is incorrect.

The mass of the precipitate before heating is 8 g, so the mass of the precipitate after heating is 6 g. Therefore, option D is incorrect.

Therefore, the answer is C.
Answer: C

Figure 31: Cases of errors in the chemistry part of VisScience for several classic close-source models.

Question

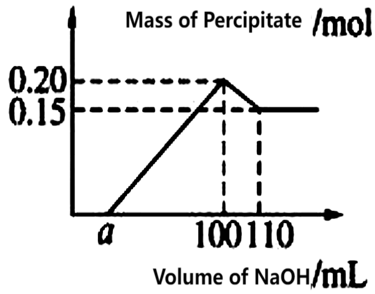
A certain experimental group conducted the following analysis on an unknown solution containing Al^{3+}:

(1) A small amount of sodium hydroxide was added, with no obvious change observed;
(2) More sodium hydroxide solution was added, resulting in a white precipitate;
(3) An excess amount of sodium hydroxide was added, and the white precipitate significantly decreased.

The experimental group performed a quantitative analysis and obtained the relationship between the amount of precipitate and the volume of sodium hydroxide added, as shown in the figure.

Which of the following statements is incorrect?

A: The unknown solution contains at least three types of cations.
B: The molarity of the added sodium hydroxide solution is 5 $mol \cdot L^{-1}$.
C: If the other ion is a divalent cation, then a = 10.
D: If the final precipitate is filtered, washed, and calcined, its mass must be 6 g.



Ground Truth：D

LLaVA-1.5_13B   Knowledge Error

The incorrect statement is B: The concentration of the added NaOH solution is 5 $mol \cdot L^{-1}$. The actual concentration of the NaOH solution is 10 $mol \cdot L^{-1}$.

GLM-4V   Knowledge Error

The graph shows that the precipitate decreases after the addition of an excess of sodium hydroxide, which indicates that the precipitate is not due to the presence of a single cation. Therefore, option A is incorrect as it states that the unknown solution contains at least 3 cations.

The graph does not provide any information about the concentration of the NaOH solution. It only shows the volume of NaOH added and the corresponding change in the precipitate. Therefore, option B is incorrect as it states the concentration of the NaOH solution is 5 mol·L^-1.

Option C is incorrect because the graph does not provide enough information to determine the value of a if another ion is a divalent cation.

Option D is correct because if the precipitate is filtered, washed, and ignited, its mass will be the mass of the hydrated oxide, which is the product of the original cation and the hydroxide ion. Since the graph shows that the precipitate decreases after the addition of an excess of sodium hydroxide, it is reasonable to assume that the precipitate is hydrated oxide. Therefore, the mass of the precipitate will be the product of the concentration of the original cation and the hydroxide ion, which is 6 g.

Therefore, the answer is D

InternVL 1.5   Knowledge Error

First, we need to understand the information represented in the chart. The chart shows the relationship between the amount of precipitate in an unknown solution and the volume of NaOH added. According to the chart, the amount of precipitate reaches its maximum when the volume of NaOH added is between 0.15 mL and 0.20 mL.

Now, let's analyze each option:

A: The unknown solution contains at least three types of cations. This statement is correct because the increase in the amount of precipitate with the addition of NaOH indicates the presence of at least three cations, as Al^{3+} requires three NaOH molecules to completely precipitate.

B: The concentration of the added NaOH solution is 5 $mol \cdot L^{-1}$. This statement is incorrect because the chart does not provide any information about the concentration of the NaOH solution.

C: If the other ion is a divalent cation, then a = 10. This statement is incorrect because the chart shows that the amount of precipitate reaches its maximum when the volume of NaOH added is 0.15 mL, indicating the presence of a divalent cation, but the value of a is not 10.

D: If the final precipitate is filtered, washed, and calcined, its mass must be 6 g. This statement is incorrect because the chart does not provide any information about the mass of the precipitate.

Therefore, option B is incorrect.

Figure 32: Cases of errors in the chemistry part of VisScience for several classic open-source models.

