SciIR: A Large-scale Training Dataset and Benchmark for Scientific Image Reasoning Generation

Zhiyuan Ma[1]*, Zhengfeng Shi[1,2*], Yuning An[1], Peize Li[1], Jiabao Wei[1], Ruijie Li[1], Junhao Xiao[1], Jianjun Li[1]†, and Bowen Zhou[3]

[1] School of Computer Science and Technology, Huazhong University of Science and Technology, China
[2] School of Airspace Science and Engineering, Shandong University, China
[3] Department of Electronic Engineering, Tsinghua University, China
{mzyth,jianjunli}@hust.edu.cn, zhengfengshi@mail.sdu.edu.cn

https://github.com/MAIR-Lab-HUST/SciIR

https://huggingface.co/datasets/MAIR-Lab-HUST/SciIR-82k

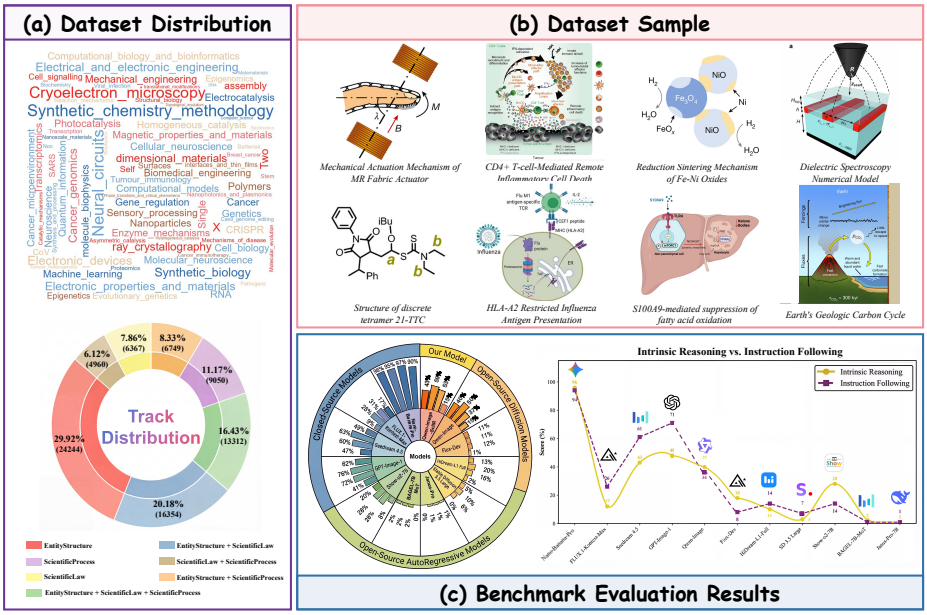

Fig. 1: Overview of SciIR. (a) SciIR-82k: keyword word cloud and distribution across semiotic-oriented image generation tracks. (b) Example figures from diverse domains. (c) Illustration of SciIR-Bench results across various open- and closed-source models with a comparison of Intrinsic Reasoning vs. Instruction Following.

* Equal contribution.
† Corresponding author.

Abstract. While Text-to-Image (T2I) models have shown remarkable success in generating photorealistic visual content, they still struggle with the rigorous semantic alignment and logical reasoning required for scientific imagery. Inspired by Peirce's Semiotic Triad, we introduce Scientific Image Reasoning (SciIR), a comprehensive resource for training and evaluation of scientific image generation. We formalize scientific reasoning into three core dimensions: Entity Structure (*Icon*), Scientific Process (*Index*), and Scientific Law (*Symbol*). Specifically, to overcome the scarcity of training data in scientific image generation, we elaborately create SciIR-82k, a large-scale dataset containing over 80,000 high-quality scientific image-text pairs from cutting-edge publications. The dataset is hierarchically organized according to the semiotic dimensions and incorporates a Scientific Reasoning Chain-of-Thought (Sci-RCoT) to explicitly model underlying visual logic. For evaluation, we propose SciIR-Bench, which aligns with these three semiotic levels and employs an Atomic Checklist to convert the outcome-oriented scientific accuracy into process-oriented, verifiable, fine-grained questions. Our extensive experiments reveal significant deficiencies in current models' scientific reasoning capabilities. Furthermore, by fine-tuning on the SciIR-82k dataset, we developed the Qwen-Image-SciIR model, which achieves a substantial improvement on the SciIR-Bench, increasing the final score from 35% to 43%, laying a solid foundation for future advances in scientific image generation.

Keywords: Text-to-Image generation · Scientific Image Reasoning

1 Introduction

Recent advances in Text-to-Image (T2I) generation have yielded high-quality models: diffusion-based approaches [8, 10, 24, 25, 30, 42, 45] deliver exceptional visual realism and stylistic diversity, autoregressive methods [6, 38, 40, 48] excel at semantic alignment and complex instruction following, and emerging reasoning-augmented techniques integrate chain-of-thought (CoT) strategies [20, 28] to help resolve ambiguities and map abstract concepts to concrete visual attributes. Despite these gains, T2I models remain fundamentally constrained when required to perform rigorous reasoning under complex, multi-constraint scenarios—most notably in scientific imagery, which must strictly adhere to physical laws, accurate topology, and causal logic. Even state-of-the-art closed-source systems (*e.g.*, `Nano-Banana`) that achieve high perceptual fidelity and object-level consistency frequently violate domain-specific logical constraints, producing results that are "visually plausible but factually incorrect".

