SCIDIAGBENCH: A TYPE-SPECIFIC BENCHMARK FOR HUMAN-ALIGNED EVALUATION OF SCIENTIFIC DIAGRAMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent agents have demonstrated strong capabilities in generating images from textual prompts, motivating their utilization for scientific diagram generation. Scientific diagrams naturally span diverse types, each characterized by distinct visual structures and content organization. Consequently, during peer review and other scientific evaluation, human experts naturally consider type-specific criteria when assessing their quality. However, existing VLM-based evaluation benchmarks fail to reflect the above human evaluation practices for two main reasons: 1) benchmark datasets typically lack a systematic categorization of different diagram types and thus adopt the same generic criteria for all diagrams, and 2) their evaluation criteria are often designed heuristically, without explicit alignment to human evaluation preferences. Motivated by these limitations, we introduce SCIDIAGBENCH, which consists of a benchmark dataset that systematically categorizes scientific diagrams into 11 types and a human preference-guided evaluation method. Specifically, SCIDIAGBENCH contains 165 scientific diagrams across 10 broad domains, with types including graphical abstracts, technical roadmaps, and illustrative cases. Based on this dataset, SCIDIAGBENCH develops an evaluation method that leverages a preference-guided calibration process to incorporate human preferences into the generation of type-specific criteria. Ultimately, SCIDIAGBENCH enables more reliable automated evaluation that better aligns with human evaluation across diverse scientific diagrams.

1 INTRODUCTION

Scientific diagrams, as an important complement to scientific texts, play a crucial role in communicating complex technical concepts through visual representations (Zhu et al., 2026b). Building on the advances in multimodal generative models (e.g., Gemini 3 Pro Image (Google, 2025) and GPT-Image-2 (OpenAI, 2026b)), recent studies have developed scientific diagram generation agents, such as PaperBanana (Zhu et al., 2026a) and AutoFigure (Zhu et al., 2026b), for automated scientific diagram generation. These agents enable the direct processing of long scientific manuscripts to extract essential information and synthesize target diagrams by coordinating multiple stages of diagram creation.

As scientific diagram generation agents become increasingly capable, a critical question is how to reliably evaluate the quality of generated diagrams in practical scientific settings, such as peer review. Compared with general images, scientific diagrams are particularly challenging to evaluate, as they span diverse types designed for different communicative purposes. Consequently, human experts naturally apply different evaluation criteria when assessing different types of scientific diagrams, making type-specific evaluation essential. For example, while technical roadmaps place greater emphasis on structural criteria to ensure protocol completeness and logical correctness, illustrative cases prioritize content criteria to faithfully represent data instances and execution traces.

On the dataset side, rather than systematically organizing diagrams according to their communicative purposes and visual structures, existing datasets often define their categories based on disciplinary domains or specific generation tasks. For example, benchmarks centered on artificial intelligence primarily cover diagrams such as methodological frameworks and flowcharts (Mondal et al.,

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

2024; Zhu et al., 2026a), while multidisciplinary benchmarks typically organize their samples by scientific domains rather than by diagram types (Chang et al., 2025; Luo et al., 2025). As a result, they provide limited support for evaluation that accounts for the diverse quality requirements of different scientific diagrams and reflects human evaluation practices.

On the evaluation side, manual assessment of scientific diagrams is often labor-intensive and time-consuming, motivating recent efforts toward automated evaluation. Current evaluation methods generally fall into two main paradigms. The first paradigm derives criteria from individual reference diagrams. For instance, some studies evaluate generated diagrams by matching extracted structural information (Luo et al., 2025; Chang et al., 2026) or by applying tailored checklists (Li et al., 2026a; Zhao et al., 2026; Li et al., 2026b; Sun et al., 2026; Pang et al., 2025) specific to each reference. While offering fine-grained assessments, these diagram-specific methods are difficult to generalize across instances and scale efficiently. The second paradigm, in contrast, applies a shared set of criteria across all diagrams, typically constructed from predefined dimensions and heuristic rules (Zhu et al., 2026a;b; Mondal et al., 2024; Pang et al., 2025). Although this approach facilitates large-scale evaluation, these overly generic criteria lack explicit alignment with the evaluation principles and preferences of human experts.

