Scientific Image Synthesis: Benchmarking, Methodologies, and Downstream Utility

Honglin Lin[1,2*], Chonghan Qin[2,3*], Zheng Liu[4], Qizhi Pei[2], Yu Li[2], Zhanping Zhong[2], Xin Gao[2], Wei Li[2,5], Wentao Zhang[4], Yanfeng Wang[1], Conghui He[2], and Lijun Wu[2†]

[1] Shanghai Jiao Tong University, Shanghai, China
[2] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[3] The University of Hong Kong, Hong Kong SAR, China
[4] Peking University, Beijing, China
[5] East China Normal University, Shanghai, China

Abstract. While synthetic data has proven effective for improving scientific reasoning in the text domain, multimodal reasoning remains constrained by the difficulty of synthesizing scientifically rigorous images. Existing Text-to-Image (T2I) models often produce outputs that are visually plausible yet scientifically incorrect, resulting in a persistent visual–logic divergence that limits their value for downstream reasoning. Motivated by recent advances in next-generation T2I models, we conduct a systematic study of scientific image synthesis across generation paradigms, evaluation, and downstream use. We examine both direct pixel-based generation and programmatic synthesis, and instantiate ImgCoder as a structured implementation of the code-driven workflow, following an explicit "understand → plan → code" prompting strategy to encourage clearer structural specification. To rigorously assess scientific correctness, we introduce SciGenBench, which evaluates generated images based on information utility and logical validity. Our evaluation reveals systematic failure modes in pixel-based models and highlights a fundamental expressiveness–precision trade-off. Finally, we show that fine-tuning Large Multimodal Models (LMMs) on rigorously verified synthetic scientific images yields consistent reasoning gains, with potential scaling trends analogous to the text domain, validating high-fidelity scientific synthesis as a viable path to unlocking massive multimodal reasoning capabilities.

Keywords: Text-to-Image Generation · Data Synthesis · Multimodal Reasoning

1 Introduction

With the advancement of Large Multimodal Models (LMMs) [1,14,34], enabling robust reasoning in scientific domains such as mathematics, physics, and engineering has become an increasingly important goal [23,24,26,43,46]. In the text-only setting, large-scale synthetic data has proven effective in alleviating data

bottlenecks and improving scientific reasoning [17, 21, 22, 27, 28, 36, 38]. However, multimodal scientific reasoning remains significantly underdeveloped. Beyond data scarcity, a fundamental challenge lies in the difficulty of synthesizing scientifically rigorous images. Unlike natural images, scientific visuals must satisfy strict structural and relational constraints, which existing Text-to-Image (T2I) models often fail to enforce, producing visually plausible yet scientifically incorrect results.

The recent emergence of next-generation T2I models [35, 40, 45], such as `Nanobanana-Pro` [16], has renewed interest in this challenge. These models promise improved semantic understanding and visual control, raising a natural question: *can modern T2I systems overcome the long-standing limitations of scientific image synthesis and unlock scalable multimodal data generation for reasoning?* To answer this question, we propose a systematic investigation with methodological innovation, benchmarking, and downstream reasoning.

From a methodological perspective, scientific image synthesis can be broadly categorized into two paradigms. Pixel-based T2I models generate images end-to-end [2, 4, 7, 44, 45] and offer strong visual expressiveness, but often struggle to reliably satisfy strict structural constraints. In contrast, programmatic approaches [30, 33, 48] generate images through explicit, executable codes, providing stronger structural control. Building on this paradigm, we instantiate **ImgCoder** as a structured implementation of the code-driven workflow, adopting an "Understand→Plan→Code" prompting strategy to separate reasoning from rendering. Through controlled experiments, we observe a consistent *precision–expressiveness* trade-off between the two paradigms.

To rigorously evaluate scientific image generation, we observe that existing benchmarks and metrics—largely focused on pixel similarity or coarse semantic alignment—are insufficient for assessing scientific correctness, where correctness hinges on precise structural and relational validity. We therefore introduce **SciGenBench**, a specialized benchmark comprising 1.4K problems that are designed to evaluate the information utility and logical validity of generated scientific images across 5 domains and 25 image types. SciGenBench adopts a hybrid evaluation protocol that combines fine-grained LLM-as-Judge scoring with an automated inverse validation pipeline, shifting the focus from visual aesthetics to information utility and logical rigor. Our evaluation uncovers a pervasive visual–logic divergence in current pixel-based models, while code-driven approaches achieve higher structural precision at the cost of reduced expressiveness. We further categorize these failure modes, providing a systematic understanding of where and why different paradigms break down.

Finally, we examine the downstream utility of synthetic scientific images for multimodal reasoning. By grounding image synthesis in scientifically verified textual sources, the resulting visual–text pairs provide reliable supervision for training. Experiments demonstrate that fine-tuning LMMs on such rigorously verified synthetic data yields consistent improvements in scientific reasoning, with evidence of potential scaling trends similar to those observed in text-only

settings. These findings suggest that high-fidelity scientific image synthesis offers a viable and scalable pathway toward advancing multimodal scientific reasoning.

2 Related Work

2.1 Scientific Diagram Generation

Text-to-Image Models. Following traditional UNet-based Latent Diffusion Models (LDMs) [11, 19, 39], Diffusion Transformers (DiT) [2, 11, 37] and Autoregressive Transformers [7, 9, 44] have recently emerged to dominate text-to-image generation. While proprietary systems [16, 35] and open-source models [4, 45] achieve high visual fidelity, generating structured diagrams directly remains challenging [52,54], revealing a gap between visual realism and structural correctness.

Code-based Methods. Early approaches generate visualization code from captions [3], while later works scale instruction-based plotting via curated code corpora [33]. Some methods rely on manually defined templates [48, 49], at the cost of limited output diversity.

2.2 Benchmarking Text-to-Image Generation

From Fidelity to Reasoning. Early benchmarks [8, 10, 13, 18, 20, 50] focus on visual fidelity and semantic alignment via auto metrics (e.g., FID, CLIP), while recent works shift toward *Reasoning-Informed* evaluation [6], assessing physical plausibility or contextual consistency in natural scenes (e.g., RISEBench [53], T2I-ReasonBench [42]). ScImage [51] constructs synthetic text-to-image tasks designed to evaluate spatial, numerical, and attribute understanding in scientific contexts. SridBench [5], on the other hand, benchmarks scientific image generation using figures collected from academic papers, which primarily serve explanatory or illustrative purposes.

The Gap: Reasoning-Oriented Synthesis. In contrast, scientific image synthesis requires explicitly materializing abstract axioms into precise visual structures (e.g., circuit topologies). Unlike natural scenes with implied logic, these tasks demand active computation under strict domain laws.

3 Scientific Image Generation

We present the first formalization of text-to-image generation task in the context of reasoning data synthesis, and investigate scientific image synthesis through a comparative study of two paradigms. As shown in Fig. 1 (Left), pixel-based models synthesize images end-to-end from text, while code-driven approaches generate executable specifications that are deterministically rendered.