Model	Physics								
	ALL	Mech	Ele	Therm	Comp	Opt	Vib & Waves	Mod Phys	Mech Motion
				Close Source Models (APIs)					
Text-only LLMs									
Zero-shot ChatGPT	22.70	22.08	19.94	23.53	4.62	40.98	29.79	19.05	23.33
Zero-shot GPT-4	30.40	34.26	30.21	33.33	15.38	40.98	34.04	42.86	20.00
Zero-shot Claude-2	22.00	24.62	23.56	25.49	12.31	27.87	21.28	28.57	23.33
Zero-shot Claude3-Opus	30.80	34.26	32.02	33.33	10.77	39.34	31.91	42.86	10.00
Zero-shot Claude3.5-Sonnet	35.30	40.36	35.95	35.29	15.38	40.98	34.04	47.62	26.67
Zero-shot GPT-4o	38.00	43.91	38.67	45.10	9.23	49.18	38.30	52.38	23.33
2-shot CoT Claude2	21.70	24.87	22.96	25.49	10.77	18.03	23.40	28.57	10.00
2-shot CoT ChatGPT	18.60	20.30	20.54	13.73	12.31	22.95	23.40	23.81	13.33
2-shot CoT GPT-4	31.50	35.03	32.02	37.25	12.31	44.26	29.79	47.62	23.33
Multi-modal LLMs									
Gemini-1.0-Pro	23.70	26.97	23.03	17.65	6.15	31.15	34.04	19.05	10.00
Gemini-1.5-Pro	38.10	46.56	33.74	47.06	20.00	45.00	34.04	52.38	43.33
GPT-4o	38.20	41.37	39.27	56.86	23.08	42.62	36.17	42.86	43.33
GPT-4o-mini	29.80	31.73	30.51	29.41	10.70	36.07	19.15	47.62	30.00
Qwen-VL-Max	30.70	36.13	26.59	39.22	9.23	34.43	31.91	28.57	30.00
Qwen-VL-Plus	26.50	31.04	24.77	33.33	6.15	36.07	36.17	23.81	16.67
Claude3.5-Sonnet	38.00	41.62	36.56	43.14	13.85	44.26	38.30	66.67	30.00
Claude3-Opus	31.10	33.25	29.91	39.22	12.31	45.90	34.04	61.90	23.33
GLM-4V	19.20	23.16	17.82	15.69	12.31	25.00	17.02	19.05	23.33
Step-1V	23.50	21.55	24.35	28.57	7.84	12.82	25.00	31.25	39.13
				Open Source Models					
General Multi-modal LLMs									
mPLUG-Owl	8.30	11.93	8.46	1.96	4.62	8.20	10.64	4.76	10.00
LLaMA-Adapter-V2	10.30	10.41	10.88	8.00	4.84	13.11	25.53	14.29	3.33
MiniCPM-Llama3-V2.5	17.90	21.57	19.64	15.69	6.15	26.23	19.15	9.52	23.33
LLaVA-1.5-13B	15.20	17.26	14.80	7.84	7.69	21.31	17.02	9.52	16.67
LLaVA-1.5-7B	13.50	15.28	15.12	11.76	3.12	15.25	15.56	5.26	17.24
DeepSeek-VL	16.80	18.77	19.33	13.73	7.69	16.67	13.04	19.05	3.45
ShareGPT4V-7B	14.00	13.71	15.41	9.80	3.08	19.67	19.15	28.57	6.67
ShareGPT4V-13B	14.90	15.23	16.92	9.80	6.15	14.75	19.15	19.05	16.67
SPHINX-Plus	15.30	16.50	18.43	17.65	4.62	11.48	12.77	19.05	13.33
InternLM-XC2	18.30	20.81	17.82	13.73	10.77	26.23	21.28	14.29	6.67
InternVL-1.2-Plus	24.80	29.69	22.94	29.41	12.31	31.67	25.53	35.00	10.00
InternVL-Chat-V1.5	20.80	23.97	20.87	23.53	9.23	25.42	17.02	14.29	17.24
CogVLM	14.50	18.02	13.29	7.84	6.15	14.75	19.15	19.05	6.67
CogVLM2	14.40	16.75	16.00	12.00	6.15	13.11	19.15	4.76	10.00
GLM-4V-9B	19.30	21.78	21.12	24.00	4.62	25.42	15.91	15.00	13.33
InternVL2-8B	22.1	25.50	22.30	24.70	9.50	28.20	21.40	30.10	18.50
Qwen2-VL-7B	6.5	5.43	6.14	9.76	2.95	18.55	8.42	0.00	6.33
InternVL2.5-8B	23.5	27.00	23.80	26.20	11.00	29.50	22.10	31.50	19.00
Qwen2.5-VL-7B	6.80	5.58	6.34	9.80	3.08	18.03	8.51	0.00	6.67

Table 7: **Results on the physics part of VisScience across different subjects.** Subjects: Mech: mechanics, Ele: electromagnetism, Threm: thermodynamics, Comp: comprehensive experiments and methods, Opt: optics, Vib & Waves: vibration and waves, Mod Phys: modern physics, Mech Motion: mechanical motion. The highest scores among close-source and open-source models are highlighted in red and blue, respectively.