To address these limitations, we introduce SCIDIAGBENCH, a comprehensive benchmark designed for automated evaluation of scientific diagram generation. Specifically, we collect scientific diagrams from diverse scholarly sources and construct a systematic taxonomy based on their visual structures, content organization, and communicative purposes, yielding 165 diagrams across 11 types and 10 domains. Building on this dataset, we develop an evaluation framework that generates type-specific criteria across multiple quality dimensions, refines the criteria with scoring instructions derived from human preferences, and adopts reference-based scoring for stable and comparable evaluation. Utilizing SCIDIAGBENCH, we systematically evaluate four representative agents (Vanilla, PaperBanana, AutoFigure, and Paper2Figure). Extensive analysis reveals a highly uneven capability profile: while current agents excel at organizing abstract conceptual relationships, they exhibit systematic deficits in generating complex topologies and precise concrete visual elements, highlighting critical bottlenecks in existing scientific diagram generation agents.

Our primary contributions are summarized as follows:

- We construct the first type-specific benchmark dataset for scientific diagram generation, with 11 diagram types and 10 scientific domains.
- We develop a human preference-guided evaluation framework that generates type-specific criteria and further refines them with scoring instructions derived from human preferences.
- We conduct extensive experiments on representative agents using SCIDIAGBENCH, revealing substantial variations across agents, diagram types, and judge models. Human evaluation further demonstrates that the proposed automated evaluation framework achieves better alignment with human judgments.

2 RELATED WORK

Scientific Diagram Generation Scientific diagram generation methods broadly fall into two categories: code-based generation and image-based generation. Code-based approaches formulate diagram creation as a text-to-code task, typically using languages such as TikZ (Belouadi et al., 2025) or Python (Ge et al., 2025; Zheng et al., 2025) to produce structured and editable representations. Recent systems further incorporate iterative reasoning and refinement, as in SciDoc2Diagrammer-MAF (Mondal et al., 2024) and SciFig (Huang et al., 2026). However, code-based approaches are often limited in visual expressiveness when handling complex scientific content (Zhu et al., 2026a). In contrast, image-based approaches directly synthesize diagrams with multimodal generative models (e.g., Gemini 3 Pro Image (Google, 2025) and GPT-Image-2 (OpenAI, 2026b)). Building on increasingly capable image generators, recent agents such as AutoFigure (Zhu et al., 2026b), PaperBanana (Zhu et al., 2026a), and Paper2Figure (the scientific-figure-generation module of Paper2Any) (Liu et al., 2025b) introduce explicit content planning and intermediate representations to improve generation quality. Despite these rapid advancements, the systematic assessment of agent outputs remains challenging.

Evaluation of Generated Diagrams Conventional metrics such as CLIPScore, FID, and CMMD measure image–text alignment or distributional similarity and are insensitive to errors in diagram entities and their logical relations (Hessel et al., 2021; Jayasumana et al., 2024). Recent studies therefore adopt more structured evaluation protocols, often supported by Vision-Language Models (VLMs). One line of work develops prompt-based benchmarks that assess images generated from carefully constructed prompts. T2I-CoreBench, KVBench, and BizGenEval associate each prompt with atomic checklists (Li et al., 2026a; Zhao et al., 2026; Li et al., 2026b). T2I-ReasonBench assesses reasoning and image quality (Sun et al., 2026). To evaluate semantics more precisely, FEPBench employs atomic decomposition (Chang et al., 2026), while MMMG uses knowledge-graph edit distance (Luo et al., 2025). Another line of work develops paper-based benchmarks to evaluate the end-to-end generation from scientific manuscripts. DiagramEval matches nodes and paths in graph representations of Scalable Vector Graphics (SVG) diagrams (Liang & You, 2025); SridBench, FigureBench, and PaperBananaBench use source contexts and reference diagrams for shared multidimensional scoring (Chang et al., 2025; Zhu et al., 2026b;a); SciFig adds paper-specific questions to common criteria (Huang et al., 2026). Overall, existing benchmarks apply uniform criteria that overlook the heterogeneous requirements of different diagram types.

