FAITHFUL, ENRICHED, AND PRECISE: BENCHMARKING NATURAL-SCIENCE ILLUSTRATION GENERATION BY T2I MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific illustrations are essential tools for communicating research findings, especially in natural science, where they visualize complex concepts and processes. As Text-to-Image (T2I) models become increasingly capable, researchers have started to use them for scientific illustration generation. However, existing benchmarks often assess outputs at a holistic level, overlooking fine-grained elements, while scientific reasoning ability and output conciseness remain under-quantified. We introduce **FEPBench**, a benchmark built from carefully selected high-quality scientific illustrations across multiple disciplines and layout types. With the assistance of multimodal large language models (MLLMs) and human experts, we provide fine-grained atom set annotations and systematically evaluate T2I models along three dimensions: instruction faithfulness, reasoning enrichment, and semantic precision. Our evaluation further decomposes model performance across visual, textual, relation, and layout elements. Results show that even state-of-the-art (SOTA) closed-source models, such as GPT Image 2 and Nano Banana Pro, still suffer from text-rendering bottlenecks, limited reasoning enrichment, and difficulty balancing generation richness with precision. These findings provide practical guidance for improving and deploying T2I models in scientific illustration generation across diverse scientific research scenarios. Our benchmark dataset, atom set annotations, and code are all available at https://anonymous.4open.science/r/FEDBench-1DDD.

1 INTRODUCTION

Scientific illustrations communicate mechanisms, structures, and processes that are difficult to convey with text alone. Unlike ordinary image generation, scientific illustration is a task of structured visual explanation: useful figures must represent intended content clearly rather than merely appear plausible, consistent with classical views of representation and information-rich graphics (Marr, 1982; Tufte, 1983). Recent T2I systems such as GPT-Image-2 and Nano Banana Pro have improved instruction following, text rendering, and visual reasoning (OpenAI, 2026b; Google DeepMind, 2025), motivating new benchmarks for scientific, academic, and knowledge-intensive image generation (Zhu et al., 2026a;b; Chang et al., 2025; Lin et al., 2026; Luo et al., 2025; Li et al., 2026b; Sun et al., 2025).

Although SciIRBench (Ma et al., 2026) uses checklists to evaluate scientific illustration generation, the effects of element type, scientific domain, illustration complexity, and prompting strategy remain insufficiently characterized. Generated illustrations commonly contain three kinds of content: elements explicitly requested by the prompt; unrequested elements that are scientifically valid, logically consistent, and appropriate to depict; and unsupported additions that should not appear. Distinguishing these outcomes requires more than a holistic quality score. Fine-grained evaluation should quantify faithful coverage, useful scientific enrichment, and unsupported generation across visual, textual, relational, and layout elements.

We therefore introduce **FEPBench** (**F**aithful, **E**nriched, **P**recise), an atomic-set benchmark for natural-science illustration generation. Each illustration is represented by textual, visual, relational, and layout atoms, since an incorrect label, missing relation, or misplaced panel can change scientific meaning despite a plausible global appearance. The supported semantic space contains instruction atoms and reasoning atoms; the latter combine reference-derived content absent from the prompt with open-world additions unanimously accepted by experts. Generated

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

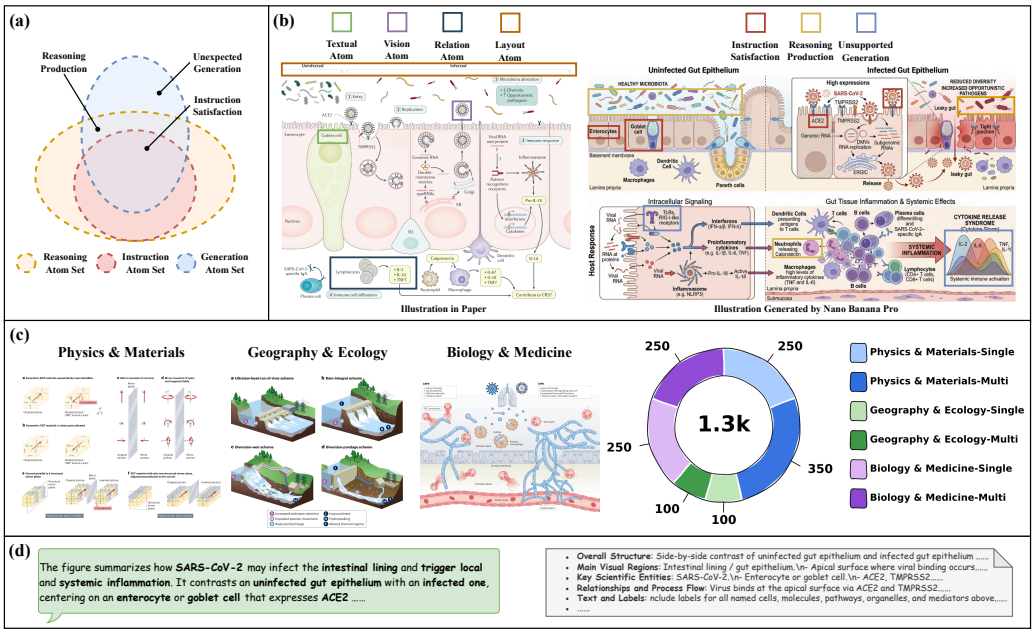

Figure 1: Overview of FEPBench: (a) atom-set representation; (b) four atom types and three outcome states; (c) scientific domains and layout distributions; (d) free-form and structured generation prompts.

atoms are categorized as instruction satisfaction, reasoning production, or unexpected generation, yielding Instruction Faithfulness (IF), Reasoning Enrichment (RE), and Semantic Precision (SP). Three scientific domains, two layout types, and two prompt regimes enable diagnosis of content coverage, enrichment, unsupported additions, and layout failure.

We evaluate nine T2I models and the Coding-Base system ImgCoder under both prompt regimes, comparing overall performance and atom-type scores. We further analyze variation across scientific domains, single- and multi-panel layouts, prompt formats, and atom-set complexity, and inspect representative generations. Finally, three experts assess 1,200 generated figures from all ten systems to validate agreement with two independent MLLM evaluators. These experiments reveal complementary strengths in faithfulness, enrichment, and precision, with persistent bottlenecks in text generation and multi-panel layout control across the evaluated systems.

Overall, our contributions are threefold:

- We introduce **FEPBench**, a benchmark of 1,300 high-quality scientific illustrations from natural-science papers, covering three major disciplines and two layout types. For each instance, we provide two types of generation prompts and human-expert-audited atom-set annotations, spanning visual, textual, relational, and layout elements as shown in Fig.1.
- An evaluation pipeline is proposed which enables fine-grained assessment of generated illustration across different element types along three key dimensions: instruction faithfulness (IF), reasoning enrichment (RE), and semantic precision (SP). The pipeline is robust across different evaluator models and shows strong agreement with human judgments.
- We conduct a comprehensive evaluation of SOTA closed-source and open-source T2I models. Our results show that current models still have substantial room for improvement across all three metrics, with text rendering and scientific reasoning emerging as the main bottlenecks.

2 Related Work

Text-to-image benchmarks and evaluation. T2I evaluation has shifted from global quality and image–text similarity toward compositional and reasoning-aware diagnostics. TIFA (Hu et al., 2023) and DSG (Cho et al., 2024) decompose prompts into question-answerable or dependency-aware semantic units; T2I-CompBench (Huang et al., 2023), VQAScore/GenAI-Bench (Lin et al., 2024), ConceptMix (Wu et al., 2024), Gecko (Wiles et al., 2025), and GeckoNum (Kajić et al.,

Table 1: Comparison with existing text-to-image and scientific illustration benchmarks. Fine-grained evaluation refers to explicit atom-level assessment of visual, textual, relational, and layout elements. ✓: explicit support; △: partial/indirect support; ✗: not explicitly evaluated.

Benchmark	Image Type	Fine-grained (V/T/R/L atoms)	Reasoning	Conciseness	Prompt Effect	Layout Effect
T2I-ReasonBench	General reasoning images	△	✓	✗	✗	✗
MMMG-Bench	Multidisciplinary knowledge images	✓	✓	✗	✗	✗
PaperBananaBench	Academic method diagrams	✗	△	✓	✗	✗
FigureBench / AutoFigure	Scientific illustrations (mainly CS)	✗	△	✗	✗	△
SridBench	Cross-disciplinary scientific illustrations	✗	✗	△	✗	✗
SciGenBench	Scientific images with downstream utility evaluation	✗	△	✗	✗	✗
SciIRBench	Scientific illustration understanding and retrieval	△	△	✗	✗	✗
FEPBench (ours)	**1,300 natural-science journal figures; three domains; single/multi-panel**	✓	✓	✓	✓	✓

2024) probe attributes, relations, compositional control, alignment, and numerical reasoning. T2I-CoReBench (Li et al., 2026a) and T2I-ReasonBench (Sun et al., 2025) further test implicit reasoning, while MMMG (Luo et al., 2025) and BizGenEval (Li et al., 2026b) extend evaluation to knowledge-intensive and structured visual content. These directions are increasingly relevant as frontier generators improve instruction following, text rendering, and visual reasoning. Yet most protocols still target natural images or broad document-style artifacts. They rarely treat natural-science illustration as a distinct scientific communication medium in which labels, entities, inter-entity relations, panel organization, and semantic precision must all be correct at once.

Scientific and academic illustration generation. Scientific-figure generation has emerged as a specialized visual-synthesis task in which the goal is not only to produce plausible images but also to communicate mechanisms, processes, and structures with scientific clarity. DiagramAgent (Wei et al., 2025) studies editable structured diagrams; PaperBanana (Zhu et al., 2026a) combines reference retrieval, planning, rendering, and self-critique for academic illustrations; AutoFigure (Zhu et al., 2026b) introduces FigureBench and an agentic refinement pipeline for publication-oriented figures; SridBench (Chang et al., 2025) evaluates scientific-illustration drawing across disciplines; SciGenBench (Lin et al., 2026) measures downstream reasoning utility and logical validity; and SciIRBench (Ma et al., 2026) investigates scientific illustration understanding through structured evaluation of generated figures. These works establish scientific illustration as an important emerging task, but their evaluation remains primarily focused on holistic multimodal assessment, scientific utility, or predefined checklist-based criteria. FEPBench instead explicitly decomposes generated figures into text, visual, relation, and layout atoms, and distinguishes prompt-required elements, reference-derived enrichment, expert-consensus open-world enrichment, and genuinely unsupported additions through atom-level matching of supported and unsupported semantic content (Tab. 1).

3 FEPBench

Benchmark overview. FEPBench evaluates scientific illustration generation from textual instructions using 1,300 illustrations across physics and materials, geography and ecology, and biology and medicine. It covers single- and multi-panel layouts, with free-form and structured prompts for each reference (Fig. 1). Expert-audited visual, textual, relational, and layout atoms support evaluation of instruction faithfulness, reasoning enrichment, and semantic precision.