By contrast, open-source alternatives [2, 7, 8, 18, 23, 41, 43] face more fundamental challenges: in knowledge-intensive contexts, they often struggle to internalize domain-specific knowledge and satisfy strict scientific constraints. This divergence underscores a vital principle: *Perceptual fidelity does not equate to reasoning robustness, and visual polish cannot compensate for the absence of scientific*

validity. These methodological gaps crystallize into three bottlenecks that impede scientific image generation:

1. *Data aspect — scarcity of logic-annotated resources.* High-quality scientific figures with explicit reasoning annotations are scarce because producing and annotating them requires deep, domain-specific expertise. Existing datasets therefore lack the "visual logic" necessary for models to learn the dependencies required to map text into scientifically accurate structures.
2. *Evaluation aspect — lack of scientific-correctness standards.* Current evaluation frameworks leave a structural gap in assessing scientific correctness. Although recent tools such as AutoFigure [50] and PaperBanana [49] facilitate automated figure generation, their evaluation emphasizes layout and workflow rather than underlying scientific logic. Moreover, benchmarks (e.g., SridBench [3]) often do not offer fine-grained semantic diagnosis, making it hard to define a verifiable ground truth for auditing the multidimensional constraints in scientific imagery.
3. *Model aspect — deficiencies in enforcing scientific constraints.* Open-source models in particular lack specialized scientific reasoning and therefore struggle to satisfy hard constraints such as topology or reaction causality, frequently producing hallucinations that violate basic physical laws.

Motivated by the above three aspects, we propose **SciIR**, a semiotic triad-based dataset and benchmark designed to promote scientific rigor in image reasoning generation, as shown in Fig. 1. In short, our main contributions are summarized as follows:

- We construct **SciIR-82k**, a large-scale refined dataset of over $80,000$ scientific image-text pairs from *Nature* and *Nature Communications*. This dataset is enhanced with Sci-RCoT annotations, which explicitly formalize latent visual reasoning pathways to train models on underlying scientific logic.
- We introduce **SciIR-Bench**, the first benchmark to systematically categorize evaluation tracks based on multidimensional scientific correctness, employing a novel atomic checklist to provide fine-grained, verifiable questions.
- We develop **Qwen-Image-SciIR**, a strong open-source baseline that boosts the final score on SciIR-Bench from 35% to 43% by fine-tuning Qwen-Image-2512 on SciIR-82k, and serves as a reliable starting point for future research on scientific image reasoning generation.

2 Related Work

2.1 Text-to-Image Datasets

Current text-to-image (T2I) datasets generally lack the cognitive depth required for complex scientific synthesis. As shown in Tab. 1, both synthetic and non-synthetic collections provide predominantly descriptive captions, omitting explicit reasoning chains or structured semantic relations. Specifically, while

Table 1: Comparison of SciIR-82k with representative T2I datasets.

Dataset	Scale	Text Length	Reasoning Type
Synthetic Image Datasets			
JourneyDB [37]	4M	Short	None
PixelProse [35]	16M	Long	None
FLUX-Reason-6M [9]	6M	Short	Visual
Science-T2I [19]	20K	Short	Outcome-oriented
Non-Synthetic Image Datasets			
CC-12M [4]	12M	Short	None
LAION-Aesthetics [33]	120M	Short	None
TextCaps [34]	28K	Short	None
DOCCI [27]	15K	Long	None
SciIR-82k (Ours)	**82K**	**Short + Long**	**Process-oriented**

synthetic datasets [9, 19, 35, 37] provide large-scale, controllable supervision, they often inherit biases from source models—prioritizing visual plausibility and stylistic diversity over rigorous logical consistency. Conversely, non-synthetic datasets [4, 27, 33, 34] source web image-text pairs to cover everyday concepts, but their typically short captions lack domain knowledge and explicit logical structures. While scientific datasets like Science-T2I [19] leverage specialized knowledge to mitigate inconsistencies in generated diagrams, such early efforts remain limited in scale and prioritize final visual correctness. Consequently, their reliance on post-hoc preference modeling for implicit, outcome-oriented reasoning fails to support the intrinsic, process-oriented reasoning required during generation. To bridge this gap, we introduce SciIR-82k. Sourced from academic publications, it is enriched with detailed Sci-RCoT annotations that formalize the logical steps for figure construction. By explicitly addressing complex scientific logic spanning Structure, Process, and Law—capturing structural relationships, causal mechanisms, and scientific principles—SciIR-82k provides process-oriented supervision. This explicit formalization allows models to learn rigorous, reasoning-conditioned visualizations from scratch, grasping not merely how scientific images look, but the underlying rationale behind their structured representations.

2.2 Text-to-Image Benchmark

Existing benchmarks evaluate generation through three evolving perspectives. *1) Perceptual Quality:* Early metrics like IS [32] and FID [13] assess distributional realism. *2) Prompt Alignment:* Benchmarks such as T2I-CompBench [15] and GenEval++ [46] quantify textual fidelity via VLM-based scoring [11, 12, 14]. *3) Semantic Plausibility:* Recent frameworks (*e.g.*, WISE [26], T2I-ReasonBench [36]) target logical reasoning using LLMs. Within the domain of scientific figure generation, PaperBananaBench [49] and FigureBench [50] concentrate on the evaluation of flowcharts and statistical diagrams. While SridBench [3] specifically

Table 2: Comparison of SciIR-Bench with representative T2I benchmarks. **SL**: Scientific Law, **ES**: Entity Structure, **SP**: Scientific Process.

Benchmark	Scale	Domain	Evaluation Dimensions				Fine-grained
			SL	ES	SP	Text	
GenEval++ [46]	280	Generic	✗	✓	✗	✗	✓
T2I-CompBench [15]	6k	Generic	✗	✓	✗	✗	✓
WISE [26]	1k	Generic	✓	✗	✓	✗	✗
R2I-Bench [5]	3068	Generic	✗	✗	✓	✗	✓
T2I-ReasonBench [36]	800	Generic	✓	✓	✗	✗	✓
ScImage [47]	–	Method. Diags.	✗	✓	✗	✓	✗
PaperBananaBench [49]	292	Method. Diags.	✗	✓	✓	✓	✗
FigureBench [50]	3300	Method. Diags.	✗	✓	✓	✓	✗
SridBench [3]	1120	Sci. Illus.	✓	✓	✗	✓	✗
SciGenBench [21]	–	Sci. Illus.	✓	✓	✗	✓	✓
SciIR-Bench (Ours)	**800**	**Sci. Illus.**	✓	✓	✓	✓	✓

targets scientific diagrams, it remains focused on broad interpretation rather than fine-grained generative constraints.