3 THE SCIDIAGBENCH BENCHMARK

In general, this section is organized into three main parts: formal task definition, dataset construction pipeline, and design of a customized evaluation framework (an overview of which is illustrated in Figure 1).

3.1 TASK DEFINITION

Following PaperBanana (Zhu et al., 2026a), we formulate the task as the end-to-end generation of a scientific diagram by an agent. Given a scientific paper P, a target diagram caption C, and a target aspect ratio R, an agent $\mathcal{A}$ directly generates the final diagram $\hat{D}$:

$$\hat{D} = \mathcal{A}(P, C, R). \tag{1}$$

The task does not prescribe any intermediate representation or generation procedure; instead, it evaluates the end-to-end capability of $\mathcal{A}$ to extract relevant information from P that falls within the scope specified by C and generate a clear, coherent, and publication-ready diagram under the specified aspect ratio constraint R.

3.2 DATASET CONSTRUCTION

Diagram Collection. To provide a foundation for defining the task scope and developing the diagram taxonomy, it is essential to ensure the diversity of visual representations. We initially collect 364 papers from diverse sources such as interdisciplinary and field-specific scholarly journals (e.g., *The Innovation* and *Reviews of Modern Physics*), major international conferences (e.g., ICIS), and the open-access arXiv preprint repository. A total of 598 figures are sampled from these papers. From this initial set, we observe that certain categories are not well suited for direct generation by agents. Specifically, we exclude two broad classes of figures based on their visual representations. Figures in the first class rely on rigid quantitative mappings or outputs produced by specialized tools, including *plots*, *tables*, *data-driven maps*, *diagrams illustrating mathematical and physical principles*, *chemical structure and reaction diagrams*, and *mechanical CAD diagrams*. Figures in the second class require direct sensor-based acquisition, primarily including *photos* (Dhote et al., 2023) and composite figures containing real-world images. Collectively, these excluded categories are inherently better suited for generation via deterministic visualization tools, specialized software, or physical sensors. Furthermore, to mitigate potential training-data contamination, we manually inspect the candidate diagrams and filter out those whose preliminary generation output reproduces the reference diagram with anomalously high similarity. Further details on the excluded categories and the contamination filtering process are provided in Appendix A. After applying filtering, we retain a total of 291 candidate diagrams for subsequent stages.

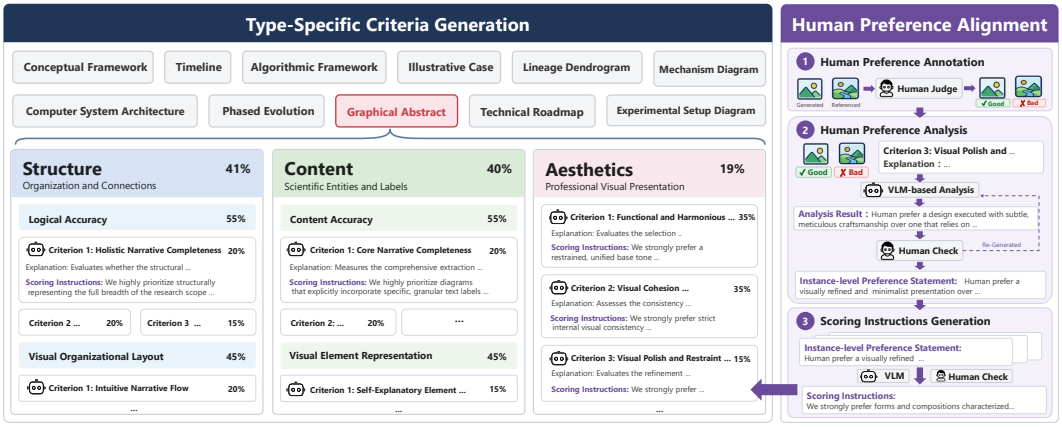


Figure 1: Overview of the dynamic hierarchical evaluation framework, combining type-specific criteria generation with human preference alignment.