Benchmark data curation. We collect 1,300 illustrations through expert screening under strict inclusion criteria. Eligible illustrations must be clear, contain both visual and textual elements, and exclude experimental results, data visualizations, and imaging outputs. To ensure scientific validity and semantic richness, the data are sourced from authoritative Nature Portfolio journals under appropriate authorization for non-commercial research evaluation. Domain experts with PhD-level training write free-form prompts using the original paper context and the corresponding illustration, following how users would naturally instruct a T2I system. GPT-5.4 then rewrites each prompt into a structured format, following Zhu et al. (2026a), under an explicit constraint not to add, remove, or alter visual, textual, relational, or layout content (Appendix Sec. D). We additionally annotate the reference aspect ratio, including landscape, portrait, and square formats, to select a generation resolution closer to the target.

As shown in Fig. 2, an MLLM extracts candidate atoms from each reference illustration and its prompt, assisted by OCR whose recognized in-figure text serves as evidence for textual element detection. Human experts inspect and correct these atoms against the reference and prompt to finalize instruction atoms and reference-derived reasoning atoms. Unanimously accepted expert proposals then expand the reasoning set with scientifically valid open-world atoms. The

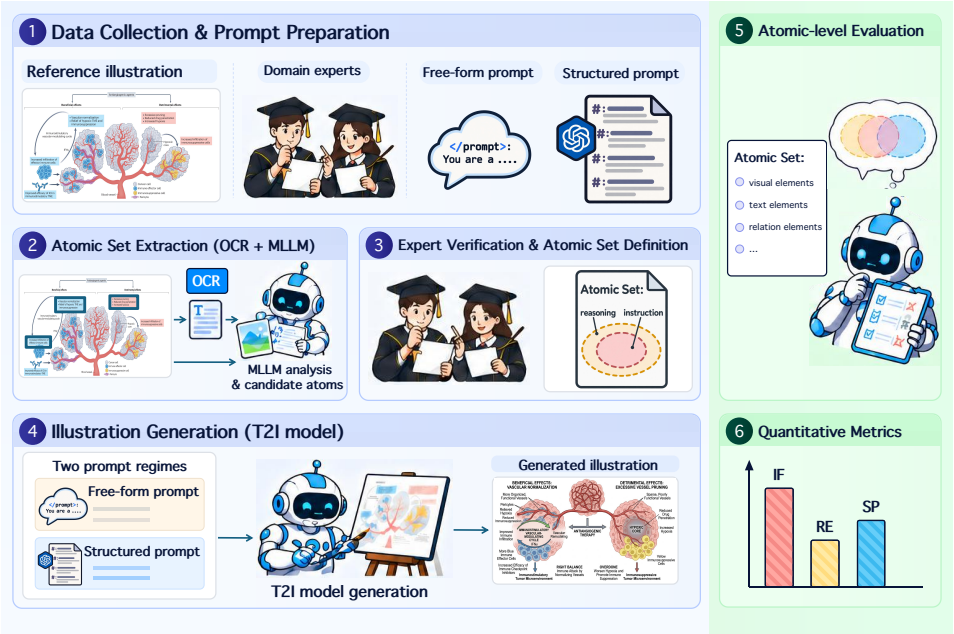

Figure 2: FEPBench construction and evaluation pipeline. Expert screening and prompt preparation are followed by OCR- and MLLM-assisted atom annotation, expert verification, and fine-grained evaluation of generated illustrations against the expert-audited supported atom sets.

extraction prompts and full human verification protocol are detailed in Appendix Secs. A and B; the resulting supported atom sets provide the basis for the evaluation defined below.

Evaluation: atom-set definition. We evaluate generated scientific illustrations with a state-based accounting framework over semantic atoms. Let $\mathcal{A}^{ref}$ denote the atom set directly supported by the reference illustration and let $\mathcal{A}^{ins}$ denote instruction atoms whose semantics are explicitly supported by the generation prompt. The reference-derived reasoning atoms are

$$\mathcal{A}^{rea,ref} = \mathcal{A}^{ref} \setminus \mathcal{A}^{ins}. \tag{1}$$

We further introduce an expert-consensus open-world set $\mathcal{A}^{rea,ow}$. An atom is admitted to $\mathcal{A}^{rea,ow}$ only when it is not required by the prompt and all participating experts agree that it is scientifically valid and reasonable to depict for the target, even if it is absent from the particular reference illustration. Together, reference-derived and expert-consensus open-world atoms define the complete reasoning set

$$\mathcal{A}^{rea} = \mathcal{A}^{rea,ref} \cup \mathcal{A}^{rea,ow}, \tag{2}$$

and the supported semantic space is $\mathcal{A} = \mathcal{A}^{ins} \cup \mathcal{A}^{rea}$. Let $\hat{\mathcal{A}}$ denote the realized atom set of a generated illustration. Each realized atom then belongs to one of three outcome states: *instruction satisfaction* $\left(\mathcal{A}^{ins} \cap \hat{\mathcal{A}}\right)$, *reasoning production* $\left(\mathcal{A}^{rea} \cap \hat{\mathcal{A}}\right)$, and *unexpected generation* $\left(\hat{\mathcal{A}} \setminus \mathcal{A}\right)$. These states induce three complementary evaluation dimensions: *Instruction Faithfulness* (IF), *Reasoning Enrichment* (RE), and *Semantic Precision* (SP).

For generation, each target illustration is paired with two prompt regimes: a free-form expert prompt p^{free} and an LLM-organized structured prompt p^{struct}. Each T2I model G generates two outputs,

$$\hat{I}^{free} = G(p^{free}), \qquad \hat{I}^{struct} = G(p^{struct}). \tag{3}$$

An MLLM verifier extracts the realized atoms $\hat{\mathcal{A}}$ from each generated figure $\hat{I}$ and matches them to the gold atom set for the three evaluation dimensions.

Atom-level matching. In our implementation, semantic atoms are instantiated as textual, visual, relation, and layout atoms, yielding

$$\mathcal{A} = \mathcal{A}_t \cup \mathcal{A}_v \cup \mathcal{A}_r \cup \mathcal{A}_l, \tag{4}$$

where $\mathcal{A}_t$, $\mathcal{A}_v$, $\mathcal{A}_r$, and $\mathcal{A}_l$ denote textual, visual, relation, and layout atoms, respectively. For each text atom $a \in \mathcal{A}_t$, we define a matching function

$$m_t(a) = \mathrm{Exact}(a) \cdot \Big(\beta \cdot \mathrm{Read}(a) + \beta \cdot \mathrm{Attach}(a)\Big) \tag{5}$$

$\mathrm{Exact}(a) \in \{0,1\}$ is indicated as textual exact equivalence, $\mathrm{Read}(a) \in \{0,1\}$ indicates readability, and $\mathrm{Attach}(a) \in \{0,1\}$ indicates correct attachment to intended object or region. $\beta = 0.5$ normalizes the two component scores. Similarly, for each visual atom $a \in \mathcal{A}_v$, we define

$$m_v(a) = \mathrm{Pres}(a) \cdot \Big(\beta \cdot \mathrm{Cnt}(a) + \beta \cdot \mathrm{Loc}(a)\Big) \tag{6}$$

where $\mathrm{Pres}(a) \in \{0,1\}$ corresponds to absent/present status, $\mathrm{Cnt}(a)$ and $\mathrm{Loc}(a)$ denote count and coarse-location correctness, respectively. For each relation atom $a \in \mathcal{A}_r$, we define

$$m_r(a) = \mathrm{Stat}(a) \cdot \Big(\beta \cdot \mathrm{Type}(a) + \beta \cdot \mathrm{Dir}(a)\Big), \tag{7}$$

Here, $\mathrm{Stat}(a) \in \{0,1\}$ represents violated/satisfied status, $\mathrm{Type}(a) \in \{0,1\}$ indicates relation-type correctness, $\mathrm{Dir}(a) \in \{0,1\}$ indicates directional correctness.

For each layout atom $a \in \mathcal{A}_l$, we use a binary local match $m_l(a) \in \{0,1\}$, where $m_l(a) = 1$ if the corresponding layout constraint is satisfied and $m_l(a) = 0$ otherwise. In our implementation, the layout atoms correspond to panel-structure match, panel-count match, and global-layout match.

Instruction Faithfulness and Reasoning Enrichment. For each semantic modality $x \in \{t, v, r\}$ with matching function $m_x(\cdot)$, we define

$$IF_x = \frac{1}{|\mathcal{A}_x^{ins}|} \sum_{a \in \mathcal{A}_x^{ins}} m_x(a),\ IF_l = \frac{1}{|\mathcal{A}_l^{ins}|} \sum_{a \in \mathcal{A}_l^{ins}} m_l(a).\ RE_x = \frac{1}{|\mathcal{A}_x^{rea}|} \sum_{a \in \mathcal{A}_x^{rea}} m_x(a) \tag{8}$$

$$\mathrm{IF} = \frac{1}{4}\left(IF_v + IF_t + IF_r + IF_l\right), \qquad \mathrm{RE} = \frac{1}{3}\left(RE_v + RE_t + RE_r\right). \tag{9}$$

Thus, IF measures how well the generated figure satisfies prompt-required scientific content, whereas RE measures scientific enrichment beyond the prompt over the union $\mathcal{A}^{rea} = \mathcal{A}^{rea,ref} \cup \mathcal{A}^{rea,ow}$. To keep each metric computable, we require a non-zero number of atoms in every evaluated category. The average number of atoms in each category for each illustration type is reported in Appendix Tab.8.

Semantic Precision. We define unsupported scientific content as realized atoms that align to neither instruction atoms nor any reasoning atom, including the expert-consensus open-world set. For modality x, let $\hat{\mathcal{A}}_x^{ex} = \hat{\mathcal{A}}_x \cap (\mathcal{A}_x^{ins} \cup \mathcal{A}_x^{rea})$ and $\hat{\mathcal{A}}_x^{unex} = \hat{\mathcal{A}}_x \setminus (\mathcal{A}_x^{ins} \cup \mathcal{A}_x^{rea})$ denote supported and unsupported realized atoms, respectively. We then define

$$SP_x = 1 - \frac{\left|\hat{\mathcal{A}}_x^{unex}\right|}{\left|\hat{\mathcal{A}}_x^{ex}\right| + \left|\hat{\mathcal{A}}_x^{unex}\right|}, \qquad \mathrm{SP} = \frac{1}{3}\left(SP_v + SP_t + SP_r\right). \tag{10}$$

Higher SP indicates a larger proportion of realized content supported by the prompt, reference-derived reasoning, or expert-consensus enrichment.

4 Experiments

Models and generation protocol. We benchmark nine text-to-image models spanning proprietary frontier systems and recent open-source generators: GPT Image 2 (OpenAI, 2026b), Nano Banana Pro (Google DeepMind, 2026), GPT Image 1.5 (OpenAI, 2025), Qwen-Image-2.0 Pro (Qwen Team, 2026b; Alibaba Cloud, 2026), Seedream 5.0 (ByteDance Seed Team, 2026), FLUX.2 [dev] (Black Forest Labs, 2025), Z-Image-Turbo (Z-Image Team et al., 2025), Ovis-Image (Wang et al., 2025), and HunyuanImage-3.0 (Cao et al., 2025). We additionally evaluate ImgCoder as a Coding-Base comparison system, using Qwen3.8-27B as its underlying model, giving ten systems in the main quantitative comparison. All systems are evaluated on the same benchmark split and prompt regimes. We do not use model-specific prompt engineering, manual image selection, or post-generation editing. For open-source image models, we use the official checkpoints and recommended inference settings; for API-based image models, we use their default generation configurations. When supported, we fix the random seed to 42; a three-seed robustness audit is reported

Table 2: Scores of different models on each atom type and the overall score evaluated by GPT-5.4. Best and second-best results are shown in **bold** and underline, respectively. Closed-source, open-source, and Coding-Base systems are shaded with different background colors.