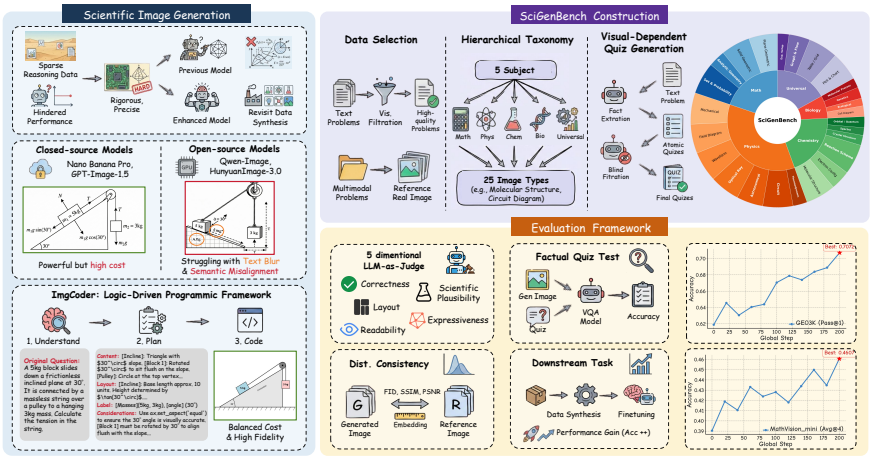

Fig. 1: Methodological Overview. The framework consists of three core components: (1) **Scientific Image Generation** (Left), where we adopt ImgCoder, a code-driven pipeline decoupling planning from implementation to outperform pixel-based baselines; (2) **SciGenBench Construction** (Top Right), a rigorously curated benchmark with a fine-grained taxonomy and atomic quizzes; (3) **Evaluation Framework** (Bottom Right), a multi-faceted assessment system combining LMM judges, inverse validation, standard metrics, and downstream performance.

3.1 Task Definition

Let $\mathcal{T}$ denote the space of scientific descriptions and $\mathcal{I}$ the image space. We formulate scientific image generation as a *conditional generation* problem with a parameterized model $\mathcal{M}_\theta : \mathcal{T} \to \mathcal{I}$. Unlike open-domain synthesis, valid images are constrained by a set of latent scientific axioms $\mathcal{A}$. Given a query $T \in \mathcal{T}$, the goal is to generate an image I^* that maximizes the likelihood of the reasoning outcome y under a downstream solver $\mathcal{S}$:

$$I^* = \underset{I \sim \mathcal{M}_\theta(\cdot|T)}{\arg\max} \; P_{\mathcal{S}}(y \mid I, T, \mathcal{A}), \tag{1}$$

where y is determined by $\mathcal{A}$ and $P_{\mathcal{S}}$ denotes the probability assigned by the solver to y. This objective encourages images that are not only visually coherent but also structurally faithful for downstream reasoning.

3.2 Pixel-based Synthesis Framework

This framework represents the direct synthesis paradigm, treating scientific image generation as an end-to-end translation task from textual descriptions to visual pixel space. We leverage SOTA T2I models as generators, including proprietary systems (e.g., `Nanobanana-Pro`, `GPT-Image-1.5`, `Seedream-4.0`, `Flux2`) known for their robust instruction following, and leading open-source models

(e.g., `Qwen-Image`, `HunyuanImage-3.0`) representing accessible multimodal capabilities. To bridge the gap between general T2I capabilities and STEM requirements, we employ a constraint-injection prompting strategy for image generation. This pipeline regulates generation by enforcing *information fidelity* (explicit visualization of all entities), applying strict *negative constraints* (preventing solution leakage like calculation steps), and ensuring *stylistic standardization* (prioritizing a clean, textbook-style aesthetic).

3.3 ImgCoder: Code-based Synthesis Framework

While pixel-based models provide flexible end-to-end synthesis, they are inherently constrained by the underlying diffusion paradigm, which makes it difficult to explicitly enforce structured constraints and logical consistency during generation. Code-driven approaches, in contrast, decouple structural specification from visual rendering, allowing deterministic control over geometry and semantics. We therefore explore a programmatic paradigm and refer to the large-model-based code generation pipeline for deterministic image rendering as **ImgCoder**. Rather than proposing a new generation paradigm, our goal is to systematically instantiate and evaluate the code-driven workflow under a unified experimental protocol. Concretely, we implement an *"Understand → Plan → Code"* baseline, where the model generates executable code (e.g., Python) for structured image rendering. To reduce layout inconsistencies commonly observed in direct text-to-code generation, ImgCoder incorporates an explicit reasoning stage prior to synthesis, encouraging the model to outline the visualization intent before emitting code. Concretely, the planning phase explicitly defines four aspects of the target figure: **(1) Image Content**, by exhaustively identifying all geometric entities, physical components, and their logical relationships; **(2) Layout**, by pre-planning coordinate systems and topological arrangements to prevent visual clutter and unintended overlaps; **(3) Labels**, by determining both the semantic content and precise anchor points of textual annotations; and **(4) Drawing Constraints**, by validating the plan against domain-specific axioms (e.g., geometric rules or physical laws) while strictly avoiding answer leakage.To examine the effectiveness of ImgCoder across model backbones, we implement two variants: `Qwen3-ImgCoder` and `Gemini3-ImgCoder`, built upon `Qwen3-235B-Instruct` [47] and `Gemini3`, respectively.

4 SciGenBench: Benchmarking Scientific Image Synthesis

We next turn to evaluation. Standard image-generation metrics are ill-suited for scientific imagery, where minor structural errors can invalidate semantics. We therefore introduce **SciGenBench** to benchmark the paradigms in Sec. 3 by measuring the *information utility* and *logical correctness* of generated images. Fig. 1 (Right) overviews our pipeline, which integrates curated data construction, a hierarchical taxonomy, and a hybrid evaluation framework.

4.1 Data Acquisition & Selection

SciGenBench is designed with two complementary data sources that serve distinct evaluation purposes: a primary instruction-driven benchmark for assessing synthesized scientific images, and a real-world visual reference set for distributional comparison.

Verified Scientific Instructions. The core of SciGenBench is constructed from high-quality scientific text corpora, including `MegaScience` [12] and `WebInstruct-verified` [32], which are selected for their logical correctness and factual rigor. To ensure suitability for visual generation, we apply a visualizability filtration to remove non-visual content (e.g., abstract derivations), retaining only descriptions with concrete, imageable structures.

Real-World Visual Reference. To contextualize synthetic scientific images against authentic visual data, we incorporate a vision-optional subset of `SeePhys` [46] as a human-authored reference set, denoted **SciGenBench-SeePhys**. This subset participates in all evaluation dimensions and serves as the ground truth for reference-based standard image metrics used in real–synthetic comparison.