Taxonomy Construction. Existing taxonomies of scientific figures, such as DocFigure and ACL-FIG-PILOT, are primarily designed for downstream tasks including figure retrieval and classification. Consequently, they typically categorize figures based on broad graphic modalities (e.g., photographs and tables) or quantitative charting formats (e.g., bar charts and line graphs) (Dhote et al., 2023; Jobin et al., 2019; Karishma et al., 2023). To comprehensively evaluate scientific diagram generation, we construct a fine-grained taxonomy based on the diverse visual structures, content organization, and communicative purposes of target diagrams. By analyzing the remaining 291 heterogeneous scientific diagrams, we identify recurring combinations of these characteristics and categorize them into 11 types. Specifically, some diagrams are organized along temporal or evolutionary axes: *Timeline* (TL), *Lineage Dendrogram* (LD), and *Phased Evolution* (PE) distinguish chronological sequences of milestones, branching lineage relationships, and transitions among discrete system states, respectively. Other types emphasize overall structural frameworks: *Algorithmic Framework* (AF), *Computer System Architecture* (CSA), and *Conceptual Framework* (CF) illustrate computational modules and their execution dependencies, layered hardware–software components and interfaces, and abstract constructs and their relations, respectively. The remaining types are defined by specific communicative objectives: *Technical Roadmap* (TR) for staged methodological procedures and decision paths, *Graphical Abstract* (GA) for paper-level summaries, *Experimental Setup Diagram* (ESD) for hardware and instrumentation setups, *Illustrative Case* (IC) for explanatory examples, and *Mechanism Diagram* (MD) for causal interactions and state transformations. Detailed definitions and characteristics of each type are provided in Appendix B.

Dataset Partitioning. To facilitate human preference alignment and support comprehensive evaluation and analyses, we partition the scientific diagrams into a development set and a test set. Given the pronounced category imbalance among the remaining 291 scientific diagrams, we first construct a balanced subset by sampling 15 diagrams for each type. Then, these diagrams are randomly partitioned into a development set (5 per type) and a test set (10 per type), with strict controls to ensure no overlap in source papers. Each finalized instance is represented as a quadruple (P, C, R, D), consisting of a scientific paper, a caption, an aspect ratio, and a reference diagram. For the scientific paper P, we parse each source PDF using MinerU (Niu et al., 2025), retaining the main body text while removing author metadata, references, and other irrelevant materials. The target aspect ratio R is derived by mapping the width-to-height ratio of the reference diagram D to the closest standard aspect ratio (e.g., 3:2).

3.3 Evaluation Framework

Type-Specific Criteria Generation. Unconstrained criteria generation may produce criteria that target irrelevant aspects rather than essential qualities (Du et al., 2026). We guide this process using three complementary top-level dimensions. *Structure* concerns the spatial organization of a diagram; *Content* concerns the domain-specific information conveyed by its constituent elements; and

Aesthetics concerns graphic design quality. However, the top-level dimensions alone do not distinguish perceptual execution from semantic validity. A diagram may exhibit a clear layout or visually recognizable elements while containing incorrect dependencies or factual statements. Since visual presentation is salient while logical or factual errors are less immediately perceptible, a VLM judge may conflate the two, leading to overestimated overall scores. We therefore decompose Structure into *Visual Organizational Layout* and *Logical Accuracy*, and Content into *Visual Element Representation* and *Content Accuracy*, while retaining Aesthetics as a single dimension (detailed in Appendix C). Conditioned on these dimensions and the target diagram type τ, we leverage Gemini 3.1 Pro to dynamically generate tailored criteria $\mathcal{C}_{d,\tau}$ for dimension d under diagram type τ, alongside two-level normalized weights comprising intra-dimension weights w_c and top-level dimension weights $W_{d,\tau}$. Notably, criteria for Structure and Content are generated directly under their respective sub-dimensions, but their intra-dimension weights w_c are normalized within the corresponding top-level dimension. Dimension-level weights $W_{d,\tau}$ are averaged over 5 independent trials to improve stability. Each generated criterion is manually reviewed for quality and relevance.