	Instruction Faithfulness				Reasoning Enrichment			Semantic Precision			Mean		
Model	$IF_v \uparrow$	$IF_t \uparrow$	$IF_r \uparrow$	$IF_\ell \uparrow$	$RE_v \uparrow$	$RE_t \uparrow$	$RE_r \uparrow$	$SP_v \uparrow$	$SP_t \uparrow$	$SP_r \uparrow$	$IF \uparrow$	$RE \uparrow$	$SP \uparrow$
GPT Image 2	**0.809**	0.389	0.573	**0.887**	**0.639**	**0.182**	**0.557**	0.866	0.574	0.954	0.664	**0.459**	0.798
Nano Banana Pro	0.772	0.333	0.541	0.870	0.491	0.135	0.471	0.911	0.611	0.970	0.629	0.366	0.831
GPT Image 1.5	0.793	0.361	0.529	0.710	0.426	0.098	0.401	0.920	0.634	0.972	0.598	0.308	0.842
Seedream 5.0	0.748	0.363	0.504	0.685	0.341	0.097	0.364	**0.930**	0.694	0.970	0.575	0.268	0.865
Qwen-Image-2.0 Pro	0.751	0.234	0.432	0.642	0.409	0.056	0.344	0.927	**0.743**	**0.976**	0.515	0.270	**0.882**
ImgCoder	0.753	**0.454**	**0.611**	0.884	0.249	0.083	0.290	0.913	0.703	0.959	**0.676**	0.207	0.858
FLUX.2 [dev]	0.724	0.289	0.424	0.688	0.309	0.041	0.274	0.892	0.466	0.965	0.531	0.208	0.775
Z-Image-Turbo	0.560	0.179	0.259	0.562	0.195	0.024	0.140	0.810	0.411	0.904	0.390	0.120	0.708
HunyuanImage-3.0	0.634	0.085	0.267	0.668	0.219	0.012	0.143	0.725	0.248	0.914	0.414	0.125	0.629
Ovis-Image	0.478	0.101	0.172	0.584	0.123	0.010	0.077	0.752	0.236	0.918	0.334	0.070	0.635

in Appendix Sec. E. Otherwise, we use the model's default stochastic generation process. Output resolution and aspect ratio are matched as closely as each image model allows to the reference figure. Image-model generation settings are listed in Tab.9.

Evaluation pipeline. We build gold atom sets with an OCR–MLLM pipeline. We first use MinerU-Diffusion-V1-0320-2.5B (Dong et al., 2026), locally deployed on a single NVIDIA RTX 4090 GPU, to extract textual labels and layout cues from each reference illustration. GPT-5.4 (OpenAI, 2026a) is then used as the multimodal annotator to parse the reference figure, OCR results, and prompt into visual entities, text elements, and semantic relations. We use both GPT-5.4 and Qwen3.5-397B-A17B (Qwen Team, 2026a) as MLLM evaluators. Qwen3.5-397B-A17B is locally deployed, whereas all GPT-5.4 calls are made with temperature fixed to 0. Unless otherwise specified, reported evaluation results use GPT-5.4. Additional human-expert annotation details and protocols will be released with the benchmark materials to support reproducibility and independent auditing.

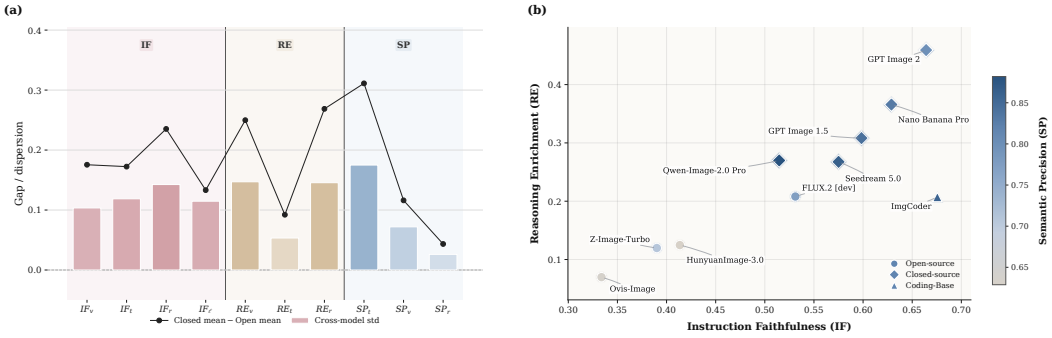

Figure 3: Overall model comparison from fine-grained metric dispersion to the capability frontier. **(a)** Cross-model dispersion and the closed–open performance gap across fine-grained IF, RE, and SP submetrics. Bars show the standard deviation across all ten evaluated systems; the line shows the mean difference between the five closed-source and four open-source T2I models, with Coding-Base ImgCoder excluded from this group-gap calculation. **(b)** Model-level capability frontier in the IF–RE space, with color indicating SP and marker shape denoting open-source, closed-source, or Coding-Base systems. Both panels report GPT-5.4 evaluation using the shared atom-level rubric.

Main Results Tab. 2 reveals a three-way capability frontier rather than a single dominant system. ImgCoder achieves the highest overall Instruction Faithfulness ($IF = 0.678$), driven by the best text faithfulness ($IF_t = 0.458$) and relation faithfulness ($IF_r = 0.613$). GPT Image 2 ranks second in IF (0.664) but is decisively strongest in Reasoning Enrichment ($RE = 0.459$), leading all three RE submetrics ($RE_v = 0.639$, $RE_t = 0.182$, $RE_r = 0.557$). Nano Banana Pro is the next strongest enrichment model ($RE = 0.366$). In contrast, Qwen-Image-2.0 Pro achieves the highest Semantic Precision ($SP = 0.882$), followed by Seedream 5.0 (0.865) and ImgCoder (0.859). These results show that faithful realization, scientific enrichment, and semantic restraint are distinct capabilities: systems that generate richer scientifically plausible content do not necessarily achieve the highest

precision. Text remains the most difficult semantic channel, especially for enrichment, where even the best RE_t is only 0.182. Relation faithfulness also remains substantially below visual faithfulness for most image generators. Additional sample-level analyses are provided in Appendix Sec.C.

Table 3: Fine-grained comparison under free-form and structured prompts across three scientific domains evaluated by GPT-5.4. Overall score is weighted by the number of samples. Within each prompt block, the best and second-best results are marked in **bold** and underline, respectively. Closed-source, open-source, and Coding-Base systems are shaded with different background colors.

	Physics & Materials			Geography & Ecology			Biology & Medicine			Overall		
Model	$IF\uparrow$	$RE\uparrow$	$SP\uparrow$	$IF\uparrow$	$RE\uparrow$	$SP\uparrow$	$IF\uparrow$	$RE\uparrow$	$SP\uparrow$	$IF\uparrow$	$RE\uparrow$	$SP\uparrow$
Free-form prompt												
GPT Image 2	0.656	**0.459**	0.815	0.643	**0.422**	0.819	0.679	**0.461**	0.797	0.663	**0.454**	0.809
Nano Banana Pro	0.636	0.408	0.816	0.607	0.367	0.815	0.651	0.377	0.827	0.637	0.390	0.820
GPT Image 1.5	0.595	0.335	0.827	0.584	0.305	0.844	0.600	0.272	0.854	0.595	0.306	0.840
Seedream 5.0	0.575	0.301	0.852	0.560	0.267	0.872	0.590	0.241	0.870	0.578	0.273	0.862
Qwen-Image-2.0 Pro	0.491	0.274	**0.879**	0.489	0.281	**0.890**	0.527	0.272	**0.877**	0.505	0.274	**0.880**
ImgCoder	**0.684**	0.244	0.856	**0.649**	0.201	0.861	**0.698**	0.177	0.866	**0.684**	0.212	0.860
FLUX.2 [dev]	0.509	0.216	0.784	0.489	0.212	0.779	0.531	0.185	0.780	0.515	0.204	0.782
Z-Image-Turbo	0.366	0.120	0.732	0.376	0.132	0.702	0.367	0.078	0.677	0.368	0.106	0.706
HunyuanImage-3.0	0.356	0.107	0.667	0.335	0.111	0.741	0.369	0.068	0.557	0.358	0.093	0.636
Ovis-Image	0.344	0.088	0.671	0.376	0.110	0.694	0.322	0.054	0.618	0.341	0.078	0.654
Structured prompt												
GPT Image 2	**0.665**	**0.478**	0.787	**0.651**	**0.431**	0.785	0.672	**0.461**	0.788	0.666	**0.464**	0.787
Nano Banana Pro	0.622	0.367	0.834	0.593	0.337	0.837	0.628	0.313	0.853	0.620	0.342	0.842
GPT Image 1.5	0.596	0.338	0.836	0.587	0.305	0.828	0.614	0.279	0.861	0.602	0.311	0.844
Seedream 5.0	0.572	0.286	0.859	0.555	0.260	0.871	0.579	0.235	0.877	0.572	0.262	0.868
Qwen-Image-2.0 Pro	0.513	0.275	**0.881**	0.504	0.266	**0.894**	0.547	0.254	**0.883**	0.525	0.265	**0.884**
ImgCoder	0.663	0.233	0.852	0.605	0.186	0.855	**0.696**	0.173	0.862	**0.667**	0.203	0.856
FLUX.2 [dev]	0.536	0.218	0.764	0.531	0.235	0.767	0.568	0.197	0.771	0.547	0.213	0.767
Z-Image-Turbo	0.400	0.139	0.719	0.419	0.167	0.708	0.423	0.113	0.699	0.412	0.134	0.710
HunyuanImage-3.0	0.476	0.183	0.648	0.429	0.175	0.646	0.477	0.118	0.579	0.469	0.157	0.621
Ovis-Image	0.337	0.065	0.628	0.353	0.099	0.675	0.305	0.041	0.579	0.327	0.061	0.616

As shown in Fig. 3 (a), the largest closed–open gaps occur in textual semantic precision (SP_t, about 0.31), relational reasoning enrichment (RE_r, about 0.27), visual reasoning enrichment (RE_v, about 0.25), and relational instruction faithfulness (IF_r, about 0.24). These are also among the most dispersed fine-grained dimensions, indicating that current systems differ most strongly in whether they can enrich scientific content and keep generated text semantically grounded. Fig. 3 (b) makes the resulting trade-off explicit. GPT Image 2 defines the high-RE frontier, Qwen-Image-2.0 Pro occupies a high-SP region, while Coding-Base ImgCoder reaches the highest IF but only moderate RE. Among open-source image models, FLUX.2 [dev] remains the strongest overall competitor, but it trails the leading closed-source systems especially on RE.