4.2 Hierarchical Taxonomy

To establish a structured data distribution, we organize SciGenBench using a rigorous two-level *"Subject–Image Type"* taxonomy. We employ `Gemini-3-Flash` as a unified data curator to perform visualizability filtration and fine-grained classification jointly within a single inference pass, enabling scalable and consistent annotation. The resulting taxonomy is defined along two hierarchical dimensions. **(1) 5 Subjects** comprise Mathematics, Physics, Chemistry, Biology categories, and a *Universal* category that captures cross-domain visual structures. **(2) 25 Image Types** include fine-grained categories reflecting domain-specific visual conventions, such as *molecular structure* in Chemistry or *circuit diagram* in Physics, as well as generic types under the Universal subject, including *chart & graph* and *experimental setup*. Detailed definitions of all image types are provided in supplementary material.

4.3 Visual-Dependent Quiz Generation

To ensure that generated images provide high information utility and are strictly indispensable for problem solving, we design a *"Generate–Filter–Select"* pipeline to construct atomic, visually grounded quizzes.

Fact-based Question Formulation. Using `Gemini-3-Flash` with a structured quiz-generation prompt, we first analyze each source instruction to extract verifiable factual elements, such as numerical values, geometric relations, or domain-specific properties (e.g., electron configurations). These elements are then converted into atomic questions, each targeting a single, concrete piece of visual information.

Blind Filtration for Visual Necessity. To eliminate pseudo-multimodal questions that can be solved without visual input, we introduce a *blind filtration*

mechanism. Specifically, `GPT-5-nano` is employed as a blind solver that receives only the question text, without access to image. Each question is evaluated across 4 independent trials; those answered correctly in all trials are discarded, as they indicate text leakage or reliance on commonsense rather than visual evidence.

Density-based Selection and Human Review. To prioritize information-rich samples, we retain images associated with the highest number of valid atomic quizzes within each image type. The resulting quiz set is further reviewed by expert annotators to ensure logical consistency, scientific rigor, and the absence of hallucinations.

Table 1: Evaluation Dimensions for LMM-as-Judge. Gemini-3-Flash scores images (0-2) across five criteria.

Dimension	Evaluation Criteria & Focus
Correctness & Fidelity	Strict prompt adherence; detects compositional errors (quantity/attribute mismatch).
Layout & Precision	Precision of geometric construction, topological accuracy, and coordinate system alignment.
Readability & Occlusion	Clarity of textual labels, checking for occlusion, garbled text, or background interference.
Scientific Plausibility	Conformity to domain-specific axioms (e.g., valency balance, Newton's laws).
Expressiveness & Richness	Ensures holistic context and scenario completeness, avoiding isolated elements.

4.4 Evaluation Framework

Evaluating scientific image generation poses unique challenges, as traditional pixel-level metrics fail to capture logical correctness and scientific factuality. To establish a rigorous standard, we adopt a hybrid evaluation strategy comprising automated judge, inverse validation, traditional metrics and downstream utility assessment, as shown in Fig. 1 (Bottom Right).

Multi-dimensional LMM-as-Judge. Following the structured evaluation rubric described in the supplementary material, we employ `Gemini-3-Flash` as an automated evaluator. For each generated image, the judge produces a reasoning critique along with a score $s \in \{0, 1, 2\}$ for each dimension specified in Tab. 1, enabling fine-grained and holistic assessment of visual logical correctness.

Inverse Quiz Validation. To quantify whether a generated image faithfully encodes its intended information, we introduce an inverse validation metric. Let $\mathcal{Q}_I$ denote the set of atomic quizzes associated with image I, and $\mathcal{V}(I, q) \in \{0, 1\}$ indicate the correctness of a strong VQA model's response to question $q \in \mathcal{Q}_I$. We define the **inverse validation rate** ($\mathcal{R}_{\text{inv}}$) as the proportion of images for which *all* associated quizzes are answered correctly:

$$\mathcal{R}_{\text{inv}} = \frac{1}{|\mathcal{D}|} \sum_{I \in \mathcal{D}} \mathbb{I} \left(\sum_{q \in \mathcal{Q}_I} \mathcal{V}(I, q) = |\mathcal{Q}_I| \right), \tag{2}$$

Table 2: Overall results on SciGenBench. We report the inverse validation rate ($\mathcal{R}_{\text{inv}}$) and the LMM-as-Judge scores. Standard-metrics are computed on the real-image SeePhys subset.

Model	$\mathcal{R}_{\text{inv}}$ (%) ↑	LMM-as-Judge (0–2) ↑					Standard-metrics			
		C&F	L&P	R&O	SP	E&R	PSNR ↑	SSIM ↑	CLIP ↑	FID ↓
Open-source T2I Models										
HunyuanImage-3.0 [4]	30.79	0.39	0.78	1.44	0.56	0.81	12.21	0.82	25.01	93.27
Qwen-Image [45]	38.86	0.24	0.70	1.48	0.30	0.76	9.63	0.78	25.02	120.42
Closed-source T2I Models										
GPT-Image-1 [35]	42.97	0.57	1.37	1.90	0.84	1.19	13.07	0.84	25.14	**77.31**
Seedream-4.0 [40]	52.67	0.44	0.94	1.67	0.55	0.95	10.65	0.74	25.02	98.22
Nanobanana [15]	57.75	0.43	0.92	1.60	0.60	1.15	14.12	0.85	25.13	104.70
Flux2-flex [25]	58.83	0.48	1.06	1.70	0.67	1.20	14.11	0.85	25.10	96.74
GPT-Image-1.5 [35]	63.52	0.98	1.70	1.97	1.17	1.62	**14.79**	**0.88**	25.16	112.52
Nanobanana-Pro [16]	73.41	1.59	1.87	**1.98**	1.72	**1.93**	12.02	0.81	25.01	87.72
Code-based Generation										
AutomaTikZ [3]	8.80	0.09	0.41	0.89	0.22	0.06	15.11	0.84	25.25	107.21
Qwen3-ImgCoder [47]	56.38	1.21	1.30	1.62	1.39	1.29	14.71	0.86	**25.21**	121.55
Gemini-3-Flash-ImgCoder [14]	76.93	1.80	1.88	1.88	1.92	1.91	14.63	0.85	25.18	117.83
Gemini-3-Pro-ImgCoder [14]	**77.87**	**1.82**	**1.93**	1.91	**1.93**	1.90	14.59	0.86	25.06	107.67

where $\mathbb{I}(\cdot)$ is the indicator function and $\mathcal{D}$ denotes the evaluation set. To ensure that validation failures primarily reflect informational deficiencies in the image rather than limitations of the solver, we use `Gemini-3-Flash` as the VQA engine.

Reference-based Standard Metrics. For comparability with prior T2I research, we also report conventional automated metrics (FID, PSNR, SSIM) computed against ground-truth images. These metrics quantify the visual domain gap between synthetic and authentic scientific images, but serve as auxiliary references due to the sparse pixel distribution of scientific diagrams. These metrics are applied only to images from SciGenBench-SeePhys.