Human Preference Alignment. Preliminary experiments suggest that relying solely on the generated criteria leads to inconsistent scoring behavior in certain diagram types. This indicates that the judge VLM requires more specific guidance, particularly on how human experts interpret and apply the criteria during scoring. Therefore, we align the generated criteria with human preferences using the development set, so that the resulting criteria can better capture how human experts apply them in practice. By comparing reference diagrams against baseline generations (detailed in Section 4) under each criterion, we elicit human preferences and employ Gemini 3.1 Pro to distill them into unified preference statements. These statements are then appended to their corresponding criteria as additional scoring instructions.

Reference-Based Scoring. Since absolute scoring can be subjective and prone to scoring drift, we instead adopt a reference-based scoring strategy by providing the original human-created diagram D as a reference anchor. To mitigate potential bias toward the reference diagram, the generated and reference diagrams are anonymized and presented in a randomized order alongside the type-specific criteria $\mathcal{C}_\tau$:

$$\Big(\{s_{\hat{D},c}\}_{c\in\mathcal{C}_\tau}, \{s_{D,c}\}_{c\in\mathcal{C}_\tau}\Big) = \mathrm{JudgeVLM}\Big(P, C, D, \hat{D}, \mathcal{C}_\tau\Big). \tag{2}$$

Based on the two-level weights generated in the initial stage, the aggregated absolute score $S(\cdot)$ (where $\cdot$ represents either $\hat{D}$ or D) is computed by aggregating its corresponding criterion-level scores $s_{\cdot,c}$ (i.e., either $s_{\hat{D},c}$ or $s_{D,c}$) as:

$$S(\cdot) = \sum_{d\in\mathcal{D}} W_{d,\tau} \sum_{c\in\mathcal{C}_{d,\tau}} w_c \cdot s_{\cdot,c}, \tag{3}$$

where $\mathcal{D} = \{\text{Structure}, \text{Content}, \text{Aesthetics}\}$. Preliminary experiments reveal that the absolute score of the reference diagram fluctuates depending on the quality of the generated diagram it is paired with. To provide a stable evaluation anchor and ensure comparability, the final score of the generated diagram is computed as a relative ratio:

$$S_{\text{final}}(\hat{D}) = \frac{S(\hat{D})}{S(\hat{D}) + S(D)}. \tag{4}$$

Accordingly, $S_{\text{final}}(\hat{D}) = 0.5$ indicates parity with the reference diagram, while values below and above 0.5 indicate inferior and superior performance relative to the reference, respectively.

4 Experiments

4.1 Experimental Setup

Baselines and Evaluation Protocol. We evaluate four representative scientific diagram generation agents on SciDiagBench: Vanilla, PaperBanana (Zhu et al., 2026a), AutoFigure (Zhu et al., 2026b), and Paper2Figure (Liu et al., 2025b). Vanilla is a simple two-stage baseline in which a

Table 1: Results across different diagram types on SCIDIAGBENCH.

Agent	TL	LD	PE	AF	CSA	CF	TR	GA	ESD	IC	MD	Overall
Vanilla	0.4406	0.4168	0.5626	0.4670	**0.4947**	**0.5820**	0.5094	**0.6228**	**0.5258**	0.3740	0.4566	0.4957
PaperBanana	**0.5272**	**0.4917**	0.5975	**0.5665**	0.4926	0.5703	**0.5361**	0.6195	0.4610	0.4824	**0.4985**	**0.5312**
Paper2Figure	0.4714	0.4514	**0.6060**	0.4314	0.3337	0.4594	0.4057	0.6034	0.4555	**0.4865**	0.3241	0.4571
AutoFigure	0.2172	0.1162	0.1189	0.2132	0.2222	0.1919	0.2905	0.2956	0.1604	0.1476	0.2137	0.1989

VLM first converts the parsed paper into a generation prompt, and then a multimodal generative model generates the diagram. PaperBanana, AutoFigure, and Paper2Figure employ more complex multi-stage pipelines with explicit planning. More details of these agents are introduced in Section 2. Despite their different pipeline designs, all these agents rely on two major backbone components: a VLM to understand the paper and a multimodal generative model to generate the diagram. Unless otherwise specified, we use GPT-5.4 (OpenAI, 2026a) as the default VLM and Gemini 3 Pro Image (Google, 2025) as the default multimodal generative model for all agents. We adopt Gemini 3.1 Pro (Google, 2026) as the default judge VLM. The temperature of all models is set to 0 to minimize stochastic variation.