Factor-wise Analysis: Subject, Layout, and Prompting Table 3 shows that the three-way capability ordering is stable across domains. Under free-form prompts, ImgCoder obtains the best IF in all three subjects, while GPT Image 2 consistently leads RE and Qwen-Image-2.0 Pro leads SP. Under structured prompts, GPT Image 2 becomes the best IF model in Physics & Materials and Geography & Ecology, whereas ImgCoder remains strongest in Biology & Medicine and has the highest weighted overall IF (0.670 versus 0.666 for GPT Image 2). GPT Image 2 remains the RE leader in every subject and both prompt regimes, while Qwen-Image-2.0 Pro remains the SP leader. Prompt rewriting is therefore not a universal improvement: it increases overall IF for FLUX.2 [dev], Z-Image-Turbo, and especially HunyuanImage-3.0, but produces only small or mixed changes for stronger closed-source systems and slightly reduces ImgCoder's aggregate scores. The revised results support a coverage–enrichment–precision view rather than a single monotonic model ranking.

Fig.4 (a) shows that multi-panel composition primarily damages layout faithfulness: IF_ℓ decreases for every system, with drops ranging from -0.13 for ImgCoder and -0.17 for GPT Image 2 to approximately -0.49 for Qwen-Image-2.0 Pro and Ovis-Image. Text and relation faithfulness also usually decline. In contrast, some strong systems show small positive changes in visual or textual RE on multi-panel figures, indicating that additional panel structure can create more opportunities for plausible en-

richment even while faithful layout control becomes harder. Semantic precision, especially SP_t, generally decreases on multi-panel inputs. Fig.4 (b) shows that prompt rewriting has a much smaller and more model-dependent effect. FLUX.2 [dev], Z-Image-Turbo, and HunyuanImage-3.0 gain IF from structured prompts, whereas changes for frontier closed-source systems are modest. HunyuanImage-3.0 exhibits the strongest trade-off: rewriting improves several IF and RE terms but sharply reduces SP_t (-0.17). Overall, structural complexity is a more consistent source of failure than prompt format.

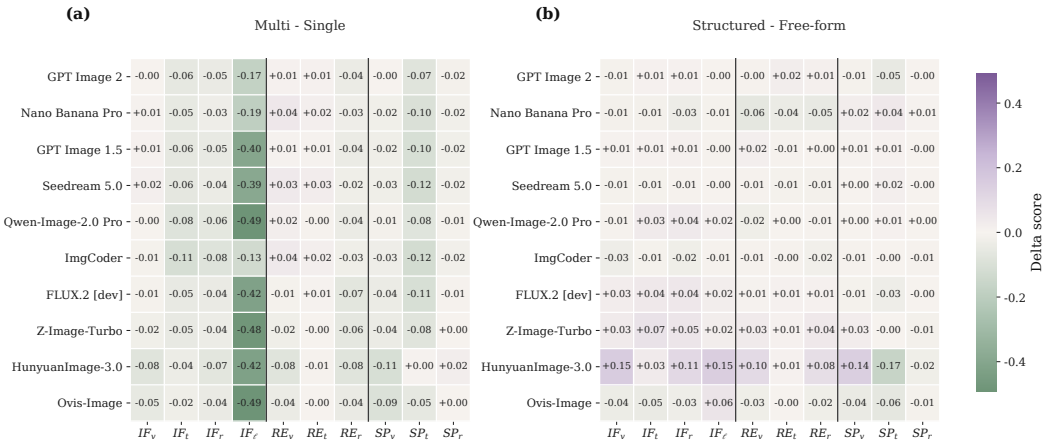

Figure 4: Effects of illustration layout and prompt format on fine-grained model performance evaluated by GPT-5.4. **(a)** Score differences between multi-panel and single-panel illustrations. **(b)** Score differences between structured and free-form prompts. Positive values indicate that the former setting improves performance over the latter. Columns cover fine-grained IF, RE, and SP submetrics, with darker colors indicating larger condition-induced changes across the ten evaluated systems.

Effect of atom set complexity. We further analyze how performance changes as scientific illustrations become compositionally more complex, measured by the total number of supported semantic atoms, including the expanded reasoning set. As shown in Fig. 5, increasing atom count generally reduces both IF and RE, but the rate of degradation differs substantially across systems. GPT Image 2 and Nano Banana Pro retain comparatively high RE as complexity grows, whereas the open-source models and ImgCoder show a steeper decline in enrichment. ImgCoder starts from the highest IF region and remains competitive at high atom counts, while GPT Image 2 and Nano Banana Pro exhibit relatively stable IF curves. SP is comparatively stable for the stronger closed-source systems and ImgCoder, but degrades more noticeably for weaker open-source models. Semantic complexity thus challenges both faithful coverage and scientifically grounded enrichment.

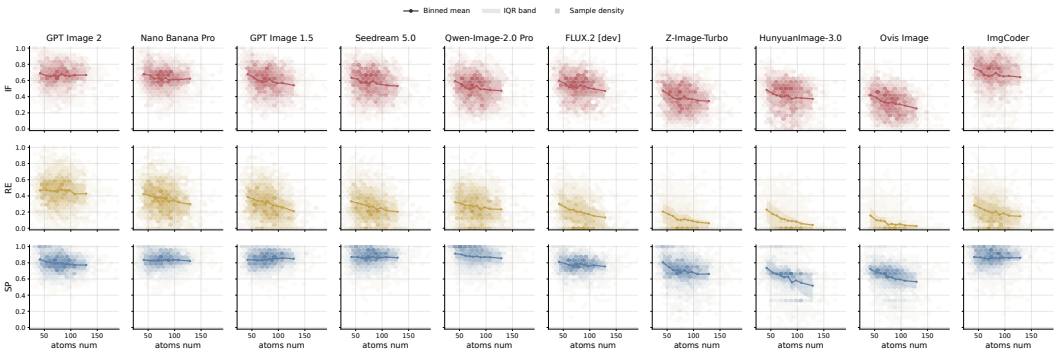

Figure 5: Model robustness to increasing semantic complexity evaluated by GPT-5.4. For each model, we plot IF, RE, and SP against the total number of semantic atoms. Density shows the empirical sample distribution; curves and shaded regions indicate quantile-binned means and IQRs, respectively, where IQR denotes the 25th–75th percentile range of per-sample scores within each bin.

Fig 6 provides a qualitative comparison of the nine T2I generators under the two prompt formats; the Coding-Base ImgCoder comparison is not included in this visualization. The examples mirror the quantitative frontier: GPT Image 2 tends to realize richer scientific content, consistent with its high RE, whereas Qwen-Image-2.0 Pro and Seedream 5.0 produce more conservative outputs that align with their higher SP. Structured prompts can improve organization and make requested regions or labels more explicit, especially for weaker generators, but they do not uniformly improve enrichment or precision. This qualitative behavior reinforces the need to report IF, RE, and SP separately rather than collapsing them into a single score. Additional visualizations of the leading IF, RE, and SP systems are provided in Appendix Sec. G.

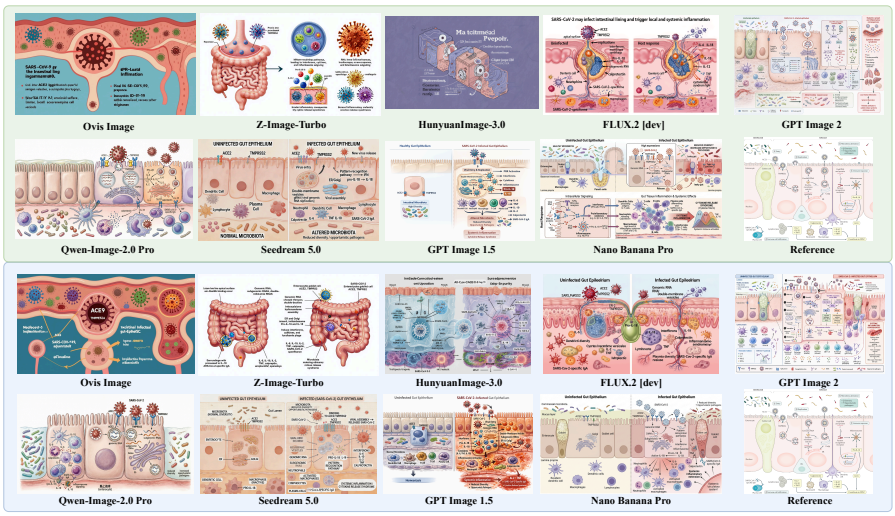

Figure 6: Qualitative comparison of nine T2I generators for the same reference, using free-form prompts (upper group) and structured prompts (lower group). ImgCoder is shown in Appendix Sec.F.

Human Agreement Validation. We uniformly sample 10 references from each discipline $\times$ figure type $\times$ prompt-regime cell and evaluate all 10 systems, including ImgCoder, yielding $3 \times 2 \times 2 \times 10 = 120$ reference cases and $120 \times 10 = 1200$ generated figures. Three domain experts independently score each figure with the same atom-level rubric. Tab.4 compares aggregated human scores with GPT-5.4 and Qwen3.5-397B-A17B on IF, RE, and SP, reporting human-human ICC and Spearman's ρ for human–judge and judge–judge agreement. The consistently high values indicate that automatic evaluation closely tracks expert judgment across independent judge models.

Table 4: Agreement among human experts, GPT-5.4, and Qwen3.5-397B-A17B. We report human-human reliability, human–judge rank correlation, and judge–judge correlation.

Metric	Human-Human ICC↑	GPT-Human $\rho\uparrow$	Qwen-Human $\rho\uparrow$	GPT-Qwen $\rho\uparrow$
IF	0.91	0.89	0.86	0.87
RE	0.88	0.84	0.79	0.83
SP	0.94	0.88	0.83	0.85

5 Conclusion

We present **FEPBench**, an atomic benchmark that separates prompt faithfulness, scientific enrichment, and semantic precision in scientific-figure generation. Its open-world RE definition includes both reference-only content and additional atoms unanimously judged by experts to be scientifically valid, so plausible enrichment is not mislabeled as hallucination. Experiments reveal a three-way frontier: ImgCoder leads IF, GPT Image 2 leads RE, and Qwen-Image-2.0 Pro leads SP; textual enrichment and multi-panel layout remain major bottlenecks. Future systems should therefore balance prompt coverage with grounded enrichment and semantic restraint. Expert-consensus open-world annotation is still conservative and labor-intensive and may miss valid domain-specific alternatives, motivating broader support spaces with reliable expert agreement.

AI USE STATEMENT

Generative AI tools were used in this work both as research subjects and as components of the benchmark construction and evaluation pipeline. Specifically, the text-to-image models evaluated in FEPBench were used to generate scientific illustrations under the benchmark prompts, while multimodal large language models, including GPT-5.4 and Qwen3.5-397B-A17B, were used to assist with fine-grained atom-set annotation and automated evaluation. We additionally used generative AI tools for language polishing, code assistance, and figure preparation during manuscript development.