Data Utility for Training. As the ultimate measure of quality, we evaluate the performance gains of LMMs fine-tuned on synthetic images produced by different generation paradigms. This directly assesses the functional utility and logical consistency of synthesized data from a downstream reasoning perspective (Sec. 6). To prevent textual shortcuts during training, we apply a *multimodal adaptation* strategy, which masks textual cues in original problem and enforces reliance on visual evidence.

5 Benchmarking Generative Capabilities

In this section, we aim to address pivotal research questions (RQs) regarding scientific image synthesis through comprehensive benchmarking. **RQ1- Generative Capability:** How do current SOTA models perform in scientific image generation task? **RQ2 - Paradigm Comparison:** What are the respective trade-offs between generative models and programmatic frameworks? Detailed experimental setup are provided in supplementary material.

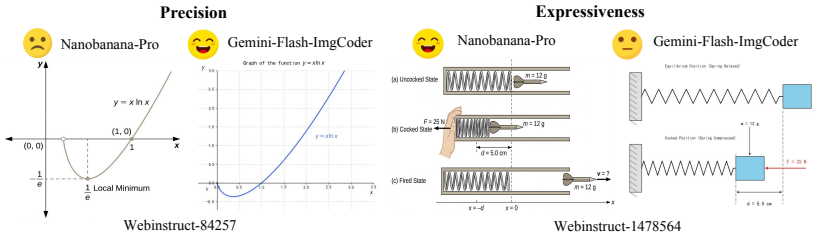

Fig. 2: Precision vs. Expressiveness Trade-off. *Left (a):* When plotting the function $y = x \ln x$, pixel-based models produce visually smooth but mathematically inaccurate plots, while code-based methods ensure exactness via execution. *Right (b):* Conversely, for physical scenarios like a spring system, pixel-based models offer richer visual expressiveness, whereas code-based outputs remain schematic.

5.1 Quantitative Results

Tab. 2 presents the comprehensive benchmarking results across our proposed taxonomy. The quantitative analysis reveals several key insights:

Closed- vs. Open-source Performance Gap. Closed pixel-based models consistently outperform open-source T2I systems across both inverse validation rate and LMM-as-Judge evaluations, benefiting from large-scale proprietary data and optimized training pipelines. For instance, `Nanobanana-Pro` achieves a $\mathcal{R}_{\text{inv}}$ of 73.41% with strong Judge scores on C&F (1.59) and L&P (1.87), whereas open-source models such as `HunyuanImage-3.0` and `Qwen-Image` remain below 40% $\mathcal{R}_{\text{inv}}$ and score under 0.8 on these structure-sensitive dimensions. This persistent gap indicates that scaling model size alone is insufficient for scientific diagram generation without stronger inductive biases or explicit constraint mechanisms.

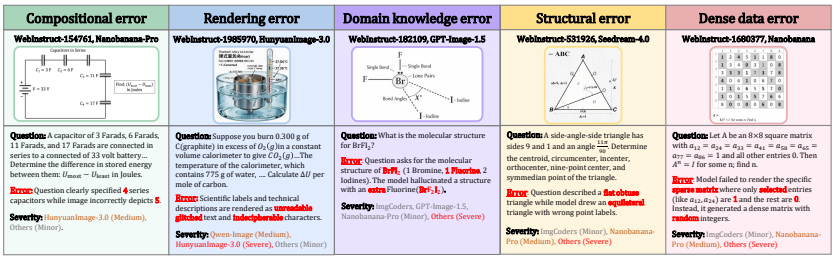

Fig. 3: Qualitative Error Taxonomy. We categorize failures into five modes ranging from low-level visual artifacts to high-level semantic hallucinations. The specific errors are annotated in red.

Advantages of ImgCoder. Interestingly, ImgCoder models consistently achieve the strong inverse validation rates and LMM-as-Judge performance, particularly on structure- and reasoning-sensitive dimensions such as C&F, L&P, and SP.

Notably, `Gemini-3-Pro-ImgCoder` attains the top $\mathcal{R}_{\text{inv}}$ (77.87%), surpassing all pixel-based models. Besides, `Qwen3-ImgCoder` achieves a large improvement over `Qwen-Image` (56.38% vs. 38.86%) despite being fully open-source, confirming that code-based execution—rather than closed data—drives the observed gains.

Disparity between Evaluation Metrics. We observe a clear divergence between perceptual standard metrics and reasoning-oriented evaluations. Models with competitive PSNR or FID scores do not necessarily achieve high inverse validation rates or judge scores. A closer inspection of judge dimensions reveals that perceptual fidelity primarily correlates with E&R, while exhibiting weak alignment with C&F and L&P, which require strict logical consistency and structural correctness. This discrepancy underscores the limitation of previous pixel-level metrics in evaluating scientific diagrams, where visual similarity can mask subtle yet critical factual/relational errors.

Spiral Co-evolution Hypothesis. We hypothesize a complementary co-evolution between code-based and pixel-based paradigms. On one hand, code-based methods enforce structural and logical validity, and as LMMs improve, such structured reasoning can transfer to pixel-based T2I models through shared backbones or by using code-rendered synthetic images as training data. Empirically, `Nanobanana-Pro` (`Gemini-3-Pro-Image`) and `Gemini-3-ImgCoder`, which share the `Gemini-3` backbone, exhibit highly similar diagram construction strategies across many samples (Fig. 2, more examples in appendix), supporting this transfer. On the other hand, pixel-based models contribute visually diverse and expressive imagery that enriches multimodal training data, which in turn benefits code-based reasoning and LMMs learning, forming a mutually reinforcing spiral between reasoning and synthesis.

5.2 Qualitative Analysis

Error Taxonomy. As illustrated in Fig. 3, we categorize observed failures into five primary classes based on a systematic audit. **(1) Compositional error:** Misalignment among multiple visual elements, including attribute leakage (e.g., incorrect label binding), spatial relation confusion, or incorrect object counts. **(2) Rendering error:** Low-level fidelity issues such as illegible text glyphs, blurred lines, or visually corrupted labels. **(3) Structural error:** Violations of geometric logic or topological integrity, including distorted shapes, non-closed curves, or logically inconsistent intersections (e.g., broken parallelism). **(4) Dense data error:** Failures in high-information-density scenarios (e.g., tables or matrices), manifested as coordinate drift, axis misalignment, or collapsed rows and columns. **(5) Domain knowledge error:** Scientific hallucinations where images appear visually plausible but violate domain-specific laws (e.g., incorrect electron configurations or optical paths). We observe a hierarchy in failure modes. Traditional text-to-image errors (compositional, rendering) are largely resolved in SOTA models, persisting only in weaker baselines. Domain knowledge error serves as a watershed: while open-source models struggle, closed-source models have significantly reduced scientific hallucinations. However, structural and dense data errors remain the most persistent bottlenecks. These tasks demand high-precision