Human Preference Annotation and Evaluation. To assess whether the proposed evaluation framework aligns with human preferences, we conduct human preference annotations based on the test set. For each instance, we generate a diagram using the Vanilla agent and pair it with the corresponding reference diagram for human comparison, resulting in a human preference dataset termed SCIDIAGBENCH-VanillaPref. Three annotators, all Ph.D. students, independently assess these pairs using 5-point preference labels: strong reference-preferred, slight reference-preferred, unable to judge, slight generated-preferred, and strong generated-preferred. To resolve disagreements among annotators, we further design a set of aggregation rules to derive the final preference labels. After aggregation, we retain 105 high-agreement pairs, comprising 24 strong reference-preferred, 46 slight reference-preferred, 29 slight generated-preferred, and 6 strong generated-preferred pairs. The remaining 5 pairs are excluded from subsequent evaluation due to conflicting directional judgments among annotators. The annotation guidelines and aggregation rules are provided in Appendix E.

We evaluate human preference alignment from two perspectives: *directional agreement* and *strength alignment*. For directional agreement, we use Directional Agreement Rate (DAR), which measures whether the preference direction indicated by the automatic relative score is consistent with the human judgment, using 0.5 as the direction threshold. For strength alignment, we measure the correspondence between automatic evaluation scores and human preference strength using Kendall's τ_b (Kendall, 1938) and Spearman's ρ (Spearman, 1904). Kendall's τ_b evaluates ordinal concordance while explicitly accounting for ties, whereas Spearman's ρ measures the monotonic association between the two rankings.

4.2 MAIN RESULTS

Results across Diagram Types. To examine the overall performance of different scientific diagram generation agents across diagram types, we report the relative score of each agent on every diagram type as well as its overall score across all types. As shown in Table 1, although PaperBanana achieves the highest overall score (0.5312), the relative performance of these agents varies considerably across diagram types, and no single pipeline design proves uniformly optimal. Generally, current agents perform relatively well on diagram types that emphasize high-level information organization and abstract conceptual relationships. For instance, top-performing agents even surpass the human-parity threshold (0.5) on types such as Graphical Abstracts and Conceptual Frameworks. By contrast, for types such as Illustrative Case and Lineage Dendrogram, which demand concrete and precise structures, current agents still fall short.

Results across Evaluation Dimensions. To examine the performance of different agents across evaluation dimensions, we report their scores on Structure, Content, and Aesthetics, together with the corresponding sub-dimensions, in Table 2. As shown in the table, PaperBanana achieves the highest scores across all dimensions and is the only agent that consistently surpasses the human-parity threshold. Comparing top-level dimensions, the Content scores are consistently lower than

Table 2: Results across different evaluation dimensions on SCIDIAGBENCH.

Method	Structure			Content			Aesthetics
	VOL	LA	Weighted	VER	CA	Weighted	
Vanilla	0.5083	0.4756	0.4909	0.5057	0.4650	0.4829	0.5393
PaperBanana	**0.5412**	**0.5259**	**0.5326**	**0.5318**	**0.5158**	**0.5224**	**0.5544**
Paper2Figure	0.4881	0.4092	0.4495	0.4916	0.3991	0.4433	0.5108
AutoFigure	0.2318	0.1892	0.2106	0.1549	0.1943	0.1808	0.2020