All AI-assisted benchmark annotations were reviewed under our annotation protocol, and the evaluation procedure was validated through human-expert agreement analysis. AI-assisted code used for data processing, metric computation, statistical analysis, and visualization was inspected, executed, and checked by the authors. Numerical results reported in the paper were obtained from the corresponding evaluation outputs and were independently checked against the underlying data and aggregation rules. References, scientific claims, experimental interpretations, and the final manuscript were reviewed by the authors rather than accepted directly from generative AI outputs.

We did not use generative AI to fabricate experimental observations, human annotations, citations, or benchmark samples beyond the model-generated outputs that constitute the object of evaluation in this study. The authors take full responsibility for all text, claims, code, figures, annotations, and experimental results presented in this work.

ETHICS STATEMENT

FEPBench is designed for the evaluation of scientific illustration generation and does not involve human subjects, personally identifiable information, or sensitive personal data. The benchmark is constructed from published scientific illustrations selected under predefined inclusion criteria and used for non-commercial research evaluation under the applicable access and usage conditions. The released benchmark materials, annotations, and metadata are intended solely for research and evaluation purposes, and users are expected to respect the copyright and licensing conditions associated with the original scientific content.

The benchmark covers multiple natural-science disciplines and illustration layouts, but it does not represent all scientific domains, cultural conventions, or visualization practices. We therefore discourage using FEPBench scores as a universal measure of scientific correctness or as the sole basis for high-stakes scientific decisions. We will document the intended research use, evaluation protocol, and known limitations together with the released benchmark materials.

REPRODUCIBILITY STATEMENT

We provide detailed information required to reproduce FEPBench and its evaluation protocol throughout the main paper and appendix. The benchmark construction procedure, data inclusion criteria, illustration categories, prompt construction, atom-set definitions, and the definitions of Instruction Faithfulness (IF), Reasoning Enrichment (RE), and Semantic Precision (SP) are described in the benchmark and evaluation sections. The appendix further reports dataset statistics, model configurations, generation settings, annotation details, and additional quantitative analyses.

For model evaluation, we specify the evaluated systems, prompt regimes, generation settings, evaluator models, and the OCR–MLLM annotation pipeline. Unless otherwise stated, the main results are evaluated using GPT-5.4, with Qwen3.5-397B-A17B and human experts used for complementary validation. Fine-grained scores are first computed at the corresponding benchmark strata and are then aggregated using the predefined benchmark composition weights, so that the same aggregation protocol is used consistently across reported tables and analyses. Random seeds are fixed where supported by the evaluated models, and API-based systems use the documented generation configurations available at evaluation time.

Benchmark data, atom-set annotations, evaluation code, and supporting materials are available through the anonymous project repository at https://anonymous.4open.science/r/FEDBench-1DDD. These materials include the information needed to reproduce the benchmark evaluation and verify the reported results.

REFERENCES

Alibaba Cloud. Supported Models and Capabilities Overview: Qwen Image Models. https://www.alibabacloud.com/help/en/model-studio/models, 2026. Accessed: 2026-05-06.

Black Forest Labs. FLUX.2: Next Generation Image Generation. https://bfl.ai/models/flux-2, 2025. Accessed: 2026-05-06.

ByteDance Seed Team. Seedream 5.0 Lite. https://seed.bytedance.com/en/seedream5_0_lite, 2026. Accessed: 2026-05-06.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, et al. HunyuanImage 3.0 Technical Report. *arXiv preprint arXiv:2509.23951*, 2025.

Yifan Chang, Yukang Feng, Jianwen Sun, Jiaxin Ai, Chuanhao Li, S. Kevin Zhou, and Kaipeng Zhang. Sridbench: Benchmark of scientific research illustration drawing of image generation model. *arXiv preprint arXiv:2505.22126*, 2025.

Jaemin Cho, Yushi Hu, Jason Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *International Conference on Learning Representations*, 2024.

Hejun Dong, Junbo Niu, Bin Wang, Weijun Zeng, Wentao Zhang, and Conghui He. MinerU-Diffusion: Rethinking Document OCR as Inverse Rendering via Diffusion Decoding. *arXiv preprint arXiv:2603.22458*, 2026.

Google DeepMind. Introducing nano banana pro. Google Blog, 2025. URL https://blog.google/technology/google-deepmind/nano-banana-pro/. Accessed 2026-04-27.

Google DeepMind. Gemini 3 Pro Image – Nano Banana Pro. https://deepmind.google/models/gemini-image/pro/, 2026. Accessed: 2026-05-06.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20406–20417, 2023.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ivana Kajić, Olivia Wiles, Isabela Albuquerque, Matthias Bauer, Su Wang, Jordi Pont-Tuset, and Aida Nematzadeh. Evaluating numerical reasoning in text-to-image models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

Ouxiang Li, Yuan Wang, Xinting Hu, Huijuan Huang, Rui Chen, Jiarong Ou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Fuli Feng. Easier painting than thinking: Can text-to-image models set the stage, but not direct the play? In *International Conference on Learning Representations*, 2026a.

Yan Li, Zezi Zeng, Ziwei Zhou, Xin Gao, Muzhao Tian, Yifan Yang, Mingxi Cheng, Qi Dai, Yuqing Yang, Lili Qiu, Zhendong Wang, Zhengyuan Yang, Xue Yang, Lijuan Wang, Ji Li, and Chong Luo. Bizgeneval: A systematic benchmark for commercial visual content generation. *arXiv preprint arXiv:2603.25732*, 2026b.

Honglin Lin et al. Scientific image synthesis: Benchmarking, methodologies, and downstream utility. *arXiv preprint arXiv:2601.17027*, 2026.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384, 2024.

Yuxuan Luo, Yuhui Yuan, Junwen Chen, Haonan Cai, Ziyi Yue, Yuwei Yang, Fatima Zohra Daha, Ji Li, and Zhouhui Lian. Mmmg: A massive, multidisciplinary, multi-tier generation benchmark for text-to-image reasoning. *arXiv preprint arXiv:2506.10963*, 2025.

Zhiyuan Ma, Zhengfeng Shi, Yuning An, Peize Li, Jiabao Wei, Ruijie Li, Junhao Xiao, Jianjun Li, and Bowen Zhou. Sciir: A large-scale training dataset and benchmark for scientific image reasoning generation. In *European Conference on Computer Vision*, pp. 377–394. Springer, 2026.

David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, San Francisco, 1982.

OpenAI. GPT Image 1.5 Model. https://developers.openai.com/api/docs/models/gpt-image-1.5, 2025. Accessed: 2026-05-06.

OpenAI. GPT-5.4 Model. https://developers.openai.com/api/docs/models/gpt-5.4, 2026a. Accessed: 2026-05-06.

OpenAI. Gpt image 2 model. OpenAI API Documentation, 2026b. URL https://developers.openai.com/api/docs/models/gpt-image-2. Accessed 2026-04-27.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents. https://qwen.ai/blog?id=qwen3.5, 2026a. Accessed: 2026-05-06.

Qwen Team. Qwen-Image-2.0: Professional Infographics, Exquisite Text, and More. https://qwen.ai/blog?id=qwen-image-2.0, 2026b. Accessed: 2026-05-06.

Kaiyue Sun, Rongyao Fang, Chengqi Duan, Xian Liu, and Xihui Liu. T2i-reasonbench: Benchmarking reasoning-informed text-to-image generation. *arXiv preprint arXiv:2508.17472*, 2025.

Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 1983.

Guo-Hua Wang, Liangfu Cao, Tianyu Cui, Minghao Fu, Xiaohao Chen, Pengxin Zhan, Jianshan Zhao, Lan Li, Bowen Fu, Jiaqi Liu, and Qing-Guo Chen. Ovis-Image Technical Report. *arXiv preprint arXiv:2511.22982*, 2025.

Jingxuan Wei, Cheng Tan, Qi Chen, Gaowei Wu, Siyuan Li, Zhangyang Gao, Linzhuang Sun, Bihui Yu, and Ruifeng Guo. From words to structured visuals: A benchmark and framework for text-to-diagram generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13315–13325, 2025.

Olivia Wiles, Isabela Albuquerque, Ivana Kajic, Jordi Pont-Tuset, Matthias Bauer, Su Wang, and Aida Nematzadeh. Revisiting text-to-image evaluation with gecko. In *International Conference on Learning Representations*, 2025.

Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. Conceptmix: A compositional image generation benchmark with controllable difficulty. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Shijie Huang, et al. Z-Image: An Efficient Image Generation Foundation Model with Single-Stream Diffusion Transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Dawei Zhu, Rui Meng, Yale Song, Xiyu Wei, Sujian Li, Tomas Pfister, and Jinsung Yoon. Paperbanana: Automating academic illustration for ai scientists. *arXiv preprint arXiv:2601.23265*, 2026a.

Minjun Zhu, Zhen Lin, Yixuan Weng, Panzhong Lu, Qiujie Xie, Yifan Wei, Sifan Liu, Qiyao Sun, and Yue Zhang. Autofigure: Generating and refining publication-ready scientific illustrations. In *International Conference on Learning Representations*, 2026b.

A Prompt used in MLLM

Here, we provide the prompts used for atom-set scoring and unexpected-atom detection during evaluation in Fig.7. Due to space limitations, the prompts for constructing atom sets will be included in the released code repository. The prompt used for rewriting free-form prompts into structured prompts is in Fig.8

B Human Annotation and Verification Protocol

The following subsections specify the human procedures used to construct, verify, and freeze prompt annotations and semantic atom sets. The protocol is designed to prevent prompt-condition leakage, inconsistent atom granularity, and model-specific expansion of the reasoning space.