To address this structural opacity, we posit that scientific correctness cannot be treated as a monolithic metric but requires systematic decomposition. Drawing inspiration from Peirce's Semiotic Triad [29], we decompose scientific correctness into *law*, *structure*, and *process*. However, we observe that no existing benchmark comprehensively covers all three dimensions (as shown in Tab. 2), failing to diagnose specific atomic violations (*e.g.*, broken causal links). To bridge this gap, we propose SciIR-Bench, a diagnostic framework that holistically assesses these dimensions through explicit, verifiable criteria.

2.3 Methods for Scientific Image Generation

Recent research has also advanced the automated generation of scientific illustrations. For example, AutoFigure [50] focuses on producing publication-ready illustrations from long-form text, emphasizing layout and aesthetics, while PaperBanana specializes in automating and improving the visualization quality of methodological workflows [49]. Alternatively, ImgCoder [21] prioritizes programmatic synthesis (code generation) to circumvent pixel-level reasoning challenges, though it lacks the visual expressivity required for rendering nuanced or complex graphic details. Despite these advances, such efforts remain primarily confined to procedural and architectural diagrams. In contrast, our work targets scientific schematics that encode underlying natural principles (*e.g.*, physical laws and causal mechanisms). Moving beyond layout fidelity, we explicitly model and evaluate *scientific correctness*, requiring models not only to reproduce structural relationships but also to internalize and faithfully represent the domain-specific principles governing the depicted phenomena.

3 SciIR-82k Dataset

To systematically evaluate the scientific reasoning abilities of T2I models, we introduce SciIR-82k, a large-scale dataset of over $80,000$ high-quality scientific image-text pairs with complete annotations. As shown in Fig. 2, our SciIR-82k is grounded in a semiotic triad and constructed through a multi-stage automated pipeline for promoting scientific fidelity.

3.1 Theoretical Taxonomy: Semiotic Triad

Scientific images are not merely visual imagery but abstract structures encoding logical relations and physical constraints. To effectively formalize these, we ground our dataset in Peirce's Semiotic Triad [29], *i.e.*, *Icon*, *Index*, and *Symbol*, which respectively correspond to the cognitive layers for scientific reasoning: *1) Entity Structure* corresponds to *iconic* representation via topological fidelity, which evaluates the geometric hierarchical reconstruction and spatial alignment of scientific entities. *2) Scientific Process* corresponds to *indexical* representation indicating causal or temporal correlations, involving state transitions, experimental workflows, or causal chains. *3) Scientific Law* corresponds to *symbolic* representation governed by abstract rules, ensuring adherence to fundamental laws (*e.g.*, conservation of energy, molecular valence). This taxonomy transcends simple visual-text alignment, establishing a structured framework for deep scientific reasoning.

3.2 Corpus Construction

Our corpus comprises articles licensed under CC BY 4.0 from *Nature* and *Nature Communications* to ensure authority; comprehensive compliance and provenance details are outlined in Appendix A. From about 360k raw figures, we employ Ultralytics-YOLO11 [16] as an automated layout analyzer to decompose multi-panel figures into semantically independent subfigures, which are then standardized to a 1024×1024 resolution. Afterwards, a two-stage filtering pipeline—VLM-based screening with InternVL3.5 [39] to retain corresponding schematics, followed by manual verification—finally extracts over 80k high-quality subfigures and their corresponding captions and content. Data processing details are provided in the Appendix B.

3.3 Semiotic Stratification

We implement a stratification strategy to align the raw data with our semiotic taxonomy (Sec. 3.1). Using Qwen3-VL [1] as a domain-specific evaluator, we assess each sample's relevance score $s \in [1, 10]$ to the three reasoning tracks (`Entity Structure`, `Scientific Process`, or `Scientific Law`). This categorization serves as a routing mechanism, directing images to targeted annotation pipelines according to their dominant semiotic attributes.

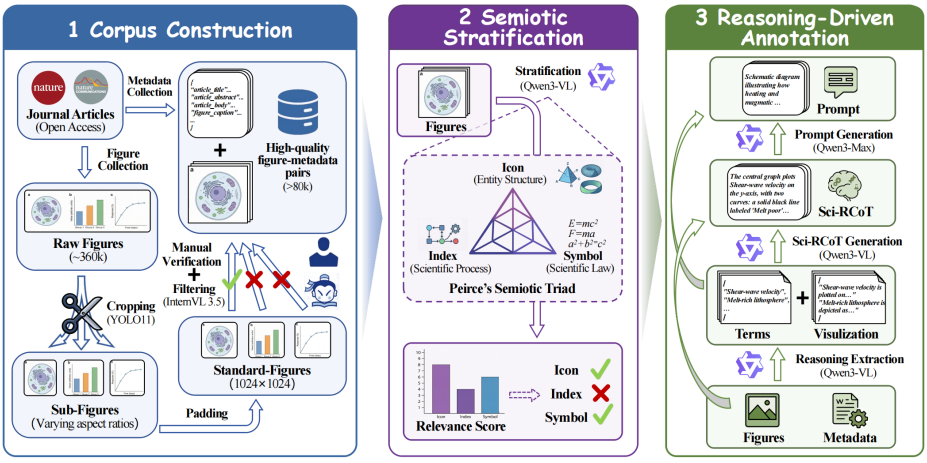

Fig. 2: Overview of the SciIR-82k pipeline grounded in Peirce's Semiotic Triad [29]. The framework comprises three stages: (1) Corpus Construction, employing YOLO11 and InternVL 3.5 for sub-figure extraction and filtering; (2) Semiotic Stratification, categorizing samples into Icon, Index, and Symbol tracks; and (3) Reasoning-Driven Annotation, which leverages Qwen3 models to reverse-engineer Sci-RCoT and prompts.