Model	ALL	Chem Exp	Chem React	Inorg Chem	Electrochem	Org Chem	Chem Equil	Sub Comp
	Chemistry							
Close Source Models (APIs)								
Text-only LLMs								
Zero-shot ChatGPT	18.60	26.35	23.86	23.26	23.75	35.43	24.64	30.89
Zero-shot GPT-4	33.10	40.54	30.68	38.76	32.50	36.22	30.43	31.71
Zero-shot Claude-2	24.40	24.32	26.36	31.71	23.86	20.29	0.30	25.98
Zero-shot Claude3-Opus	32.50	37.16	30.68	31.78	31.25	36.22	30.43	39.84
Zero-shot Claude3.5-Sonnet	36.90	34.80	36.93	39.53	46.25	45.67	23.19	47.15
Zero-shot GPT-4o	39.60	42.57	40.34	44.96	35.00	41.73	26.09	54.47
2-shot CoT Claude2	23.90	23.99	26.70	22.48	30.00	26.77	27.54	27.64
2-shot CoT ChatGPT	21.30	19.93	23.30	20.93	22.50	22.83	26.09	30.08
2-shot CoT GPT-4	32.40	29.05	32.39	32.56	32.50	42.52	28.99	53.66
Multi-modal LLMs								
Gemini-1.0-Pro	27.80	24.03	26.70	26.36	31.25	35.54	31.82	37.19
Gemini-1.5-Pro	47.00	43.46	47.43	51.59	50.00	57.02	35.29	61.16
GPT-4o	41.60	43.58	46.02	38.76	46.25	43.31	43.48	50.41
GPT-4o-mini	28.40	22.30	27.27	27.13	30.00	34.65	20.29	42.09
Qwen-VL-Max	42.50	46.28	41.48	51.94	35.00	41.73	36.23	53.66
Qwen-VL-Plus	37.70	33.78	40.34	44.19	41.25	48.03	33.33	41.80
Claude3.5-Sonnet	43.10	40.54	41.48	42.64	50.00	42.52	33.33	59.35
Claude3-Opus	34.10	35.47	30.11	31.78	31.25	40.16	33.33	51.22
GLM-4V	25.00	23.65	25.86	21.71	28.75	27.78	31.88	32.52
Step-1V	25.00	32.51	27.48	25.26	25.45	17.72	13.33	21.95
Open Source Models								
General Multi-modal LLMs								
mPLUG-Owl	9.50	7.77	11.36	7.75	12.50	12.60	13.04	9.76
LLaMA-Adapter-V2	10.80	7.77	13.64	8.53	12.66	12.80	17.65	17.07
MiniCPM-Llama3-V2.5	19.50	20.96	26.29	26.61	18.18	24.00	28.79	30.83
LLaVA-1.5-13B	18.80	15.54	16.48	24.03	20.00	22.05	23.19	19.51
LLaVA-1.5-7B	16.00	13.49	17.14	19.20	16.25	20.49	26.09	10.74
DeepSeek-VL	21.00	18.84	20.57	20.16	21.25	23.62	36.76	20.66
ShareGPT4V-7B	19.00	13.85	19.32	26.36	18.75	23.62	28.99	15.45
ShareGPT4V-13B	18.40	13.51	21.02	19.38	23.75	22.83	13.04	19.51
SPHINX-Plus	20.40	20.27	21.02	24.03	22.50	22.83	27.54	21.95
InternLM-XC2	25.60	22.64	27.27	26.36	21.25	33.86	26.09	24.39
InternVL-1.2-Plus	31.20	22.29	31.82	33.58	31.46	39.57	32.47	38.84
InternVL-Chat-V1.5	23.70	20.07	25.00	25.20	22.37	28.80	25.00	28.46
CogVLM	17.00	15.54	20.45	10.85	16.25	22.05	20.29	17.07
CogVLM2	21.00	13.10	21.39	25.78	20.51	31.45	22.73	30.17
GLM-4V-9B	22.50	21.00	25.44	26.23	23.08	26.83	17.39	25.83
InternVL2-8B	26.5	23.8	27.2	25.6	24.7	28.3	26.1	27.5
Qwen2-VL-7B	4.3	3.8	4.5	4.1	3.9	4.7	4.0	4.6
InternVL2.5-8B	27.7	24.5	28.5	26.8	25.9	29.2	27.3	28.8
Qwen2.5-VL-7B	4.94	2.03	3.97	3.10	3.75	3.93	0.00	18.70

Table 8: **Results on the chemistry part of VisScience across different subjects.** Subjects: Chem Exp: chemical experiment, Chem React: chemical reaction, Inorg Chem: inorganic chemistry, Electrochem: Electrochemistry, Org Chem: organic chemistry, Chem Equil: chemical equilibrium, and Sub Comp: substance composition. The highest scores among close-source and open-source models are highlighted in red and blue, respectively.