A. Details of Training Data

To train SCIEval, we utilize the ArXivCap dataset [21] as our foundational source, from which we generate dimension-specific positive and negative examples. ArXivCap is an extensive figure-caption collection derived from 572K ArXiv papers across 32 distinct scientific domains. It provides a robust training base, consisting of approximately 6.4M images and 3.9M captions.

To construct accuracy negatives, we generate a negative caption C_A by introducing minor revisions to the gold caption C_T. This process relies on the attribute dictionary D from ScImage [47], which categorizes essential elements of scientific figures such as objects (e.g., squares, circles), attributes (e.g., color, size), spatial relations (e.g., left, right), and numeric values. The revision process follows a structured sampling and template-based approach:

- Dictionary sampling. We first identify the relevant word class within D and randomly select a replacement item from that specific list.
- Template application. We utilize a set of predefined query templates—consisting of one or more sentences with placeholders—designed to house these objects, attributes, or relations.
- Final revision. By populating a chosen template with elements from D, we create a modified version of C_T, which is then formally recorded as the negative caption C_A.

B. Details of SCIEval-Bench

B.1. Evaluation sample generation

To evaluate generative LMMs within SCIEval-Bench, we employ three distinct output modes:

First, in the Direct Text-to-Image mode, the model synthesizes an image directly from a specific text prompt, such as "Please generate a scientific figure according to the following requirements: {Generation Instruction}."

Second, the Text-to-Code-to-Image mode introduces an intermediate step where the LMM generates executable code (either Python or TikZ) based on a query, such as "Please generate a scientific figure according to the following requirements: {Generation Query}. Your output should be in [Python/Tikz] code. Do not include any text other than the [Python/Tikz] code." This generated code is subsequently compiled to produce the final figure.

Third, in the Image-to-Caption mode, the LMM is tasked with generating an analysis rather than a visualization. Using a prompt such as "Generate a scientific-style caption for the given image", the model produces a descriptive caption for a provided figure.

B.2. Human Scoring Criteria

As shown in Table 4, we establish comprehensive scoring guidelines for evaluating instances within SCIEval-Bench.

B.3. More Examples

As illustrated in Figures 8 and 9, we provide more qualitative examples from SCIEval-Bench for each task.

Our analysis of the Sci-T2I setting (Fig.8) reveals several key findings:

- Instruction density. There is a direct correlation between prompt detail and output quality; more granular instructions (e.g., I1) consistently yield superior images.
- Performance gaps. Significant room for improvement remains in Sci-T2I generation. Even with straightforward prompts like I2, prominent LMMs—including Llama_tikz, DALL-E, and Stable Diffusion—might produce images with low relevance and accuracy.
- Leading architecture. Among the four candidates evaluated, Llama_python demonstrates the strongest performance, consistently generating the highest-quality scientific figures.

Our analysis of the Sci-IC setting (Fig.9) yields several insights into the captioning capabilities of current LMMs:

- Baseline performance. Most LMMs perform reliably on straightforward visual inputs (e.g., I1), producing satisfactory and accurate captions.
- Verbosity trends. Models such as DeepSeek-VL and Qwen-VL consistently produce more detailed and lengthy captions compared to IDEFICS-2 and LLaVA-1.6, as seen in examples I1 and I2.
- Inherent complexity. High-complexity images (e.g., I3) featuring visually indistinguishable curves represent a significant hurdle; under these conditions, all tested LMMs fail to exceed moderate-quality outputs.
- Top performer. Among the four candidates, DeepSeek-VL distinguishes itself by generating the highest-quality captions overall.

C. Details of Experimental Settings

C.1. Details of Evaluation Mechanism

To demonstrate statistical significance, we report the p-value, a metric that represents the probability of obtaining the observed results assuming that the hypothesis is true.

For the evaluation of rationale text, we employ an LMM-as-Judge approach (utilizing models such as GPT-4o), supplemented by rigorous manual inspections of both individual generated instances and aggregate performance. Detailed scoring criteria for automated and human assessment are provided in Table 5.