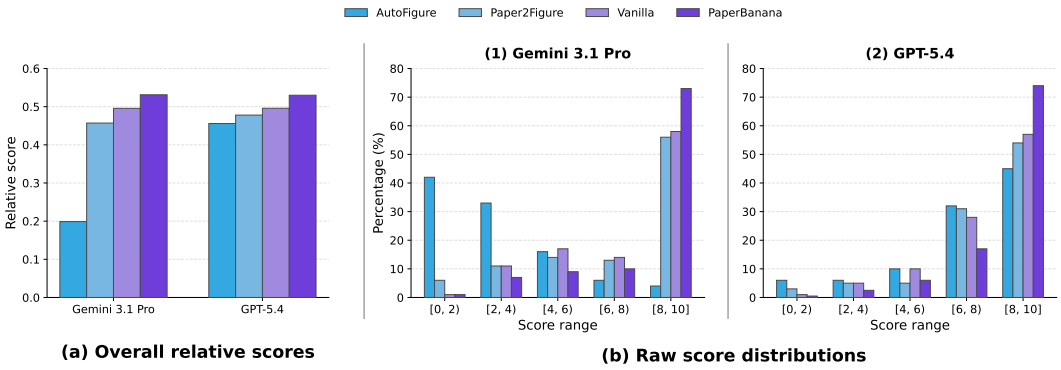


Figure 2: Sensitivity of evaluation results to different judge models. (a) Overall relative scores of different generation agents. (b) Raw score distributions produced by Gemini 3.1 Pro and GPT-5.4.

the Structure scores across all agents, indicating that accurately representing scientific content remains more challenging than organizing the overall diagram structure for existing agents. Moreover, for Vanilla, PaperBanana, and Paper2Figure, Aesthetics consistently receives the highest top-level score, suggesting that current agents are relatively effective at producing visually appealing diagrams. A similar pattern can be observed within the Structure and Content dimensions. For Vanilla, PaperBanana, and Paper2Figure, VOL outperforms LA, while VER outperforms CA. This suggests that current agents are generally stronger at visually organizing and presenting information than at preserving rigorous logical and semantic accuracy.

AutoFigure Failure Analysis. As AutoFigure exhibits substantially lower performance, we further analyze its generated diagrams to identify the underlying failure patterns. As shown in Tables 1 and 2, its performance is particularly low on LD (0.1162) and PE (0.1189), as well as on the VER dimension (0.1549), suggesting difficulties in handling complex topological structures, continuous transformations, and faithful visual-element representation. Further analysis of the generated diagrams reveals several failure patterns, including replacing visual elements with lengthy text blocks, using overly generic geometric primitives or literal placeholders, and producing unclear connections between diagram components. Representative cases are provided in Appendix F. Such difficulties may partly reflect the challenge of encoding complex visual structures through its SVG-based intermediate representation.

4.3 SENSITIVITY ANALYSIS OF JUDGE MODELS

Since the capability and scoring behavior of the judge model can substantially affect automatic scientific diagram evaluation, we first examine the sensitivity of evaluation results to different judge models. Specifically, in addition to the default Gemini 3.1 Pro, we use GPT-5.4 as an alternative judge.

Main Results with Different Judge Models. To examine the effect of the judge model on automatic evaluation, we re-evaluate the outputs of all generation agents on the test set using GPT-5.4. As shown in Figure 2(a), although the two judges preserve the same ordering among generation agents, the score gaps among them are substantially reduced under GPT-5.4. Among the four agents,

Table 3: Human preference alignment evaluation on SCIDIAGBENCH-VanillaPref. Ref. and Gen. denote reference-preferred and generated-preferred, respectively. Δ denotes the judge score minus 0.5, which indicates the direction and strength of the judge preference.

Human stratum	GPT	Δ GPT	Gemini	Δ Gemini	GPT DAR	Gemini DAR
Strong Ref.	0.4690	-0.0310	0.3806	-0.1194	0.5417	0.8750
Slight Ref.	0.4882	-0.0118	0.4357	-0.0643	0.4783	0.7826
Slight Gen.	0.5210	+0.0210	0.6249	+0.1249	0.6552	0.9310
Strong Gen.	0.5511	+0.0511	0.6838	+0.1838	1.0000	1.0000

AutoFigure shows the largest score increase under GPT-5.4, while the scores of Vanilla and PaperBanana remain relatively stable, and Paper2Figure shows a little increase.