Atom Set Verifier Prompt

```
You are a strict verifier for generated scientific figures.
Return valid JSON only. Use double quotes for all object keys and string values.
Do not use markdown fences, comments, Python dict syntax, or explanations.
You are a strict verifier for generated scientific figures.
You are given:
1. a generated scientific figure image
2. compact gold graph atoms from final_annotation_json:
- text entities
- visual entities
- relations
- layout constraints
Your task is to verify the whole gold graph against the generated image in ONE pass.
For each gold text entity, output only:
- entity_id
- presence_status: "present" | "absent"
- exact_match: 1 | 0
- readable: 1 | 0
- attachment_match: 1 | 0
For each gold visual entity, output only:
- entity_id
- presence_status: "present" | "absent"
- count_match: 1 | 0
- coarse_location_match: 1 | 0
For each gold relation, output only:
- relation_id
- relation_status: "satisfied" | "violated"
- type_match: 1 | 0
- direction_match: 1 | 0
For layout, output only:
- panel_structure_match: 1 | 0
- panel_count_match: 1 | 0
- global_layout_match: 1 | 0
Rules:
- Verify only against the generated image and the compact gold graph.
- Do not use or infer anything from the original generation prompt.
- Do not output importance, confidence, notes, or explanations.
- Return exactly one check item for every input text entity, visual entity, and relation.
- For text: exact_match = 1 only if the target text appears exactly as surface_text or normalized_text.
- For text: if exact text is absent, set exact_match = 0, readable = 0, attachment_match = 0.
- For visual: presence_status is "present" only when the gold visual atom is clearly realized.
- For visual: if absent, set count_match = 0 and coarse_location_match = 0.
- For relation: relation_status is "satisfied" only if source entity, target entity, and relation are clearly realized.
- For relation: if either endpoint is absent, the relation should normally be violated.
- For direction_match: use 1 for symmetric/non-directional relations if the relation is otherwise correct.
- Be conservative when uncertain.
- Return valid JSON only; do not wrap it in markdown fences.
Compact gold graph:
<<<GOLD_GRAPH_START>>>
{gold_graph_json}
<<<GOLD_GRAPH_END>>>
Return JSON only:
{{
"text_checks": [
{{
"entity_id": "t1",
"presence_status": "present",
"exact_match": 1,
"readable": 1,
"attachment_match": 1
}}
],
"visual_checks": [
{{
"entity_id": "v1",
"presence_status": "present",
"count_match": 1,
"coarse_location_match": 1
}}
],
"relation_checks": [
{{
"relation_id": "r1",
"relation_status": "satisfied",
"type_match": 1,
"direction_match": 1
}}
],
"layout_check": {{
"panel_structure_match": 1,
"panel_count_match": 1,
"global_layout_match": 1
}}
}}
```

Precision Verifier Prompt

```
You are a strict verifier for generated scientific figures.
Return valid JSON only. Use double quotes for all object keys and string values.
Do not use markdown fences, comments, Python dict syntax, or explanations.
You are a strict verifier for semantic precision of generated scientific figures.
You are given:
1. a generated scientific figure image
2. compact allowed atoms from the gold graph:
- required/optional allowed texts
- allowed visual entities
- allowed relations
- gold visual entity count limits
Inspect the generated image directly and identify unsupported scientific content in ONE pass.
Output:
1. supported_texts: realized scientific texts that align to required or optional gold text atoms
2. unsupported_texts: realized scientific texts that align to neither required nor optional gold text atoms
3. unsupported_visual_entities: salient generated visual entities with scientific meaning that are not allowed by gold atoms
4. unsupported_relations: generated scientific relations that are not allowed by gold relations
5. generated_visual_entity_counts: generated counts for supported visual entity kinds
Rules:
- Do not use or infer anything from the original generation prompt.
- Report only content with clear scientific meaning.
- Ignore harmless decoration, layout fillers, watermark-like noise, and unreadable artifacts.
- Do not output importance, confidence, notes, or explanations.
- Be conservative when uncertain.
- Return valid JSON only; do not wrap it in markdown fences.
Allowed atoms:
<<<ALLOWED_ATOMS_START>>>
{allowed_atoms_json}
<<<ALLOWED_ATOMS_END>>>
Return JSON only:
{{
"supported_texts": [
{{"surface_text": "ACE2"}}
],
"unsupported_texts": [
{{"surface_text": "Random label"}}
],
"unsupported_visual_entities": [
{{
"name": "",
"category": "",
"panel": "single_panel"
}}
],
"unsupported_relations": [
{{
"relation_type": "",
"source_name": "",
"target_name": "",
"panel": "single_panel"
}}
],
"generated_visual_entity_counts": [
{{
"name": "",
"category": "",
"panel": "single_panel",
"generated_count": 0
}}
]
}}
```

Figure 7: Prompts for Atom Set Verifier and Precision Verifier of MLLM.

B.1 Free-form Prompt Authoring and Verification of Structured Rewriting

Prompt authors use the reference figure, its caption, and, when necessary, the relevant main-text context to describe in natural language the generation target, major objects, key scientific relations, and necessary layout constraints. Prompts are written to reflect plausible real-world user requests. They must not rely on context unavailable to the model, and core content must not be intentionally omitted merely to enlarge the scoring space for Reasoning Enrichment (RE).

After a free-form prompt is drafted, a second expert checks its scientific accuracy, executability, ambiguity, and correspondence to the reference figure. Not every detail in the reference figure is required to appear in the prompt; details omitted from the prompt but directly supported by the reference figure may constitute reference-derived reasoning atoms.

Structured rewriting is allowed to change only organization and wording. The reviewer compares, item by item, the objects, quantities, textual labels, relation types, directions, conditions, layout constraints, and any other scoring-relevant requirements, and verifies that no semantic requirement has been added, removed, or strengthened. Any discrepancy is corrected and re-

```
You are an expert prompt rewriter for scientific figure generation.
Your task is to transform a user's natural, conversational description of a scientific illustration into a concise, semi-structured prompt for a text-to-image model.
The rewritten prompt should resemble a well-organized scientific figure description, with short titled sections and compact bullet points when helpful. It should be clear, faithful, and generation-oriented, but not too long.
Requirements:
Preserve all information from the original description.
Do not add any new entities, labels, mechanisms, layout components, colors, styles, or details that are not explicitly mentioned or directly implied.
Do not remove any entities, labels, textual elements, visual elements, relationships, or mechanisms mentioned in the input.
Write in fluent English.
Do not output JSON.
Do not include explanations, apologies, or meta-commentary.
If some information is not provided in the input, omit it rather than inventing it.
Preferred output style:
Start with 1-2 sentences summarizing the figure's overall type, layout, and scientific topic.
Then organize the content into short sections using this style when applicable:
1. Overall Structure
2. Main Visual Regions
3. Key Scientific Entities
4. Relationships and Process Flow
5. Text and Labels
6. Legend and Visual Encoding
7. Style
Only include sections that are supported by the input.
Keep each section compact.
Use short bullet points inside sections if useful.
Emphasize:
core topic and research object
key visual and textual entities
logical relations, mechanisms, and process flow
layout and panel structure
color palette and visual style
Output quality target:
The result should read like a compact, publication-aware scientific illustration prompt: structured, precise, visually grounded, and easy for a text-to-image model to follow.
Input description:
<<<INPUT_START>>>
{input_text}
<<<INPUT_END>>>
```

Figure 8: Prompt used for rewriting free-form prompts into structured prompts

reviewed. The fact that a rewriting model was instructed to "preserve the original semantics" is not, by itself, accepted as evidence of equivalence.

The two prompt forms share the same instruction-atom set only when they have been verified to have identical scoring semantics. If equivalence cannot be guaranteed, the two prompts are annotated separately and the content difference between prompt conditions is disclosed in the paper. We retain the original prompt, rewritten draft, final version, verification result, and revision history.

B.2 Atom Granularity, Categories, and Evidence Provenance

A semantic atom represents one independently verifiable unit of scientific content or one independently verifiable layout constraint. Compound propositions are decomposed into units that can be judged separately, while synonymous repetitions are not counted more than once. Atom granularity is kept consistent across images, models, and prompt conditions.

Table 5: Semantic atom categories and the minimum constraints recorded for verification.

Atom category	Annotated content	Required constraints
Text	Scientifically meaningful labels, terms, formulas, etc.	Canonical form, pre-approved equivalents, associated object or region
Visual	Identifiable scientific entities or structures	Identity, required quantity, coarse location, and allowed depiction forms
Relation	Causal, functional, connective, containment, or other scientific relations	Source/target entities, relation type, direction, and applicable conditions
Layout	Panel structure, panel count, and global organization	Operational criteria for determining satisfaction

For a statement such as "activate A and inhibit B," activation and inhibition are recorded as two separate relation atoms. Multiple synonymous names of the same object are represented by a single semantic entity. Whether an attribute is encoded as an independent atom or as a constraint on an entity follows the fixed annotation manual. Entity types and entity instances are encoded separately, and quantity constraints must not be artificially upweighted by duplicating equivalent entities.

Each atom is assigned one of the following evidence provenances:

- **Instruction atom**: its semantics are directly supported by the generation prompt. Ordinary linguistic paraphrases are accepted as equivalent, whereas content that requires additional scientific reasoning to infer is not treated as an explicit instruction.
- **Reference-derived reasoning atom**: it is directly supported by the reference figure but is not an instruction atom. Captions and main text may be used to disambiguate visible

content, but content appearing only in the main text is not represented as if it were directly visible in the reference figure.

- **Open-world reasoning atom**: it is neither required by the prompt nor directly covered by the reference figure, but is accepted as a scientifically valid supplementary element according to the review procedure in Sec. B.4.

A prompt may contain requirements that are absent from the reference figure; therefore, we do not assume that all instruction atoms are a subset of reference-derived atoms. Set membership is deduplicated by semantics: the same atom cannot receive credit both as a satisfied instruction and as a reasoning enrichment, and overlaps between reference-derived and open-world reasoning atoms are likewise removed.

Each annotation record contains at least the sample ID, atom ID, category, canonical description, set membership, supporting text span or image region, relation endpoints where applicable, accepted equivalent expressions, applicability constraints, proposer, reviewer, version, and reason for revision.

B.3 OCR/MLLM-assisted Annotation and Human Review

OCR and MLLM outputs are used only as candidate drafts. Annotators inspect them item by item and correct omissions, incorrect labels, hallucinated objects, erroneous relations, duplicate atoms, and inappropriate atom splitting, while also actively checking for content that was not proposed by the automated tools.

A reviewer then audits the revised annotation against the original materials. Initial judgments and reasons for disagreement are retained. When third-person arbitration is needed, the decision must cite evidence from the image, prompt, or relevant scientific source. Content that remains unresolved is marked as such and excluded from the frozen ground truth. If the unresolved issue affects the core interpretation of a sample, the sample is returned for revision or excluded according to the predefined rules, with the reason recorded.

Routine review is not treated as independent duplicate annotation. On a pre-sampled, stratified quality-audit subset, two experts independently construct annotations from the same raw materials without seeing each other's results. The fraction of this subset is **[to be filled]**, and the sampling seed is **[to be filled]**. Using a predefined semantic matching rule, we compute atom-extraction precision, recall, and F1, and report category agreement and set-provenance agreement for matched atoms. Quality statistics are computed from the independent annotations before disagreement resolution.

We record the versions of the automated tools, their prompts and parameters, and the types and counts of manual edits. Human review must consist of item-level verification and may not be replaced by direct acceptance of automated outputs.

B.4 Proposal and Unanimous Review of Open-world Atoms

For each candidate open-world atom, the proposer records the atom description, applicability conditions, scientific justification, and why the element is appropriate for the current illustration. Three experts then independently judge its scientific correctness, task relevance, contextual compatibility, appropriateness of expression, and verifiability.

Each dimension is recorded as *accept*, *reject*, or *uncertain*. A candidate is included only when all three experts accept all required conditions. Abstentions, uncertain judgments, and missing votes do not count as acceptance. Discussion may be used to clarify misunderstandings, but any revised candidate must subsequently be independently confirmed again. Initial votes, revisions, final votes, and reasons for exclusion are retained.

A candidate is not included merely because it is scientifically possible; it must also be appropriate for the specified object, species, scale, and conditions. Mutually exclusive mechanisms are not simultaneously treated as required scoring targets. If equivalent-alternative groups are used, their matching rule and denominator calculation must be explicitly implemented in code. When the current implementation does not support such groups, mutually exclusive candidates are not jointly included in the RE denominator.