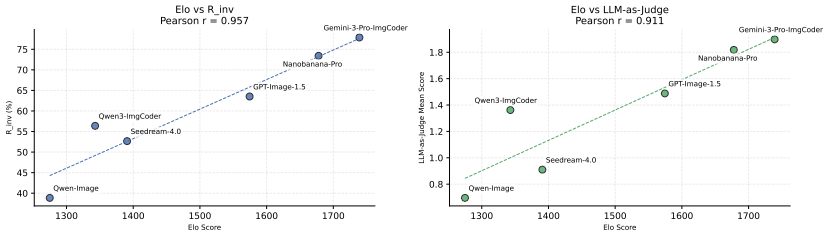

Fig. 4: Correlation between human-derived Elo scores and automatic evaluation metrics ($\mathcal{R}_{\text{inv}}$ and LMM-as-Judge). Both metrics show strong alignment with human preferences in terms of absolute scores and relative rankings.

rendering and information density that challenge the probabilistic nature of diffusion models, causing even `Nanobanana-Pro` to falter. Notably, `ImgCoder` outperforms pixel-based SOTAs in these high-precision regimes due to its rigorous code-driven grounding.

Precision-Expressiveness Trade-off. In Fig. 2, the two paradigms exhibit distinct trade-offs. While code-based methods may fall slightly short in visual expressiveness—often producing schematic or "flat" diagrams compared to the rich renderings of `Nanobanana-Pro` (e.g., the textured spring and hand illustration)—they offer an intrinsic advantage in precision. Pixel-based models, constrained by their probabilistic generation nature, fundamentally struggle to guarantee strict mathematical accuracy. This is evident in the function plotting task ($y = x \ln x$), where `Nanobanana-Pro` produces a plausible-looking curve but fails to capture the correct intercepts and extrema. In contrast, code-based methods leverage deterministic execution engines to ensure the rigorous alignment of geometric shapes and data coordinates, effectively solving the "hallucination" problem in quantitative visualization.

5.3 Human–Judge Consistency Analysis

Since automated metrics rely on model capability, we conducted a human evaluation study to assess the reliability of LMM-as-Judge. Ten annotators with STEM backgrounds performed pairwise preference comparisons on the outputs generated by different methods under the same evaluation criteria, the evaluation covered 500 different samples. To balance evaluation coverage and annotation cost, we selected a representative subset of models for comparison. Annotators were asked to choose a winner, or select one of the following options: "both good," "both bad," or "problematic question." Questions marked as problematic were filtered out from the final dataset (Sec. 4.3).

We then computed the Elo scores of these models ($\sigma = 400$), with initial scores set to 1500 and $K = 32$. Since Elo computation is order-dependent, we shuffled the win–loss pairs and repeated the calculation 100 times, and reported the average scores. Finally, we conducted a correlation analysis between the Elo scores derived from human preferences and the $\mathcal{R}_{\text{inv}}$ and LMM-as-Judge scores.

As shown in Fig. 4, the scores and rankings of the two metrics are highly consistent with human preferences. Under different evaluation settings, the relative ranking among methods remains stable. These results indicate that, for scientific diagram generation tasks, these metrics can serve as a reasonable proxy for human comparative evaluation, and the selected `Gemini-3-Flash` model demonstrates reliable discriminative capability.

5.4 Domain-specific Breakdown

Table 3: Breakdown of Performance by Subject. We report the inverse validation rate ($\mathcal{R}_{\text{inv}}$) and the LMM-as-Judge mean score (Judge) for each discipline. The last row reports the average performance across all baselines.

Model	Math		Physics		Chemistry		Biology		Universal	
	$\mathcal{R}_{\text{inv}}$ (%)	Judge	$\mathcal{R}_{\text{inv}}$ (%)	Judge	$\mathcal{R}_{\text{inv}}$ (%)	Judge	$\mathcal{R}_{\text{inv}}$ (%)	Judge	$\mathcal{R}_{\text{inv}}$ (%)	Judge
Open-source T2I Models										
HunyuanImage-3.0	13.70	0.56	25.39	0.94	22.92	0.86	33.33	1.06	12.58	0.68
Qwen-Image	30.14	0.57	28.12	0.85	40.28	0.66	43.14	0.95	15.23	0.62
Closed-source T2I Models										
GPT-Image-1	31.51	1.07	32.42	1.26	38.19	1.20	49.02	1.38	29.14	1.09
Flux2-flex	41.78	0.83	53.52	1.18	<u>60.42</u>	1.06	58.82	1.17	36.42	0.86
Seedream-4.0	42.47	0.78	51.56	0.98	46.53	1.11	41.18	1.12	37.09	0.82
Nanobanana	43.84	0.78	51.17	1.09	51.39	1.00	66.67	1.20	31.13	0.77
GPT-Image-1.5	52.05	1.43	60.16	1.61	56.94	1.48	72.55	1.58	47.68	1.41
Nanobanana-Pro	64.38	1.80	62.50	1.84	54.86	**1.80**	74.51	**1.89**	60.93	1.82
Code-based Generation										
AutomaTikZ	0.00	0.38	2.34	0.33	4.17	0.26	17.65	0.32	2.65	0.43
Qwen3-ImgCoder	44.52	1.54	43.36	1.39	52.08	1.14	41.18	0.96	43.05	1.34
Gemini-3-Flash-ImgCoder	<u>66.44</u>	<u>1.89</u>	<u>70.31</u>	<u>1.91</u>	**63.19**	<u>1.76</u>	**80.39**	1.82	<u>71.52</u>	<u>1.89</u>
Gemini-3-Pro-ImgCoder	**69.86**	**1.94**	**75.39**	**1.93**	59.03	<u>1.76</u>	<u>76.47</u>	<u>1.84</u>	**72.85**	**1.91**
Average	45.52	1.20	50.36	1.36	49.62	1.26	57.93	1.36	41.60	1.20

As shown in Tab. 3, We analyze performance across domains to understand how generation paradigms interact with domain-specific requirements. Mathematics, physics, and universal diagram types demand strict geometric, symbolic, and layout accuracy, where code-based ImgCoder models consistently outperform pixel-based approaches. In contrast, biology and visually rich chemistry subdomains favor pixel-based models, benefiting from organic shapes and dense textures. Chemistry represents a boundary case: molecular structures favor code-based reasoning, while crystal and reaction diagrams are better handled by pixel-based models. Overall, domain performance depends more on structural constraints and information density than on model scale, highlighting the potential of hybrid approaches for scientific image synthesis. Additional per-domain results and qualitative examples are provided in the supplementary material.

5.5 Additional Validation and Ablation

Evaluator Robustness. To address the potential circularity of LMM-based evaluation, we further conduct a cross-family judge validation by replacing the

Table 4: Cross-family judge validation with GPT-5-mini.