3.4 Reasoning-Driven Annotation

In contemporary text-to-image and multimodal generation frameworks, the process typically follows a forward pipeline: Abstract Prompt → Logical Deduction → Visual Output, where high-level semantic intent is progressively instantiated into concrete visual representations. For high-fidelity reasoning annotation, we invert this chain via *logical reverse engineering*: reconstructing latent scientific reasoning (*i.e.*, Sci-RCoT) from ground-truth images to derive precise prompts, with Qwen3-VL [1] for visual grounding and Qwen3-Max [44] for semantic abstraction.

Taxonomy-Driven Reasoning Extraction. In the first phase, Qwen3-VL [1] extracts taxonomy-guided structured information, prioritizing visual evidence (Image > Caption > Text). The model parses input into JSON, decoupling: *1) Terms:* Image-validated entities and nomenclatures. *2) Visualization:* Descriptions of visual grounding (*e.g.*, geometry, layout). This taxonomy-driven extraction enforces semantic decomposition of scientific content, reduces hallucination, and converts free-form multimodal understanding into controllable, fine-grained reasoning units suitable for downstream transformation.

Sci-RCoT Generation. Afterwards, Qwen3-VL [1] synthesizes a Scientific Reasoning CoT (Sci-RCoT) by re-examining the image to integrate visual style (*e.g.*, schematic diagrams) and text rendering requirements with "Visualization" entries. By transforming discrete reasoning units into a continuous visual reconstruction process, Sci-RCoT, as an explicit reasoning trace, bridges symbolic reasoning and holistic scene composition, elucidating the causal logic underlying the mapping of abstract concepts to concrete spatial arrangements.

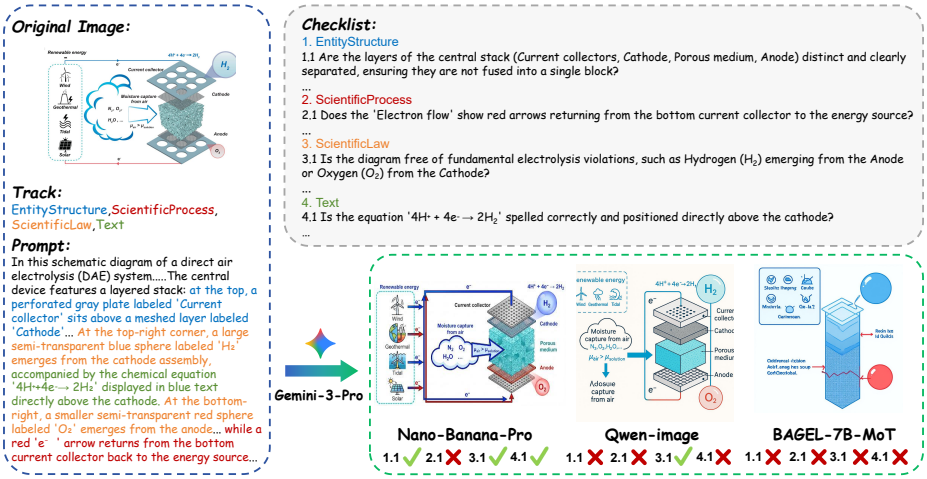

Fig. 3: An evaluation instance from SciIR-Bench. Prompt from a sample covering all four tracks is used to guide various models in generating images. The output is then scrutinized by Gemini-3-Pro using a dimension-specific atomic checklist.

Prompt Generation. Eventually, Qwen3-Max [44] distills Sci-RCoT into a concise prompt via a "Term-Substitution" strategy: Visualization descriptions are replaced with canonical scientific terms while preserving the original visual style as the leading phrase, and only explicitly required textual renderings are retained. This produces a synchronized pair of abstract prompt and retained text. This step removes redundant visual detail while preserving scientific semantics, yielding a compact yet information-complete prompt representation. The resulting abstraction enhances controllability, improves semantic consistency across samples, and provides high-quality structured supervision for process reasoning-aware image generation and evaluation.

Overall, this pipeline performs logical reverse engineering from images to prompts by progressively transforming visual evidence into structured reasoning traces and finally into compact prompt representations. This design enforces explicit alignment between image structures, textual elements, and reasoning units, producing controllable and logically grounded prompts. The resulting annotations provide reliable supervision for evaluating and training process reasoning-aware scientific image generation models.

4 SciIR-Bench

To systematically evaluate the scientific reasoning and generation capabilities of current text-to-image models, we develop the SciIR-Bench (Fig. 3). This benchmark moves beyond traditional holistic image quality metrics and instead measures whether models can faithfully instantiate structured scientific content—such as correctly rendering labeled entities, preserving spatial and

topological relations, and accurately depicting multi-stage processes—without introducing unsupported elements or logical contradictions in the visualization.

4.1 Evaluation Benchmark

Candidate Selection and Filtration. From the massive corpus of processed samples, we distilled a high-quality evaluation benchmark consisting of 800 test instances. To ensure the benchmark challenges the upper limits of current models, we enforced a rigorous protocol focusing on both the breadth of scientific domains and the depth of reasoning complexity. We prioritized samples exhibiting High Term Density (term count > 3) to guarantee sufficient semantic content. Details of the screening process are provided in the Appendix C. Furthermore, to verify multimodal reasoning capabilities, we selected candidates that necessitate compound reasoning across our theoretical dimensions, filtering out samples that do not contain valid reasoning paths in at least two of the three semiotic tracks (`Entity Structure`, `Scientific Process`, or `Scientific Law`).

Taxonomy-Based Grouping. To systematically evaluate model performance across different reasoning intersections, we exploit a four-folds strategy to categorize the filtered candidates into four distinct evaluation groups ($N = 200$ per group). The first group represents the most complex "holistic" reasoning scenario, containing samples that simultaneously encompass attributes of all three tracks: Scientific Law, Entity Structure, and Scientific Process. The remaining three groups are constructed to test pairwise reasoning capabilities, covering the specific intersections of Law-Entity, Law-Process, and Entity-Process, respectively. This combinatorial approach ensures that the benchmark evaluates not just isolated knowledge but the model's ability to synthesize conflicting constraints from multiple semiotic dimensions.