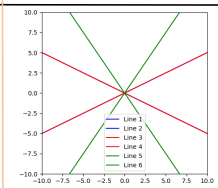
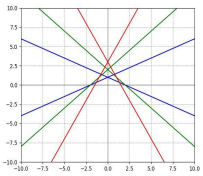
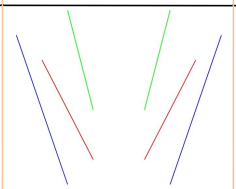
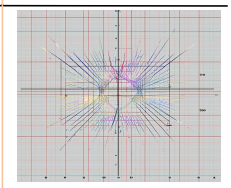
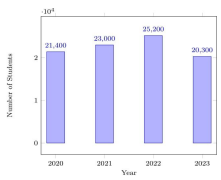
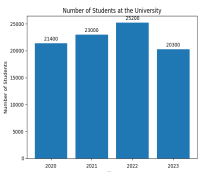
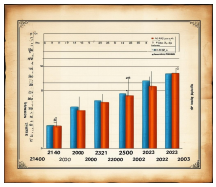
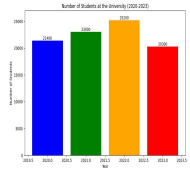
Instruction (I) \ Model (M)	Llama-python	Llama-tikz	DALL-E	Stable Diffusion
1 *Six straight lines are located in a coordinate system and are symmetrical about the Y-axis in pairs. Each pair shares the same color.*	 (I1, M1) Relevance Score: 5 4 5 Accuracy Score: 4 5 5	 (I1, M2) Relevance Score: 5 3 5 Accuracy Score: 5 4 5	 (I1, M3) Relevance Score: 4 3 2 Accuracy Score: 5 3 2	 (I1, M4) Relevance Score: 4 5 5 Accuracy Score: 3 5 4
2 *The number of students at the university was 21,400 in 2020, 23,000 in 2021, 25,200 in 2022, and 20,300 in 2023. The student numbers over these four years are represented in a bar chart. The values are also marked in the chart.*	 (I2, M1) Relevance Score: 4 5 5 Accuracy Score: 3 4 4	 (I2, M2) Relevance Score: 5 4 3 Accuracy Score: 4 3 4	 (I2, M3) Relevance Score: 3 3 1 Accuracy Score: 2 2 1	 (I2, M4) Relevance Score: 4 3 3 Accuracy Score: 4 3 3

Figure 8. Examples of Sci-T2I from SCIEval-Bench. For each line-wise example, we present an input instruction (I) and four output images generated by distinct LMMs (M), each with three relevance scores and three accuracy scores annotated by individual participants.

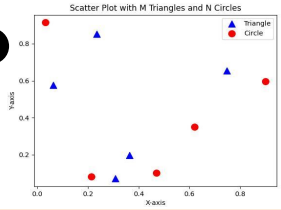
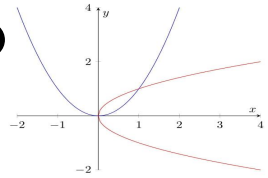
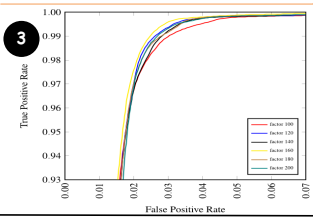
Image (I) \ Model (M)	IDEFICS-2	LLaVA-1.6	DeepSeek-VL	Qwen-VL
1 	*Scatter plot with M Triangles and N Circles.* (I1, M1) Relevance Score: 4 4 4 Accuracy Score: 3 4 3	*Scatterplots for Blue Triangles and Red Circles* (I1, M2) Relevance Score: 4 3 5 Accuracy Score: 4 4 3	*A scatterplot of a random distribution of different shapes, containing 5 red circles and 5 blue triangles.* (I1, M3) Relevance Score: 5 5 5 Accuracy Score: 5 5 5	*A scatter plot showing the random distribution of different shapes, containing 5 circles and 5 triangles.* (I1, M4) Relevance Score: 5 5 5 Accuracy Score: 4 5 4
2 	*Two intersecting curves in a coordinate system.* (I2, M1) Relevance Score: 2 3 2 Accuracy Score: 1 2 2	*Two arbitrary curves in a Cartesian coordinate system.* (I2, M2) Relevance Score: 3 3 3 Accuracy Score: 2 2 3	*In the Cartesian coordinate system, the blue curve represents the function y = x^2. The red curve may represent the function y = -1/2x, but this is uncertain.* (I2, M3) Relevance Score: 5 5 5 Accuracy Score: 4 4 5	*In the Cartesian coordinate system, an upward-pointing blue parabola and a rightward-pointing red parabola.* (I2, M4) Relevance Score: 5 4 5 Accuracy Score: 4 4 5
3 	*Curve Trend Chart for Different Factors.* (I3, M1) Relevance Score: 3 3 2 Accuracy Score: 2 2 2	*Arbitrary Curve Trend Chart for Different Factors.* (I3, M2) Relevance Score: 3 3 2 Accuracy Score: 1 2 2	*True Positive Rate vs. False Positive Rate Comparison Chart for Factors 100, 120, etc.* (I3, M3) Relevance Score: 3 3 3 Accuracy Score: 3 3 3	*Arbitrary curve trend plots for 6 different factors* (I3, M4) Relevance Score: 3 3 3 Accuracy Score: 2 3 2

Figure 9. Examples of Sci-IC from SCIEval-Bench. For each line-wise example, we present an input image (I) and four output captions generated by distinct LMMs (M), each with three relevance scores and three accuracy scores annotated by individual participants.