System	$R_{\text{inv}} \uparrow$ (%) G-Flash	GPT-5-m	Δ	Judge $\uparrow$ G-Flash	GPT-5-m	Δ
G3-Pro-ImgCoder	**77.87**	75.16	−2.71	**1.90**	1.83	−0.06
G3-Flash-ImgCoder	76.93	73.59	−3.34	1.88	1.81	−0.06
Nanobanana-Pro	73.41	**77.31**	+3.90	1.82	**1.91**	+0.09
GPT-Image-1.5	63.52	56.09	−7.43	1.49	1.72	+0.24
Flux2	58.83	57.63	−1.20	1.02	1.56	+0.54
Nanobanana	57.75	56.52	−1.23	0.94	1.48	+0.54
Qwen3-ImgCoder	56.38	51.19	−5.19	1.36	1.54	+0.18
Seedream-4.0	52.67	48.56	−4.11	0.91	1.39	+0.48
GPT-Image-1	42.97	40.09	−2.88	1.17	1.49	+0.32
Qwen-Image	38.86	36.94	−1.92	0.70	1.33	+0.63
Spearman ρ	1.00	0.927^{**}	—	1.00	0.927^{**}	—
Pearson r	1.00	0.979^{**}	—	1.00	0.962^{**}	—

$^{**}p<0.001$.

Table 5: Planned ablation of ImgCoder components.

Variant	Plan	Code	$R_{\text{inv}} \uparrow$	Tokens ↓	Runtime ↓
Gemini3-F-ImgCoder	✓	✓	76.9	7412	37.5s
w/o Plan		✓	74.8	3663	18.7s
w/o Code	✓		75.5	3205	16.9s
w/o Plan & Code (Direct)			74.3	3046	16.1s

original `Gemini-3-Flash` evaluator with `GPT-5-mini`. As shown in Tab. 4, the rankings and scores remain highly correlated across the two judge families, suggesting that our main conclusions are not artifacts of a single evaluator.

ImgCoder Components Ablation. We further ablate the two key components of ImgCoder in Tab. 5. Removing either planning or code rendering decreases $\mathcal{R}_{\text{inv}}$, while removing both leads to the largest drop. The full pipeline achieves the best performance at the cost of higher token usage and runtime, indicating a clear test-time scaling effect.

6 Data Utility for Downstream Reasoning

While Sec. 5 assessed intrinsic visual quality, this section evaluates the extrinsic utility of generated images as training data for LMMs. Specifically, we address **RQ3 - Data Utility**: Whether synthetic scientific images can improve downstream multimodal reasoning via fine-tuning?

6.1 Downstream Training Setup

To evaluate whether higher-quality synthesized diagrams improve reasoning performance, we perform reinforcement learning tuning on `Qwen3-VL-8B-Instruct` using four synthetic datasets generated by different pipelines: `Nanobanana-Pro`, `Qwen-Image`, `Qwen-ImgCoder`, and `Gemini-Flash-ImgCoder`. For a fair comparison, we additionally construct filtered subsets by removing incorrect images, resulting in `Nanobanana-Pro (Filt)` and `Qwen-Image (Filt)`. Training follows a standard rollout-based RL procedure implemented with VeRL [41]. We train for 200 global steps with multiple rollouts per prompt to stabilize reward estimation. Evaluation is conducted on $\text{MathVision}_{\text{mini}}$ [43] and $\text{Geometry3K}_{\text{test}}$ [31].

Table 6: Downstream Performance Comparison.

RL Data	GEO3K	MV	AVG
Base	61.9	39.0	54.5
Qwen-Image	68.2	45.9	57.1
Qwen-Imgcoder	67.9	46.9	57.4
Qwen-Image (Filt)	68.6	<u>47.0</u>	57.8
Nanobanana-Pro	**70.1**	46.1	<u>58.1</u>
Gemini-ImgCoder	<u>69.1</u>	46.9	58.0
Nanobanana-Pro (Filt)	68.7	**47.7**	**58.2**

To ensure stable results under limited test size, we report average performance over multiple samples. For answer verification and reward computation, we employ `Compass-Verifier-8B` [29] as a unified judge model. More implementation details are provided in supplementary material.

6.2 Effectiveness of Synthetic Image Training

As shown in Tab. 6, training with synthetic images yields substantial performance improvements over the baseline across all model variants on both GEO3K and MathVision. For instance, the best-performing `Nanobanana-Pro (Filt)` achieves an average score of 58.2, corresponding to a 3.7-point absolute gain over the baseline (54.5). Consistently, Fig. 5a shows a steadily increasing reward trend during training, indicating stable and effective optimization dynamics. Notably, ImgCoder variants show a clear advantage on the more challenging and discriminative MV benchmark compared to pixel-based counterparts, indicating that ImgCoder can serve as a low-cost and scalable data synthesis approach with strong structural rigor.

6.3 Impact of Synthetic Image Quality

We further examine the effect of synthetic image quality on downstream performance. Higher-quality teachers (e.g., `Nanobanana-Pro`) consistently outperform lower-quality pixel-based counterparts (e.g., `Qwen-Image`) on both validation and test benchmarks (58.1 vs. 57.1 AVG). Moreover, applying data filtration consistently improves both training rewards (Fig. 5a) and test accuracy (Fig. 5b), indicating that downstream performance is sensitive to the quality of synthetic images rather than data quantity alone. We note that on GEO3K, filtering does not always lead to monotonic improvements. This may stem from GEO3K's smaller scale and geometry-focused nature, which leads to higher variance across training settings. In contrast, MathVision shows consistent gains from higher-quality supervision, suggesting that structural fidelity mainly benefits more complex reasoning scenarios. We also observe a slightly slower reward increase for ImgCoder than T2I models, this suggests that richer visual expressiveness can be beneficial under reward-driven optimization.

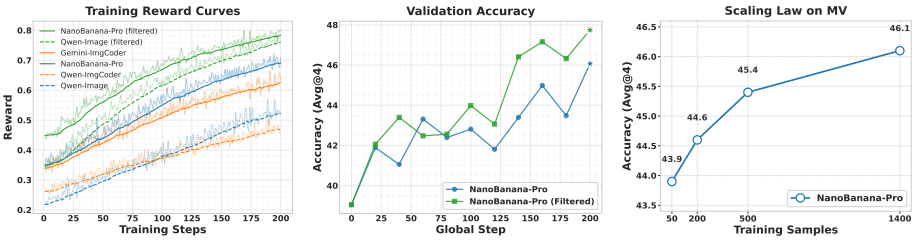

Fig. 5: Downstream Data Utility. *Left (a)*: Stronger teachers yield higher training rewards. *Middle (b)*: Filtered data consistently outperforms unfiltered data. *Right (c)*: Performance scales predictably with data size.