Adaptive Difficulty Stratification. Within each group, we bifurcated samples into two difficulty levels based on scientific term density to strictly disentangle *Instruction Following* from *intrinsic reasoning*: *1) Instruction Following (IF).* Samples with high term density ($>$ median) are paired with the detailed Sci-RCoT. For these semantically saturated images, compressing the input into a concise prompt inevitably leads to *semantic erosion*, causing the model to omit critical scientific details. By providing the exhaustive Sci-RCoT, we eliminate the ambiguity space, thereby strictly testing the model's fidelity in visualizing complex, fine-grained instructions. *2) Intrinsic Reasoning (IR).* Conversely, samples with lower term density are paired with the abstract Prompt. In this regime, the input provides only high-level nomenclature without visual cues. This information sparsity compels the model to bridge the gap using latent domain knowledge, effectively evaluating its capacity to autonomously reason out valid spatial layouts and causal logic from abstract concepts.

4.2 Evaluation Metrics

Traditional generative metrics such as FID [13] or CLIPScore [31] focus primarily on image fidelity or broad semantic similarity. However, these metrics fail to

capture the factual correctness and logical consistency essential for scientific modeling. To bridge this gap, we propose a fine-grained, interpretable evaluation protocol designed to mimic the rigor of human peer review. Unlike black-box scoring, this VLM-driven checklist operationalizes evaluation through a transparent, three-stage automated pipeline comprising ground truth extraction, atomic questioning, and evidence-based refereeing.

Atomic Checklist Generation. We utilize the structured Reasoning content (extracted in Phase 1) as the absolute ground truth, avoiding reliance on potentially noisy reference images. To ensure comprehensive coverage, the checklist generation is strictly term-driven: for every "Scientific Term" identified in the reasoning structure, a VLM (`Gemini-3-Pro`) generates a corresponding binary validation query. This enforces Atomicity, as each question is strictly scoped to verify the visual manifestation of a single semantic unit (*e.g.*, the specific morphology of a protein or the directionality of a process arrow). Moreover, to address the issue of hallucination, we generate supplementary adversarial questions tailored to each specific track to explicitly probe for domain-specific fabrications (*e.g.*, non-existent chemical bonds), ensuring the model is penalized for inventing scientifically invalid details.

Automated Evaluation. In the adjudication phase, an advanced VLM acts as a "Senior Scientific Reviewer" to evaluate the generated images against the checklist. To prevent hallucinated judgments, we enforce a Visual Evidence Retrieval protocol: the referee must explicitly locate and describe the specific visual element mentioned in the query before assigning a verdict. The evaluation logic is strictly compartmentalized by category—Text is judged on spelling and positional exactness, while scientific tracks (`Entity Structure`, `Scientific Process`, `Scientific Law`) are evaluated on topological and causal logic. This separation ensures graphical errors result in scientific penalties, while textual errors are isolated to the text score.

Accuracy Score. Distinct from standard metrics that average accuracy across all questions, we adopt a rigorous sample-level pass rate to reflect the intolerance for error in scientific communication. Formally, let $Q_{i,c}$ be the set of atomic questions generated for a sample i within a specific reasoning category c (*e.g.*, `Scientific Process`), and let $s(q) \in \{0, 1\}$ denote the binary score of a single question q. The validity of a sample i in category c, denoted as $V_{i,c}$, is defined by a strict veto mechanism:

$$V_{i,c} = \begin{cases} 1 & \text{if } \prod_{q \in Q_{i,c}} s(q) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

A sample is considered valid for a specific track only if it passes every atomic check associated with that track; a single failure renders the sample scientifically compromised for that dimension. The final accuracy score for category c is calculated as the percentage of valid samples across the dataset D:

$$\text{Accuracy Score}_c = \frac{1}{|D|} \sum_{i \in D} V_{i,c} \tag{2}$$

Table 3: Evaluation on SciIR-Bench. We report the Accuracy Score (%) for Intrinsic Reasoning (IR), Instruction Following (IF), and overall performance across four distinct tracks. **SL**: Scientific Law, **ES**: Entity Structure, **SP**: Scientific Process.

Model	SL (%)			ES (%)			SP (%)			Text (%)			Final (%)
	IR	IF	Avg.	IR	IF	Avg.	IR	IF	Avg.	IR	IF	Avg.	
Closed-Source Models													
Nano-Banana-Pro	95	97	96	98	94	95	98	97	97	92	89	90	95
FLUX.1-Kontext-Max	16	19	17	15	40	31	13	36	28	3	11	9	22
Seedream 4.5	39	57	49	56	67	63	53	64	60	25	55	47	55
GPT-Image-1	52	72	62	64	82	76	51	82	72	23	47	41	62
Open-Source Diffusion Models													
Qwen-Image-2512	38	42	40	53	46	50	42	32	37	16	14	15	35
Flux-Dev	25	6	11	23	10	11	19	14	12	3	1	1	9
HiDream-L1-Full	12	14	13	14	23	20	15	16	16	1	2	2	13
SD 3.5 Large	3	7	5	6	12	10	2	8	6	0	0	0	5
Open-Source AutoRegressive Models													
Show-o2-7B	28	12	20	42	15	28	32	25	28	12	4	8	21
BAGEL-7B-MoT	1	2	2	4	1	2	3	1	2	0	0	0	2
Janus-Pro-7B	1	2	1	1	2	1	1	1	1	0	0	0	1
Fine-tuned Models (Ours)													
Qwen-Image-SciIR	**37**	**50**	**43**	**56**	**62**	**59**	**52**	**54**	**53**	**14**	**15**	**15**	**43**

This strict metric provides a granular and uncompromising diagnostic of model performance, ensuring that high scores reflect true scientific robustness rather than partial hallucinatory success.

5 Experiments and Analyses

In this section, we provide a comprehensive analysis of model performance on SciIR-Bench. We discuss the quantitative results reported in Tab. 3, analyze the correlation between our metrics and traditional evaluation standards, and qualitatively categorize common failure modes based on the proposed semiotic taxonomy.