Table 4. Human scoring guidelines for data annotation of SCIEval-Bench.

Score	Description
	Sci-T2I Relevance Dimension
5	The image contains no redundant objects (e.g., geometric shapes, OCR, and data points, etc).
4	The image contains a few redundant objects but remains highly relevant to the text's requirements.
3	The image contains some redundant objects and some required elements.
2	The image contains more redundant objects than required elements.
1	The overall image is not relevant to the requirements.
	Sci-T2I Accuracy Dimension
5	The image fully meets all the requirements (objects, attributes, relations) with no mistakes.
4	The image meets the key requirements (objects, attributes, relations), with only minor mistakes.
3	The image meets some or half of the requirements (objects, attributes, relations), with some mistakes.
2	The image meets only a few of the text's requirements (objects, attributes, relations) and contains serious mistakes.
1	The image fails to meet the requirements (objects, attributes, relations) of the text.
	Sci-IC Relevance Dimension
5	The caption contains no redundant objects (e.g., geometric elements, text, and data, etc).
4	The caption contains a few redundant objects but remains highly relevant to the visual content.
3	The caption contains some redundant descriptions and some required elements.
2	The caption contains more redundant descriptions than required elements.
1	The overall caption is not relevant to the visual content.
	Sci-IC Accuracy Dimension
5	The caption fully depicts all visual content (objects, attributes, relations) with no mistakes.
4	The caption depicts the key visual content (objects, attributes, relations), with only minor mistakes.
3	The caption depicts some or half of visual content (objects, attributes, relations), with some mistakes.
2	The caption depicts only a little of visual content (objects, attributes, relations) and contains serious mistakes.
1	The caption fails to depict the visual content (objects, attributes, relations) of the text.

Table 5. Human scoring guidelines of rationale assessment for both Sci-T2I and Sci-IC tasks.

Score	Discription
	Correctness Dimension
5	All unfaithful elements revealed by the rationale are correct, with no mistakes.
4	The key unfaithful elements revealed by the rationale are correct, with only minor mistakes in other elements.
3	Some or half of unfaithful elements revealed by the rationale are correct, with some mistakes in other elements.
2	Only a few of unfaithful elements revealed by the rationale are correct and other elements contain serious mistakes.
1	No unfaithful elements revealed by the rationale are correct.
	Completeness Dimension
5	The rationale sufficiently reveals all unfaithful elements in the generated image without omissions.
4	The rationale reveals the key unfaithful elements in the generated image, with only minor omissions.
3	The rationale reveals some or half of unfaithful elements in the generated image, with some omissions.
2	The rationale reveals only a few of unfaithful elements in the generated image and contains serious omissions.
1	The rationale fails to reveal the unfaithful elements in the generated image.

Table 6. The versions for all models used in this paper.

Model	Version / HF Checkpoint
Closed-source LMMs	
GPT-4V [32]	*gpt-4-vision-preview*
GPT-4 Turbo [31]	*gpt-4-turbo-2024-04-09*
GPT-4o [33]	*gpt-4o-2024-05-13*
Gemini Pro Vision [1]	*gemini-pro-vision*
Gemini 1.5 Pro [37]	gemini-1.5-pro-001
Claude 3 Haiku [2]	*claude-3-haiku@20240307*
Claude 3 Opus [2]	*claude-3-opus-20240229*
Open-source LMMs	
IDEFICS-9b-Instruct [17]	*huggingfaceM4/idefics2-8b*
IDEFICS-80b-Instruct [17]	*huggingfaceM4/idefics2-80b*
Emu2 [41]	*BAAI/Emu2*
InternLM-XComposer-7b [42]	*InternLM/InternLM - XComposer - 7b*
OmniLMM-3b [34]	*OpenBMB/MiniCPM - V*
OmniLMM-12b [34]	*OpenBMB/OmniLMM - 12b*
InstructBLIP-FlanT5-xl [8]	*salesforce/instructblip- flan-t5-xl*
InstructBLIP-Vicuna-7b [8]	*salesforce/instructblip- vicuna-7b*
InstructBLIP-Vicuna-13b [8]	*salesforce/instructblip-vicuna-13b*
Qwen-VL-Chat [3]	*Qwen/Qwen-VL-Chat*
LLaVA-1.5-7b [24]	*llava-hf/llava-1.5-7b-hf*
LLaVA-1.5-13b [24]	*llava-hf/llava-1.5-13b-hf*
Underlying Components of *SCIEval (our work)*	
CLIP (trained)	*openai/clip-vit-base-patch32*
mPLUG-Owl3 (fine-tuned)	*mplugOwl3-8B*