6.4 Scaling Law of Synthetic Data

To further investigate scalability, we conduct a controlled data-scaling study using synthetic images generated by `NanoBanana-Pro`. Training sets ranging from 50 to 1.4K samples are constructed. As shown in Fig. 5c, downstream accuracy exhibits a clear and stable increase with data scale, rising from 43.9% to 46.1% (+2.2%). Notably, the performance curve follows a distinct log-linear growth trend without observable saturation, suggesting that high-fidelity synthetic images can continuously serve as effective training signals. These results suggest that high-fidelity synthetic images provide consistently effective supervision signals rather than diminishing returns, and that the well-known "Data Engine" scaling behavior observed in text-domain training extends naturally to multimodal reasoning scenarios.

7 Conclusion

In this paper, we conduct a systematic study of scientific image synthesis for multimodal reasoning, highlighting a precision–expressiveness trade-off between pixel-based and code-driven generation paradigms. We introduce SciGenBench, a dedicated benchmark designed to assess structural correctness and information utility in generated scientific diagrams. Our experiments show that code-based methods such as ImgCoder provide stronger structural precision and reduce common logical failure modes, while pixel-based models offer greater visual expressiveness and flexibility. Furthermore, we find that high-fidelity synthetic diagrams consistently enhance downstream multimodal reasoning performance. Overall, these findings suggest that scalable, logic-aware image synthesis serves as a valuable component for improving reasoning capabilities in LMMs.

Acknowledgements

This work is supported by Shanghai Artificial Intelligence Laboratory.

References

1. Bai, S., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025), https://arxiv.org/abs/2511.21631
2. Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv e-prints pp. arXiv–2506 (2025)
3. Belouadi, J., Lauscher, A., Eger, S.: Automatikz: Text-guided synthesis of scientific vector graphics with tikz. arXiv preprint arXiv:2310.00367 (2023)
4. Cao, S., Chen, H., Chen, P., Cheng, Y., Cui, Y., Deng, X., Dong, Y., Gong, K., Gu, T., Gu, X., Hang, T., Huang, D., Jiang, J., Jiang, Z., Kong, W., Li, C., Li, D., Li, J., Li, X., Li, Y., Li, Z., Li, Z., Lin, J., Linus, Liu, L., Liu, S., Liu, S., Liu, Y., Liu, Y., Long, Y., Lu, F., Lu, Q., Peng, Y., Peng, Y., Shen, X., Shi, Y., Tao, J., Tao, Y., Tian, Q., Wan, P., Wang, C., Wang, K., Wang, L., Wang, L., Wang, L., Wang, Q., Wang, W., Wen, H., Wu, B., Wu, J., Wu, Y., Xie, S., Yang, F., Yang, M., Yang, X., Yang, X., Yang, Z., Yu, J., Yuan, Z., Zhang, C., Zhang, J.W., Zhang, P., Zhang, S.X., Zhang, T., Zhang, W., Zhang, Y., Zhang, Y., Zhang, Z., Zhang, Z., Zhao, P., Zhao, Z., Zhe, X., Zhu, J., Zhong, Z.: Hunyuanimage 3.0 technical report. arXiv preprint arXiv:2509.23951 (2025)
5. Chang, Y., Feng, Y., Sun, J., Ai, J., Li, C., Zhou, S.K., Zhang, K.: Sridbench: Benchmark of scientific research illustration drawing of image generation model. arXiv preprint arXiv:2505.22126 (2025)
6. Chen, K., Lin, Z., Xu, Z., Shen, Y., Yao, Y., Rimchala, J., Zhang, J., Huang, L.: R2i-bench: Benchmarking reasoning-driven text-to-image generation. arXiv preprint arXiv:2505.23493 (2025)
7. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
8. Chen, Z., Du, Y., Wen, Z., Zhou, Y., Cui, C., Weng, Z., Tu, H., Wang, C., Tong, Z., Huang, Q., et al.: Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? arXiv preprint arXiv:2407.04842 (2024)
9. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
10. Du, S., Zou, Y., Wang, Z., Li, X., Li, Y., Shang, C., Shen, Q.: Unsupervised hyperspectral image super-resolution via self-supervised modality decoupling. International Journal of Computer Vision **134**, 152 (2026)
11. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: International Conference on Machine Learning. pp. 12606–12633. PMLR (2024)
12. Fan, R.Z., Wang, Z., Liu, P.: Megascience: Pushing the frontiers of post-training datasets for science reasoning. arXiv preprint arXiv:2507.16812 (2025)
13. Fang, R., Yu, A., Duan, C., Huang, L., Bai, S., Cai, Y., Wang, K., Liu, S., Liu, X., Li, H.: Flux-reason-6m & prism-bench: A million-scale text-to-image reasoning dataset and comprehensive benchmark. arXiv preprint arXiv:2509.09680 (2025)
14. Google: Gemini 3 Model Card. https://deepmind.google/models/gemini (2025), accessed: 2025-12-28

15. Google: Generate, transform and edit images with simple text prompts, or combine multiple images to create something new. all in gemini. (2025), https://deepmind.google/models/gemini-image/flash/
16. Google: Introducing nano banana pro: Turn your visions into studio-quality designs with unprecedented control, improved text rendering and enhanced world knowledge. (2025), https://blog.google/technology/ai/nano-banana-pro/
17. Guha, E., Marten, R., Keh, S., Raoof, N., Smyrnis, G., Bansal, H., Nezhurina, M., Mercat, J., Vu, T., Sprague, Z., Suvarna, A., Feuer, B., Chen, L., Khan, Z., Frankel, E., Grover, S., Choi, C., Muennighoff, N., Su, S., Zhao, W., Yang, J., Pimpalgaonkar, S., Sharma, K., Ji, C.C.J., Deng, Y., Pratt, S., Ramanujan, V., Saad-Falcon, J., Li, J., Dave, A., Albalak, A., Arora, K., Wulfe, B., Hegde, C., Durrett, G., Oh, S., Bansal, M., Gabriel, S., Grover, A., Chang, K.W., Shankar, V., Gokaslan, A., Merrill, M.A., Hashimoto, T., Choi, Y., Jitsev, J., Heckel, R., Sathiamoorthy, M., Dimakis, A.G., Schmidt, L.: Openthoughts: Data recipes for reasoning models (2025), https://arxiv.org/abs/2506.04178
18. Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: 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)
19. Huang, G., Mao, J., Huang, F., Liu, F., Luo, X., Liang, Y., Lu, J., Wang, X., Liu, P., Fu, R., Huang, R., Huang, S.L.: Exposure bias can alleviate itself via directional and frequency rectification in flow matching (2026), https://arxiv.org/abs/2606.28226
20. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems **36**, 78723–78747 (2023)
21. Huang, Y., Liu, X., Gong, Y., Gou, Z., Shen, Y., Duan, N., Chen, W.: Key-point-driven data synthesis with its enhancement on mathematical reasoning. arXiv preprint arXiv:2403.02333 (2024)
22. Hugging Face: Open r1: A fully open reproduction of deepseek-r1. https://github.com/huggingface/open-r1 (2025), gitHub repository
23. Jiang, T., Wu, L., Xia, S., Li, S., Yan, Z., Yang, H., Qiao, Y., Wang, Y.: Imagine before you predict: Interleaved latent visual reasoning for video event prediction. arXiv preprint arXiv:2606.05769 (2026)
24. Jiang, T., Xia, S., Xu, Y., Wu, L., Zeng, X., Wang, L., Qiao, Y., Wang, Y.: Vknowu: Evaluating visual knowledge understanding in multimodal llms. arXiv preprint arXiv:2511.20272 (2025)
25. Labs, B.F.: FLUX.2: Frontier Visual Intelligence (2025)
26. Lin, H., Liu, Z., Zhu, Y., Qin, C., Lin, J., Shang, X., He, C., Zhang, W., Wu, L.: Mmfinereason: Closing the multimodal reasoning gap via open data-centric methods. arXiv preprint arXiv:2601.21821 (2026)
27. Lin, H., Pan, Z., Li, Y., Pei, Q., Gao, X., Cai, M., He, C., Wu, L.: Metaladder: Ascending mathematical solution quality via analogical-problem reasoning transfer. arXiv preprint arXiv:2503.14891 (2025)
28. Lin, H., Pei, Q., Gao, X., Pan, Z., Li, Y., Li, J., He, C., Wu, L.: Scaling code-assisted chain-of-thoughts and instructions for model reasoning. arXiv preprint arXiv:2510.04081 (2025)
29. Liu, S., Liu, H., Liu, J., Xiao, L., Gao, S., Lyu, C., Gu, Y., Zhang, W., Wong, D.F., Zhang, S., Chen, K.: CompassVerifier: A Unified and Robust Verifier for Large Language Models. arXiv preprint arXiv:2508.03686 (2025)