5.1 Experimental Settings

Implementation of Qwen-Image-SciIR. Qwen-Image-SciIR is implemented to ensure rigorous zero-shot evaluation: we removed the 800 test instances in SciIR-Bench from the SciIR-82k corpus to avoid data leakage. As shown in Figure 4, the pipeline decouples scientific reasoning from visual synthesis via two fine-tuned modules. The first, Qwen2.5-7B-Instruct, serves as a reasoning planner and was fine-tuned on (prompt, Sci-RCoT) pairs using an all-linear LoRA

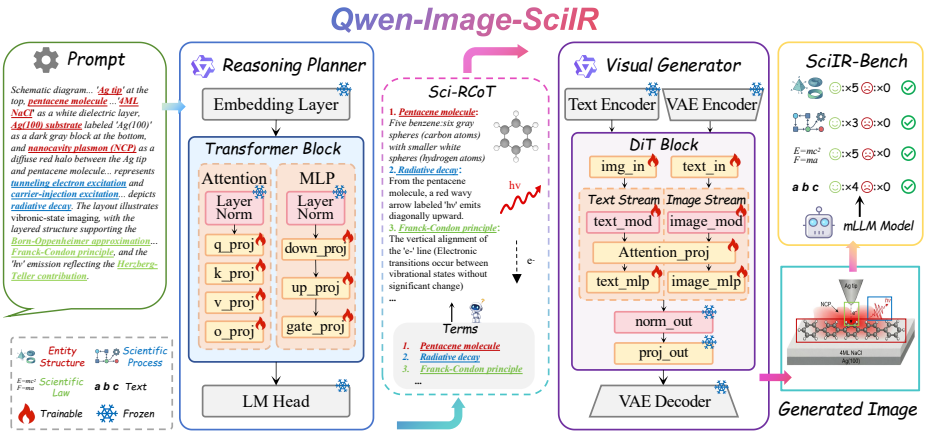

Fig. 4: Qwen-Image-SciIR model architecture.

configuration ($r = 64, \alpha = 16$). Specifically, LoRA adapters were integrated into all linear transformation layers within the Transformer blocks to maximize adaptation capacity. This module was trained with a learning rate of 1×10^{-4} and a maximum context window of 2,048 tokens for one optimization step. The second, Qwen-Image-2512 as a visual generator, was fine-tuned on (Sci-RCoT, image) pairs via LoRA ($r = 32$) applied to the diffusion transformer layers, with a learning rate of 1×10^{-4}, training resolution 1024×1024, and trained for one optimization step.

Inference Protocol. We develop a systematic inference pipeline for Qwen-Image-SciIR. Across 800 evaluation samples, encompassing both Intrinsic Reasoning (IR) and Instruction Following (IF) categories, we employ a chained generation flow. Specifically, the Reasoning Planner first infers a comprehensive Sci-RCoT from the input prompt, which is then utilized by the Visual Generator to synthesize the final image. This unified protocol ensures that the reasoning module is actively engaged for every instance, maintaining a standard reasoning-to-rendering process throughout the entire benchmark evaluation.

5.2 Main Quantitative Results

Tab. 3 presents the systematic evaluation of 12 T2I models across the SciIR-Bench. Our analysis reveals several key insights regarding the current landscape of scientific visual synthesis.

Closed- *vs.* Open-Source. Nano-Banana-pro's near-saturation performance (95%) provides strong evidence that the task is solvable, yet a 60% gap remains for open-source contenders. Critically, aesthetic-focused baselines like Flux-Dev fail (<10%) on strict tracks, confirming a fundamental misalignment: current open-source training optimizes for *perceptual fidelity*, sacrificing the logic essential for scientific accuracy.

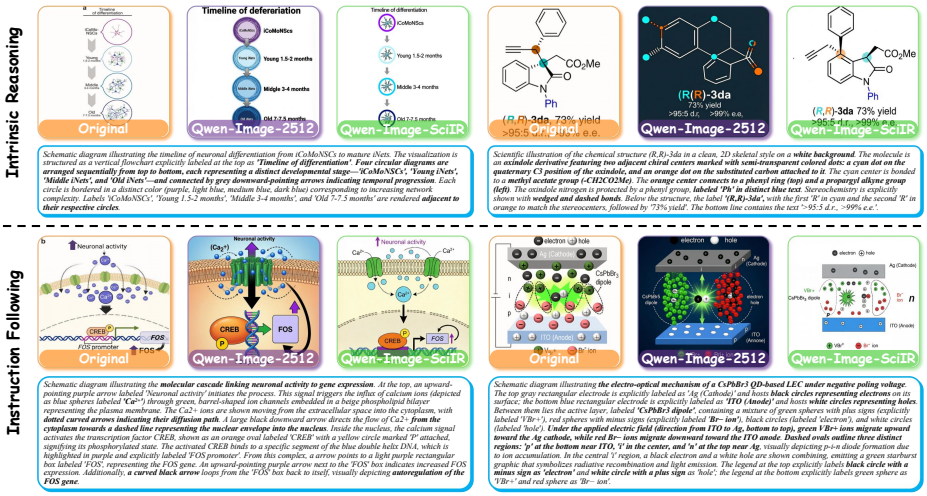

Fig. 5: Qualitative comparison of generated results.

Instruction Following *vs.* Intrinsic Reasoning. For the majority of models (*e.g.*, GPT-Image-1, Seedream 4.5), performance under explicit Sci-RCoT prompting (IF) significantly outpaces abstract prompting (IR). For instance, FLUX.1-Kontext-Max's accuracy drops from 36% to 13% without dense guidance. This confirms that while they excel at executing detailed instructions, they lack internalized scientific world models to autonomously derive constraints. However, a counter-intuitive trend emerges in some open-weights models (*e.g.*, Flux-Dev, Show-o2-7B), where IR outperforms IF. Rather than indicating superior reasoning, this highlights their deficiency in complex instruction adherence. Dense Sci-RCoT prompts overwhelm them, causing prompt overflow and attribute confusion. Thus, they paradoxically perform better on shorter abstract prompts by relying on superficial parametric memory.