C.2. Model Versions

As detailed in Table 6, we specify the exact versions of the compared models and the hugging face checkpoints utilized for all experiments in this study. In particular, we include the versions of the underlying components used in SCIEval. This ensures full reproducibility and transparency regarding the specific model architectures and weights evaluated in our benchmark.

D. Details of Compared Methods

In this paper, we conduct a comparative analysis between SCIEval and 24 competing models. These baselines are categorized into the following distinct groups:

(1) Closed-source LMMs. This category encompasses industry-leading proprietary models, categorized by their distinct architectural strengths:

The GPT-4 Family: GPT-4V [32] leverages the human-preferred output through Reinforcement Learning from Human Feedback (RLHF) to enhance its visual capabilities. GPT-4 Turbo [31] expands the ecosystem with a 128K context window and integrates multimodal support for image generation. GPT-4o [33] represents the latest iteration, which matches Turbo's reasoning performance while offering significantly improved speed, lower API costs, and superior native comprehension of vision and audio.

The Gemini Family: Gemini [1] is designed for versatile multimodal understanding across text, images, video, and audio, scaling from on-device applications to complex reasoning tasks. Gemini 1.5 Pro [37] introduces high compute efficiency and an expansive context window, capable of reasoning over millions of tokens, including multi-hour videos and extensive document sets.

The Claude 3 Series: Claude 3 Opus [2] is optimized for high-level cognitive tasks, exhibiting near-human fluency in code generation and nuanced multilingual dialogue. Claude 3 Haiku [2] serves as the high-speed, cost-effective counterpart, tailored for near-instantaneous information retrieval and data analysis.

(2) Open-source LMMs. This category highlights prominent models within the open-source community, ranging from massive web-scale learners to instruction-tuned specialists:

Large-Scale Multimodal Learners: IDEFICS (9B/80B) [17] is built on the OBELICS dataset, a filtered and interleaved collection of 141 million web pages, enabling robust integration of vision-language. Emu2 (37B) [41] utilizes a unified auto-regressive objective to master multimodal sequences, demonstrating exceptional in-context learning capabilities. InternLM-Composer (104B) [42] undergoes a multi-phase progressive pre-training process on 1.6T tokens followed by alignment of human-preference.

Instruction-Aware & Foundational Architectures: In-

structBLIP [8] advances the BLIP-2 framework by introducing an instruction-aware Query Transformer that extracts visual features specifically tailored to user prompts. Qwen-VL [3] extends the Qwen-LM foundation through a specialized three-stage training pipeline to achieve high-fidelity visual comprehension. OmniLMM (3B/12B) [34] is designed for efficient multimodal processing, delivering high-quality text output from combined image-text inputs.

Optimized Vision-Language Baselines: LLaVA-1.5 (13B) [24] refines the original LLaVA architecture by incorporating CLIP-ViT-L-336px, an MLP projection layer, and academic-task-oriented VQA data to achieve superior baseline performance.

(3) Existing Faithfulness Metrics. Assessing faithfulness involves measuring the alignment between visual and textual content. Traditional approaches utilize fundamental text-image embedding models:

Embedding-Based Metrics: CLIPScore [13] is the seminal metric in this category, calculating the cosine similarity between image and text embeddings. By focusing tightly on direct image-text compatibility, it serves as a complementary approach to reference-based metrics that emphasize text-text similarities. Similarly, BLIP2Score [19] adapts this concept, utilizing the text-image alignment capabilities inherent in the BLIP2 architecture.

However, recent evaluations have exposed significant limitations in these embedding-based methods, particularly their inability to provide reliable scores when faced with complex prompts involving compositionality, such as nuanced combinations of objects, attributes, and relations. This gap has led to the development of alternative metrics:

VQA-Based and Task-Specific Metrics: To overcome these compositional failures, VQAScore [23] utilizes a visual-question-answering (VQA) model to derive an alignment score based on the probability of a "Yes" response to the simple query: "Does this figure show [text]?" Its underlying model, CLIP-FlanT5, has demonstrated superior performance, even outperforming strong baselines relying on proprietary models like GPT-4V. Furthermore, specialized metrics have been developed that are tailored to specific generative directions.