In principle, candidate construction is completed and frozen before formal evaluation outputs are inspected. If output-driven supplementary review is used in practice, we disclose this procedure separately: candidates are collected uniformly from all systems, model identity is hidden during review, the same review procedure is applied, the updated version is frozen, and all relevant outputs are re-evaluated under the updated support space. The supported atom space must not be expanded selectively for a single model. Results under the original and updated versions are retained separately.

We report the initial unanimous-acceptance rate, candidate-disagreement rate, revision rate, and final inclusion rate. Because the final included set is defined by unanimous acceptance, its "100% agreement" is not presented as an independent measure of annotation reliability.

End-to-end annotation and evaluation workflow. Figure 9 provides a compact implementation-oriented view of how the above protocol connects benchmark construction to model evaluation. Starting from the reference illustration, OCR and an MLLM propose candidate visual, textual, and relational content, which is then checked by human experts before the supported atom sets are frozen. In parallel, expert- and LLM-assisted prompt preparation produces the verified free-form and structured prompts used to query each T2I model. The generated illustration is finally parsed by the evaluator MLLM and matched against the frozen atom sets to obtain IF, RE, and SP, so that prompt preparation, atom construction, generation, and scoring remain explicitly separated in the pipeline.

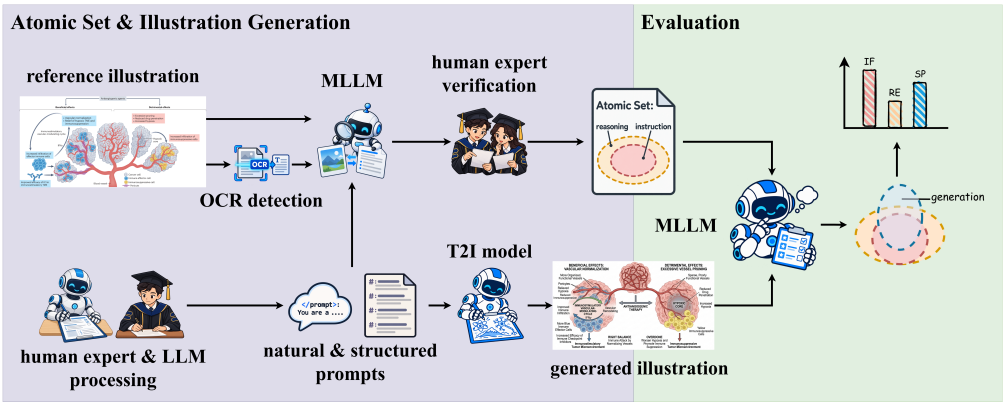

Figure 9: Supplementary overview of the FEPBench annotation and evaluation pipeline. Reference illustrations are processed by OCR and an MLLM, followed by human-expert verification to construct the supported atom sets. Verified free-form and structured prompts are used for T2I generation, and an evaluator MLLM compares the generated illustration with the frozen atom sets to compute Instruction Faithfulness (IF), Reasoning Enrichment (RE), and Semantic Precision (SP).

C Statistic Analysis of Results

Appendix Fig 10 provides a sample-level view of model performance across the three primary evaluation dimensions: Instruction Faithfulness (IF), Reasoning Enrichment (RE), and Semantic Precision (SP). Unlike aggregate tables that report only mean scores, this visualization characterizes the full distribution of scores for each model, revealing both central tendency and variability across individual scientific illustrations. For each metric, the violin shape reflects the empirical score density, the box indicates the interquartile range, the black tick marks the median, and the dot denotes the mean.

The revised distributions reinforce the aggregate findings: ImgCoder is shifted toward higher IF, GPT Image 2 has the strongest RE distribution, and Qwen-Image-2.0 Pro and Seedream 5.0 concentrate near the high-SP end. The substantial overlap among systems also shows why sample-level variability should be reported alongside means.

D Reliability of LLM-based Prompt Rewriting

To verify that structured prompt rewriting changes organization rather than scoring semantics, we conduct a stratified expert audit of 120 free-form/structured prompt pairs, with 20 samples drawn

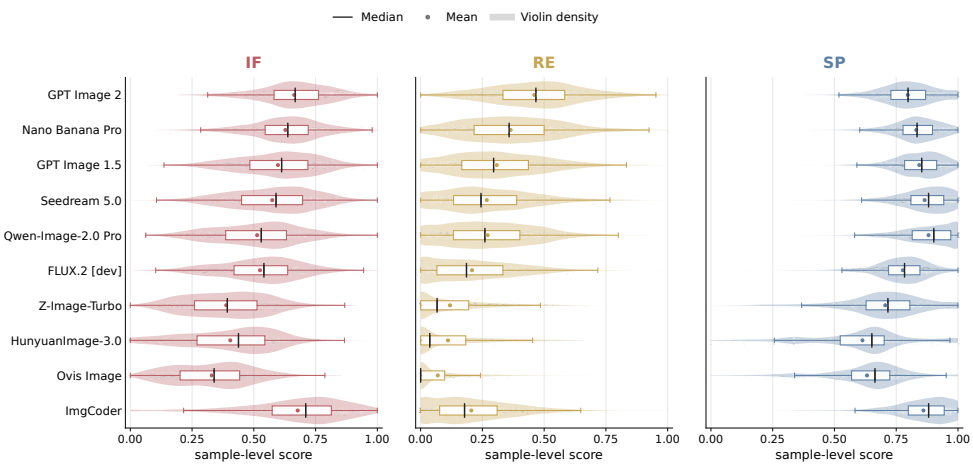

Figure 10: Sample-level score distributions across models evaluated by GPT-5.4. For each model, we show the distributions of Instruction Faithfulness (IF), Reasoning Enrichment (RE), and Semantic Precision (SP) over individual generated scientific illustrations. Violin shapes represent empirical score densities, boxes indicate interquartile ranges, black ticks denote medians, and dots denote means.

from each of the six discipline–layout combinations. Experts compare each pair under a strict semantic-preservation criterion and check whether the entities and scientific relations specified in the free-form prompt remain covered after rewriting. Table 6 summarizes the audit results.

Table 6: Reliability audit of LLM-based prompt rewriting on 120 stratified free-form/structured prompt pairs. Strict coverage measures the fraction of source-prompt semantic requirements preserved in the structured rewrite under expert verification.

Audit criterion	Strict coverage
Entity coverage	0.9923
Relation coverage	0.9639

The audit shows that LLM-based rewriting preserves nearly all entity-level requirements and the large majority of relation-level requirements. Relation coverage is slightly lower than entity coverage, indicating that semantic relations are more sensitive to reformulation than object mentions. For the benchmark, structured prompts are therefore not accepted solely on the basis of the rewriting instruction: they remain subject to the expert equivalence check described in Sec. B.1, and any detected semantic deviation is corrected before the prompt pair is finalized.

E Random-seed Robustness

To assess whether the reported scores are sensitive to generation randomness, we conduct a repeated-generation experiment on a stratified subset of the benchmark. We sample 20 examples from each of the six discipline–layout combinations, yielding 120 examples in total, and regenerate outputs with random seeds 42, 123, and 2026. Table 7 reports the resulting IF, RE, and SP scores for FLUX.2 [dev], HunyuanImage-3.0, Z-Image-Turbo, and Ovis-Image.

Table 7: Random-seed robustness on 120 stratified examples. Each model is regenerated with seeds 42, 123, and 2026 and evaluated using the same IF, RE, and SP protocol.

	Seed 42			Seed 123			Seed 2026		
Model	IF	RE	SP	IF	RE	SP	IF	RE	SP
FLUX.2 [dev]	0.534	0.210	0.759	0.528	0.216	0.762	0.533	0.205	0.751
HunyuanImage-3.0	0.415	0.138	0.628	0.410	0.130	0.624	0.402	0.129	0.619
Z-Image-Turbo	0.382	0.112	0.695	0.394	0.121	0.701	0.393	0.118	0.698
Ovis-Image	0.331	0.071	0.625	0.337	0.077	0.629	0.339	0.062	0.611

The scores are stable across the three seeds on this audited subset. Across all four models and three metrics, the largest seed-to-seed range is 0.018 (Ovis-Image SP), while most variations are close to one percentage point or less. FLUX.2 [dev], for example, varies by only 0.006 in IF, 0.011 in RE, and 0.011 in SP. These results suggest that the observed performance patterns for the tested open-source models are not driven by a particular random seed, while we still report seed-controlled generation settings for reproducibility.

F Generation Results of Model and Analysis

The examples in Figs. 11–12 provide a complementary view of how leading closed-source generators behave beyond aggregate scores. Across domains, models generally recover the global scientific theme and the dominant visual organization, but differ more noticeably in fine-grained labels, relational structure, and the placement of local annotations. These qualitative differences are consistent with the metric decomposition in the main paper: visually plausible outputs can still lose instruction-specified text or relations, while richer generations may introduce additional content that must be evaluated separately through RE and SP.

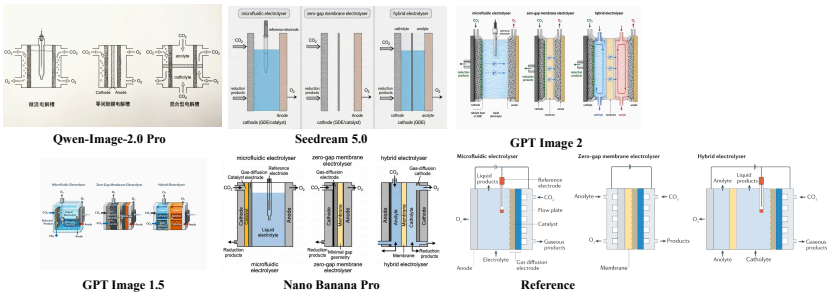

Figure 11: Closed-source model generations on Physics and Materials tasks.

In the Physics and Materials examples, the main structural motif is usually recognizable across models, whereas the density and fidelity of annotations vary substantially. Multi-region scientific layouts especially expose differences in text rendering and in preserving the requested correspondence between labels, arrows, and visual components. This supports reporting element-level IF rather than judging the figures only by overall visual plausibility.

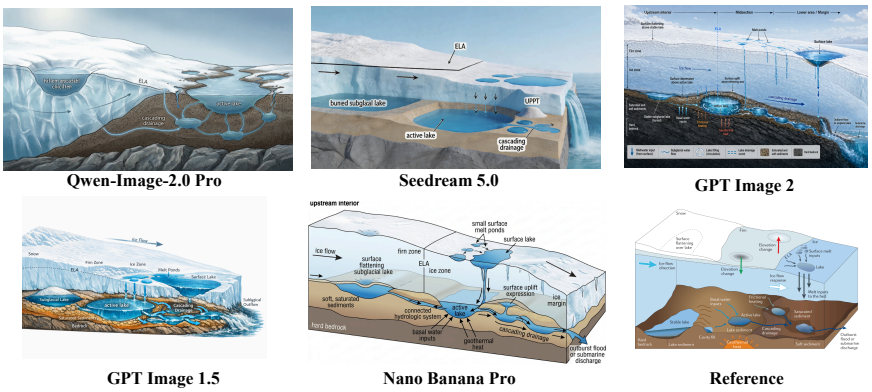

Figure 12: Closed-source model generations on Geography and Ecology tasks.