30. Liu, Z., Lin, H., Wang, X., Gao, X., Li, Y., Cai, M., Zhu, Y., Zhong, Z., Pei, Q., Pan, Z., et al.: Chartverse: Scaling chart reasoning via reliable programmatic synthesis from scratch. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7551–7577 (2026)
31. Lu, P., Gong, R., Jiang, S., Qiu, L., Huang, S., Liang, X., Zhu, S.C.: Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In: The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021) (2021)
32. Ma, X., Liu, Q., Jiang, D., Zhang, G., MA, Z., Chen, W.: General-Reasoner: Advancing LLM reasoning across all domains. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=pBFVoll8Xa
33. Ni, Y., Nie, P., Zou, K., Yue, X., Chen, W.: Viscoder: Fine-tuning llms for executable python visualization code generation. arXiv preprint arXiv:2506.03930 (2025)
34. OpenAI: Gpt-5 system card. Tech. rep., OpenAI (2025), https://cdn.openai.com/gpt-5-system-card.pdf
35. OpenAI: The new chatgpt images is here (2025), https://openai.com/index/new-chatgpt-images-is-here/
36. Pan, Z., Li, Y., Lin, H., Pei, Q., Tang, Z., Wu, W., Ming, C., Zhao, H.V., He, C., Wu, L.: Lemma: Learning from errors for mathematical advancement in llms. arXiv preprint arXiv:2503.17439 (2025)
37. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
38. Pei, Q., Wu, L., Pan, Z., Li, Y., Lin, H., Ming, C., Gao, X., He, C., Yan, R.: Mathfusion: Enhancing mathematical problem-solving of llm through instruction fusion. arXiv preprint arXiv:2503.16212 (2025)
39. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: The Twelfth International Conference on Learning Representations (2024)
40. Seedream, T., :, Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., Jian, X., Kuang, H., Lai, Z., Li, F., Li, L., Lian, X., Liao, C., Liu, L., Liu, W., Lu, Y., Luo, Z., Ou, T., Shi, G., Shi, Y., Sun, S., Tian, Y., Tian, Z., Wang, P., Wang, R., Wang, X., Wang, Y., Wu, G., Wu, J., Wu, W., Wu, Y., Xia, X., Xiao, X., Xu, S., Yan, X., Yang, C., Yang, J., Zhai, Z., Zhang, C., Zhang, H., Zhang, Q., Zhang, X., Zhang, Y., Zhao, S., Zhao, W., Zhu, W.: Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint arXiv:2509.20427 (2025)
41. Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., Wu, C.: Hybridflow: A flexible and efficient rlhf framework. arXiv preprint arXiv:2409.19256 (2024)
42. Sun, K., Fang, R., Duan, C., Liu, X., Liu, X.: T2i-reasonbench: Benchmarking reasoning-informed text-to-image generation. arXiv preprint arXiv:2508.17472 (2025)
43. Wang, K., Pan, J., Shi, W., Lu, Z., Ren, H., Zhou, A., Zhan, M., Li, H.: Measuring multimodal mathematical reasoning with math-vision dataset. Advances in Neural Information Processing Systems **37**, 95095–95169 (2024)

44. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
45. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
46. Xiang, K., Li, H., Zhang, T.J., Huang, Y., Liu, Z., Qu, P., He, J., Chen, J., Yuan, Y.J., Han, J., et al.: Seephys: Does seeing help thinking?–benchmarking vision-based physics reasoning. arXiv preprint arXiv:2505.19099 (2025)
47. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
48. Yang, Y., Patel, A., Deitke, M., Gupta, T., Weihs, L., Head, A., Yatskar, M., Callison-Burch, C., Krishna, R., Kembhavi, A., et al.: Scaling text-rich image understanding via code-guided synthetic multimodal data generation. arXiv preprint arXiv:2502.14846 (2025)
49. Yang, Y., Zhang, Z., Hou, Y., Li, Z., Liu, G., Payani, A., Ting, Y.S., Zheng, L.: Effective training data synthesis for improving mllm chart understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2653–2663 (2025)
50. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 **2**(3), 5 (2022)
51. Zhang, L., Eger, S., Cheng, Y., Zhai, W., Belouadi, J., Leiter, C., Ponzetto, S.P., Moafian, F., Zhao, Z.: Scimage: How good are multimodal large language models at scientific text-to-image generation? arXiv preprint arXiv:2412.02368 (2024)
52. Zhang, T., Lin, H., Liu, Z., Chen, C., Zhang, W.: Sciflow-bench: Evaluating structure-aware scientific diagram generation via inverse parsing. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 17747–17765 (2026)
53. Zhao, X., Zhang, P., Tang, K., Zhu, X., Li, H., Chai, W., Zhang, Z., Xia, R., Zhai, G., Yan, J., et al.: Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. arXiv preprint arXiv:2504.02826 (2025)
54. Zhuo, L., Han, S., Pu, Y., Qiu, B., Paul, S., Liao, Y., Liu, Y., Shao, J., Chen, X., Liu, S., et al.: Factuality matters: When image generation and editing meet structured visuals. arXiv preprint arXiv:2510.05091 (2025)