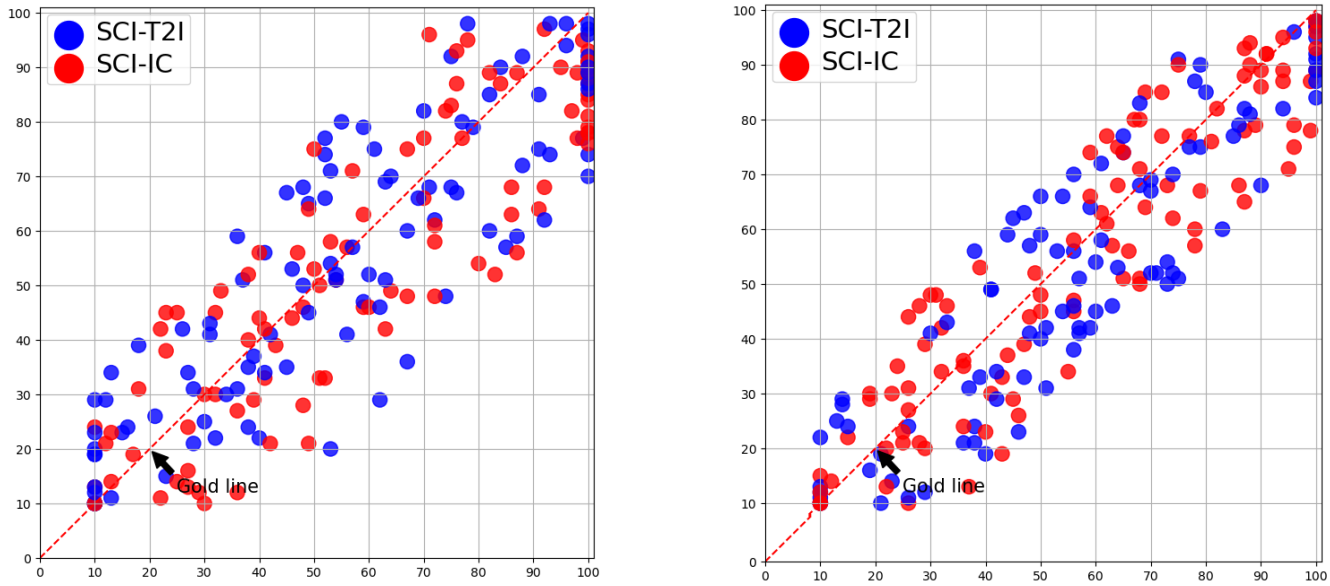
Text-to-Image Faithfulness: TIFA [15] evaluates image-to-text faithfulness by automatically generating question-answer pairs from the text input and assessing the accuracy of VQA models on those questions.

Image-to-Text Faithfulness: In contrast, VALOR-EVAL [35] measures if a generated caption is faithful to visual content. It does this by comparing the generation against gold captions to specifically detect hallucinations regarding object existence, specific attributes (e.g., color, count), and complex relations (positional and comparative).

E. Additional Experimental Results

To complement the overall correlation analysis, we provide a visual representation of the raw scores generated by the models and their direct alignment with human judgments. For this visualization, we randomly sampled 200 instances from the CS subset—evenly split between Sci-T2I and Sci-IC tasks. The relevance scores produced by GPT-4o and SCIEval were plotted against human-annotated scores on a scatter plot (Fig.10), where the X-axis denotes the predicted score and the Y-axis denotes the human reference score. Key observations from this analysis include:

- The Identity line ($y = x$). We include an "Identity" line to represent a perfect correlation where the model's prediction exactly matches the human score. The proximity of a data point to this line serves as a direct measure of scoring accuracy.
- Distribution density. As shown in Fig.10, the data points for SCIEval (right) are significantly more concentrated around the identity line than those for GPT-4o (left).
- Comparative performance. This increase in density demonstrates that SCIEval generates scores that are more closely aligned with human-expert annotations than GPT-4o within this specific benchmark.



GPT-4o **SCIEval (ours)**

Figure 10. To visually assess scoring accuracy, we randomly selected 200 samples from the CS subset and mapped their relevance scores onto a scatter plot. The X-axis represents the predicted scores generated by GPT-4o (left) or SCIEval (right), while the Y-axis reflects the human-annotated scores. To distinguish between the two primary tasks, we have color-coded the data points. This visualization allows for a direct comparison of how closely each model's scoring distribution tracks the human reference across different scientific tasks.