AutoRegressive *vs.* Diffusion. Diffusion models currently maintain a distinct advantage over AutoRegressive (AR) architectures, with Qwen-Image-2512 (35%) establishing a clear 14% performance gap over the leading AR contender, Show-o2-7B (21%).However, despite this architectural disparity, both paradigms share a critical vulnerability: severe failure in the Text track. With even the top models scoring merely 15% (Diffusion) and 8% (AR) on text generation, the results confirm a shared fundamental limitation: whether utilizing continuous denoising or discrete next-token prediction, current open-source visual generation frameworks fundamentally lack the fine-grained typographic control necessary for accurate scientific illustrating.

Efficacy of Fine-tuning. To validate the effectiveness of our proposed pipeline, we conduct a direct comparison between the fine-tuned Qwen-Image-SciIR and its backbone, Qwen-Image-2512. Quantitatively, Tab. 3 demonstrates a substantial performance gain, elevating the Final Score from 35% to 43%. This

Table 4: Correlation between metrics and expert judgments. Our Atomic Checklist shows the strongest linear and rank alignment with human experts.

Metric	Pearson's $r \uparrow$	Kendall's $\tau \uparrow$	Spearman's $\rho \uparrow$
CLIPScore [12]	$0.345^{4\text{th}}$	$0.231^{4\text{th}}$	$0.315^{4\text{th}}$
VQAScore [22]	$0.457^{2\text{nd}}$	$0.342^{2\text{nd}}$	$0.410^{2\text{nd}}$
VIEScore [17]	$0.412^{3\text{rd}}$	$0.313^{3\text{rd}}$	$0.389^{3\text{rd}}$
Atomic Checklist (Ours)	$\mathbf{0.692}^{1\text{st}}$	$\mathbf{0.596}^{1\text{st}}$	$\mathbf{0.683}^{1\text{st}}$

improvement is particularly pronounced in the `Scientific Process` and `Entity Structure` tracks, with increases of 16% and 9% respectively, indicating a robust enhancement in modeling sequential process and topological integrity.

5.3 Qualitative Comparison

Qualitatively, visual comparisons in Fig. 5 reveal that Qwen-Image-SciIR shifts from a generic artistic style to a precise scientific illustration standard. We observe that the baseline Qwen-Image-2512 is prone to scientific hallucinations across three dimensions. The first is the **inability to correctly depict scientific processes**: it fails to visually depict the morphological progression of neuron development (top-left), relying solely on text proxies. The second is **morphological and structural omission**: *e.g.*, in the top-right, the baseline generates non-academic redundant backgrounds and hallucinates bizarre molecular topologies that violate chemical valence rules, along with the missing nucleus in the cell cross-section (bottom-left). The third is the **violation of domain priors**: as seen in the bottom-right, it incorrectly grounds the charge states to their respective micro-particles (electrons, holes) and ions (V_{Br}^{+} and Br^{-}). Conversely, our model effectively minimizes scientific hallucinations by explicitly integrating reasoning planning.

5.4 Correlation Analysis

We validated our automated protocol against human expert ratings on 200 randomly sampled test cases (50 per evaluation group). Three human annotators independently performed blind scoring on these samples, with final ratings determined by averaging their scores. As shown in Tab. 4, our *Atomic Checklist* achieves strong alignment with domain experts ($r = 0.692$), substantially outperforming the best baseline metric, VQAScore ($r = 0.457$). This discrepancy suggests that embedding-based metrics may capture only superficial semantic relevance, failing to penalize subtle scientific violations (*e.g.*, impossible topologies). In contrast, our taxonomy-grounded approach effectively detects these domain-specific hallucinations, confirming that high-fidelity scientific evaluation requires verifiable atomic constraints rather than holistic visual similarity. Implementation details are provided in Appendix G.

6 Conclusion

SciIR proposes a principled approach to scientific image reasoning that narrows the gap between general text-to-image capabilities and the strict constraints of natural science. We release SciIR-82k, containing more than 80k high-quality science image–text pairs with traceable Sci-RCoT reasoning chains—and SciIR-Bench, a fine-grained benchmark that breaks scientific correctness into verifiable atomic checks (topology, causality, conservation, *etc.*). Fine-tuning on SciIR-82k yields Qwen-Image-SciIR, which raises the SciIR-Bench score from 35% to 43% and shows the largest gains on entity structure and scientific process tracks, demonstrating that reasoning-dense training data measurably improves scientific consistency beyond perceptual quality alone.

Despite these advances, some limitations remain. SciIR-82k is biased toward published, standardized figures and underrepresents atypical or unconventional diagrams, while SciIR-Bench emphasizes scientific correctness over visual aesthetics. Future work should broaden domain and style coverage, add multimodal and cross-lingual annotations, and investigate hybrid training and evaluation approaches—such as symbolic constraints, weak supervision, and adversarial or counterfactual checks—to further enhance the models' ability for scientific reasoning.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China (No. 62406161). This work was completed during the internships of the authors Zhengfeng Shi, Yuning An, Peize Li, Jiabao Wei, Ruijie Li, and Junhao Xiao at the MAIR Lab, Huazhong University of Science and Technology.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Cai, Q., Chen, J., Chen, Y., Li, Y., Long, F., Pan, Y., Qiu, Z., Zhang, Y., Gao, F., Xu, P., et al.: Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. arXiv preprint arXiv:2505.22705 (2025)
3. 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)
4. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)