The Geography and Ecology examples show a similar pattern. Models can reproduce the broad spatial organization and salient scientific objects, but smaller annotations and local semantic relations are less stable. Structured prompts often make the intended regions and components more explicit, yet the resulting gains are model dependent, matching the factor-wise analysis in the main paper.

G Qualitative Comparisons of Leading Models

To complement the quantitative results, we provide qualitative comparisons for the models achieving the highest overall scores on the three principal dimensions of FEPBench: Instruction Faithfulness (IF), Reasoning Enrichment (RE), and Semantic Precision (SP). Specifically, ImgCoder achieves the highest overall IF, GPT Image 2 achieves the highest RE, and Qwen-Image-2.0 Pro achieves the highest SP. For each model, we visualize representative examples spanning Physics & Materials, Geography & Ecology, and Biology & Medicine, with both multi-panel and single-panel layouts. Reference illustrations are shown together with generations under free-form and structured prompts to enable direct comparison across benchmark strata and prompt formulations.

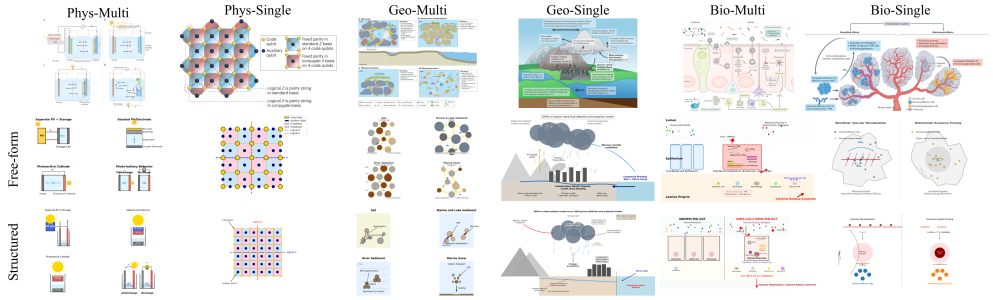

Figure 13: **Qualitative results of ImgCoder, the model with the highest Instruction Faithfulness (IF).** Representative examples cover six benchmark strata across three scientific domains and both multi-panel and single-panel layouts. Reference illustrations are compared with generations under free-form and structured prompts. ImgCoder achieves the highest overall IF (0.676), consistent with its strong preservation of prompt-specified visual, textual, and relational content.

Qualitatively, ImgCoder tends to preserve the requested organization and the correspondence among major visual components more consistently across the six strata. Its advantage is most evident when the prompt specifies dense labels or multiple relations that must remain aligned with a structured layout. At the same time, some outputs simplify local visual details, illustrating why high IF should be interpreted as faithful realization of requested content rather than as a complete measure of visual richness.

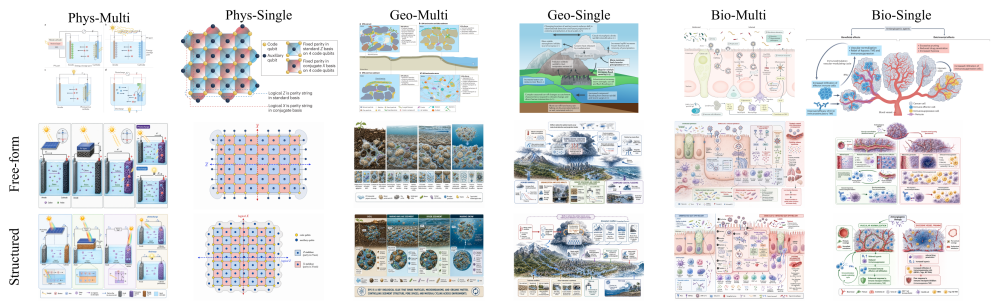

Figure 14: **Qualitative results of GPT Image 2, the model with the highest Reasoning Enrichment (RE).** The examples cover the same six combinations of scientific domain and layout type under both free-form and structured prompting. GPT Image 2 achieves the highest overall RE (0.459), reflecting its stronger tendency to introduce scientifically plausible visual entities, textual information, and semantic relations beyond those explicitly specified in the prompt.

GPT Image 2 more frequently enriches the requested illustration with additional scientifically plausible structure. The extra content can make a figure appear more explanatory or publication-like, but it also increases the importance of separating supported enrichment from unsupported additions. This behavior directly motivates FEPBench's open-world RE formulation together with a separate SP score instead of treating every prompt-external element as an error.

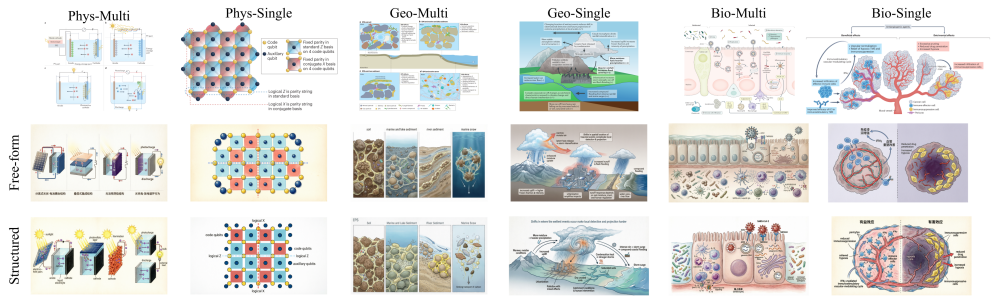

Figure 15: **Qualitative results of Qwen-Image-2.0 Pro, the model with the highest Semantic Precision (SP).** Representative generations are shown for all six benchmark strata under free-form and structured prompts. Qwen-Image-2.0 Pro achieves the highest overall SP (0.882), indicating that most generated visual entities, textual elements, and semantic relations remain supported by the prompt or scientifically plausible enrichment atoms.

Qwen-Image-2.0 Pro is comparatively restrained: its outputs tend to contain fewer unsupported semantic additions, which is reflected by its leading SP. The three sets of examples therefore make the capability frontier concrete. ImgCoder emphasizes faithful coverage, GPT Image 2 emphasizes plausible enrichment, and Qwen-Image-2.0 Pro emphasizes semantic restraint; none of these properties alone subsumes the other two.

H Details of Models and Data

To facilitate reproducibility, we provide comprehensive details on the dataset construction process, model configurations, generation protocols, and evaluation settings, enabling researchers to replicate our results and extend the benchmark in future studies.

Table 8: Mean numbers of instruction and reasoning atoms across benchmark subsets. Here, $|\mathcal{A}_x^{ins}|$ and $|\mathcal{A}_x^{rea}|$ denote the number of instruction and reasoning atoms in modality $x \in \{v, t, r\}$, corresponding to visual, text, and relation atoms, respectively.

Discipline	Type	N	Ins.				Rea.			
			$\lvert\mathcal{A}_v^{ins}\rvert$	$\lvert\mathcal{A}_t^{ins}\rvert$	$\lvert\mathcal{A}_r^{ins}\rvert$	$\lvert\mathcal{A}^{ins}\rvert$	$\lvert\mathcal{A}_v^{rea}\rvert$	$\lvert\mathcal{A}_t^{rea}\rvert$	$\lvert\mathcal{A}_r^{rea}\rvert$	$\lvert\mathcal{A}^{rea}\rvert$
Phys. & Mat.	single	250	9.40	11.75	19.57	40.72	2.15	4.72	7.69	14.56
	multi	350	12.81	15.61	24.73	53.15	2.51	9.11	8.84	20.45
Geo. & Eco.	single	100	11.74	16.94	26.28	54.96	2.49	5.77	9.49	17.75
	multi	100	12.41	18.18	26.28	56.87	2.70	8.73	9.31	20.74
Bio. & Med.	single	250	13.16	21.29	32.17	66.62	1.95	4.83	9.30	16.07
	multi	250	14.04	21.93	32.19	68.16	2.76	9.65	9.23	21.63

Table 8 also highlights the structural complexity of the benchmark. Multi-panel examples contain more instruction atoms than their single-panel counterparts in every discipline, and they generally contain more reasoning atoms as well. Biology and Medicine has the largest instruction-atom totals, whereas the reasoning-atom counts remain substantial across all domains. These statistics help explain why multi-panel generation places greater demands on layout control and relation preservation.

Table 9: Generation settings for the nine evaluated text-to-image models. For API-based models, low-level sampling parameters such as steps and guidance scale are not exposed by the interface. For open-source models, we use 50 inference steps and a guidance scale of 5.0. ImgCoder uses Qwen3.8-27B as its underlying model and is reported separately as a Coding-Base comparison; it is therefore not included in this image-model settings table.

Model	Access	Generation Size	Steps	CFG / Guidance	Neg. Prompt
GPT-image-2	API	1024×1024 / 1536×1024 / 1024×1536	Not exposed	Not exposed	None
GPT-image-1.5	API	1024×1024 / 1536×1024 / 1024×1536	Not exposed	Not exposed	None
Nano Banana Pro	API	1920×1920 / 1920×1080 / 1080×1920	Not exposed	Not exposed	None
Seedream 5.0	API	1920×1920 / 1920×1080 / 1080×1920	Not exposed	Not exposed	None
Qwen-Image-2.0 Pro	API	1920×1920 / 1920×1080 / 1080×1920	Not exposed	Not exposed	None
FLUX.2 [dev]	Open-source	1920×1920 / 1920×1080 / 1080×1920	50	5.0	None
HunyuanImage-3.0	Open-source	1024×1024 / 1536×1024 / 1024×1536	50	5.0	None
Z-Image-Turbo	Open-source	1920×1920 / 1920×1080 / 1080×1920	50	5.0	None
Ovis-Image	Open-source	1920×1920 / 1920×1080 / 1080×1920	50	5.0	None

Generation control and comparability. The generation settings in Table 9 reflect the controls exposed by each model interface. For proprietary API systems, low-level sampler choices are unavailable, so we use the provider defaults while matching output resolution and aspect ratio as closely as possible. For open-source image generators, we standardize inference to 50 steps with guidance 5.0. This protocol is intended to compare the models in their practically accessible configurations rather than to optimize a separate sampler for each system. Consequently, differences in IF, RE, and SP should be interpreted primarily as differences in end-to-end generation behavior under a common benchmark protocol.

Why the benchmark remains challenging. The atom statistics also quantify the density of information that a model must preserve. Moving from single-panel to multi-panel figures increases the total number of instruction atoms from 40.72 to 53.15 in Physics and Materials, from 54.96 to 56.87 in Geography and Ecology, and from 66.62 to 68.16 in Biology and Medicine. The corresponding reasoning-atom totals rise from 14.56 to 20.45, 17.75 to 20.74, and 16.07 to 21.63, respectively. Multi-panel examples therefore require not only more content coverage but also more opportunities for scientifically plausible enrichment and for relation or layout errors. This provides a data-level explanation for the consistent degradation in layout faithfulness observed in the factor-wise analysis.

Reproducibility scope. Together, Tables 8 and 9 specify the benchmark composition and generation-side controls. Exact pixel-level replication is not expected for stochastic or API-based systems, but the released prompts, atom annotations, evaluation code, benchmark weights, and documented settings support reproduction of the protocol and independent auditing of the aggregate results.