5. 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)
6. 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)
7. 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)
8. Esser, P., Kulal, S., Blattmann, A., Entezari, R., M"uller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Int. Conf. Mach. Learn. (2024)
9. 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)
10. Gao, P., Zhuo, L., Liu, D., Du, R., Luo, X., Qiu, L., Zhang, Y., Lin, C., Huang, R., Geng, S., et al.: Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. arXiv preprint arXiv:2405.05945 (2024)
11. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Adv. Neural Inform. Process. Syst. **36**, 52132–52152 (2023)
12. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 7514–7528 (2021)
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Adv. Neural Inform. Process. Syst. **30** (2017)
14. 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: Int. Conf. Comput. Vis. pp. 20406–20417 (2023)
15. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Adv. Neural Inform. Process. Syst. **36**, 78723–78747 (2023)
16. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLO. `https://github.com/ultralytics/ultralytics` (2024), software version 11.0.0. Accessed: 2025-12-21
17. Ku, M., Jiang, D., Wei, C., Yue, X., Chen, W.: Viescore: Towards explainable metrics for conditional image synthesis evaluation. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12268–12290 (2024)
18. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
19. Li, J., Chai, W., Fu, X., Xu, H., Xie, S.: Science-t2i: Addressing scientific illusions in image synthesis. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2734–2744 (2025)
20. Liao, J., Yang, Z., Li, L., Li, D., Lin, K., Cheng, Y., Wang, L.: Imagegen-cot: Enhancing text-to-image in-context learning with chain-of-thought reasoning. arXiv preprint arXiv:2503.19312 (2025)

21. Lin, H., Qin, C., Liu, Z., Pei, Q., Li, Y., Zhong, Z., Gao, X., Wang, Y., He, C., Wu, L.: Scientific image synthesis: Benchmarking, methodologies, and downstream utility. arXiv preprint arXiv:2601.17027 (2026)
22. Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation. In: Eur. Conf. Comput. Vis. pp. 366–384. Springer (2024)
23. Ma, Y., Liu, X., Chen, X., Liu, W., Wu, C., Wu, Z., Pan, Z., Xie, Z., Zhang, H., Yu, X., et al.: Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7739–7751 (2025)
24. Ma, Z., Zhang, Y., Jia, G., Zhao, L., Ma, Y., Ma, M., Liu, G., Zhang, K., Ding, N., Li, J., et al.: Efficient diffusion models: A comprehensive survey from principles to practices. IEEE Trans. Pattern Anal. Mach. Intell. (2025)
25. Ma, Z., Zhao, L., Qi, B., Zhou, B.: Neural residual diffusion models for deep scalable vision generation. Adv. Neural Inform. Process. Syst. **37**, 117456–117480 (2024)
26. Niu, Y., Ning, M., Zheng, M., Jin, W., Lin, B., Jin, P., Liao, J., Feng, C., Ning, K., Zhu, B., et al.: Wise: A world knowledge-informed semantic evaluation for text-to-image generation. arXiv preprint arXiv:2503.07265 (2025)
27. Onoe, Y., Rane, S., Berger, Z., Bitton, Y., Cho, J., Garg, R., Ku, A., Parekh, Z., Pont-Tuset, J., Tanzer, G., Wang, S., Baldridge, J.: DOCCI: Descriptions of Connected and Contrasting Images. In: Eur. Conf. Comput. Vis. (2024)
28. Pan, J., Ma, Z., Zhang, K., Ding, N., Zhou, B.: Self-reflective reinforcement learning for diffusion-based image reasoning generation. arXiv preprint arXiv:2505.22407 (2025)
29. Peirce, C.S.: Collected papers of charles sanders peirce, vol. 5. Harvard University Press (1934)
30. Qin, Q., Zhuo, L., Xin, Y., Du, R., Li, Z., Fu, B., Lu, Y., Yuan, J., Li, X., Liu, D., et al.: Lumina-image 2.0: A unified and efficient image generative framework. arXiv preprint arXiv:2503.21758 (2025)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. pp. 8748–8763. PmLR (2021)
32. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Adv. Neural Inform. Process. Syst. **29** (2016)
33. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Adv. Neural Inform. Process. Syst. **35**, 25278–25294 (2022)
34. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: a dataset for image captioning with reading comprehension. In: Eur. Conf. Comput. Vis. pp. 742–758. Springer (2020)
35. Singla, V., Yue, K., Paul, S., Shirkavand, R., Jayawardhana, M., Ganjdanesh, A., Huang, H., Bhatele, A., Somepalli, G., Goldstein, T.: From Pixels to Prose: A Large Dataset of Dense Image Captions. arXiv (2024)
36. 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)
37. Sun, K., Pan, J., Ge, Y., Li, H., Duan, H., Wu, X., Zhang, R., Zhou, A., Qin, Z., Wang, Y., et al.: Journeydb: A benchmark for generative image understanding. Adv. Neural Inform. Process. Syst. **36**, 49659–49678 (2023)

38. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024)
39. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
40. 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)
41. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
42. Xie, E., Chen, J., Zhao, Y., Yu, J., Zhu, L., Wu, C., Lin, Y., Zhang, Z., Li, M., Chen, J., et al.: Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. arXiv preprint arXiv:2501.18427 (2025)
43. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528 (2024)
44. 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)
45. Yang, L., Liu, J., Hong, S., Zhang, Z., Huang, Z., Cai, Z., Zhang, W., Cui, B.: Improving diffusion-based image synthesis with context prediction. Adv. Neural Inform. Process. Syst. **36**, 37636–37656 (2023)
46. Ye, J., Jiang, D., Wang, Z., Zhu, L., Hu, Z., Huang, Z., He, J., Yan, Z., Yu, J., Li, H., et al.: Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. arXiv preprint arXiv:2508.09987 (2025)
47. 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)
48. Zhang, Q., Dai, X., Yang, N., An, X., Feng, Z., Ren, X.: Var-clip: Text-to-image generator with visual auto-regressive modeling. arXiv preprint arXiv:2408.01181 (2024)
49. Zhu, D., Meng, R., Song, Y., Wei, X., Li, S., Pfister, T., Yoon, J.: Paperbanana: Automating academic illustration for ai scientists. arXiv preprint arXiv:2601.23265 (2026)
50. Zhu, M., Lin, Z., Weng, Y., Lu, P., Xie, Q., Wei, Y., Liu, S., Sun, Q., Zhang, Y.: Autofigure: Generating and refining publication-ready scientific illustrations. In: Int. Conf. Learn. Represent. (2026)