Figure 2(b) further compares the raw-score distributions produced by the two judges. GPT-5.4 assigns substantially fewer scores in the low-score ranges and concentrates more scores in the high-score ranges, while Gemini 3.1 Pro produces more differentiated distributions across different agents. For example, when evaluating AutoFigure, 41.7% of the scores assigned by Gemini 3.1 Pro fall within the $[0, 2)$ interval, compared with only 5.9% under GPT-5.4.

To further analyze this discrepancy, we check the diagrams generated by AutoFigure. Among the 110 generated diagrams, 98 contain severe defects, such as missing glyphs and overlapping text. However, GPT-5.4 fails to identify these defects in 52 cases, including 39 cases in which it positively rationalizes the defects and assigns high scores. In the remaining 46 cases, although the defects are explicitly identified, GPT-5.4 does not penalize them as strongly as prescribed by the rubric.

Evaluation of Human Preference Alignment across Judge Models. Given the substantial discrepancies between the two judge models, we further assess how well their evaluation results align with human preferences on SCIDIAGBENCH-VanillaPref. This analysis also allows us to examine whether the different scoring behaviors persist on Vanilla outputs, which contain fewer severe defects than AutoFigure outputs. As shown in Table 3, the judge scores given by GPT-5.4 exhibit substantially smaller deviations from the 0.5 parity threshold (ranging from -0.031 to $+0.051$) compared to those given by Gemini 3.1 Pro (ranging from -0.119 to $+0.184$) across the four human-preference strata, indicating a strong tendency of GPT-5.4 to compress relative scores toward 0.5. This compression makes its evaluations less consistent with human preference strength. Under the complete evaluation framework (Tables 4 and 9), Gemini 3.1 Pro shows substantially higher correlations with human judgments than GPT-5.4, with Kendall's τ_b of 0.594 vs. 0.228 and Spearman's ρ of 0.733 vs. 0.293. GPT-5.4 also exhibits a notably lower DAR with human judgments, particularly on reference-preferred instances (Table 3).

4.4 ABLATION STUDY

To investigate how Type-Specific Criteria (TSC) and Human Preference Alignment (HPA) affect the evaluation results, we compare four configurations: ours, w.o. HPA, w.o. TSC, and w.o. HPA&TSC. w.o. TSC replaces the type-specific criteria with a unified set of general criteria shared across all diagram types, while w.o. HPA removes the preference statements from the evaluation criteria while retaining the type-specific criteria. w.o. HPA&TSC removes both components. The results on SCIDIAGBENCH-VanillaPref are in Table 4. We use the Directional Agreement Rate (DAR with 0.5 as the decision threshold) as the primary metric, together with correlation metrics such as Kendall's τ_b to measure agreement with human preference annotation.

The ablation results demonstrate that both TSC and HPA contribute to human-aligned evaluation, while their combination yields the largest improvement. Removing HPA decreases the overall DAR from 0.857 to 0.762 and the reference-preferred DAR from 0.814 to 0.671, whereas removing TSC leads to a larger drop to 0.705 and 0.586, respectively. Notably, HPA alone provides only marginal gains over the configuration without either component, while its effect becomes much stronger when combined with TSC. This suggests that TSC provides a more informative type-specific evaluation basis, upon which HPA can further calibrate the evaluation toward human preferences.

Table 4: Ablation results for preference alignment evaluation on SCIDIAGBENCH-VanillaPref.

Experimental setup		Directional Agreement Rate			Strength Alignment	
TSC	HPA	Overall	Gen.-pref.	Ref.-pref.	Kendall's τ_b	Spearman's ρ
–	–	0.6952	0.9143	0.5857	0.4787	0.5979
–	✓	0.7048	0.9429	0.5857	0.2085	0.2746
✓	–	0.7619	0.9429	0.6714	0.4974	0.6205
✓	✓	**0.8571**	0.9429	**0.8143**	**0.5943**	**0.7329**

Table 5: Stability over three independent scoring runs. Between var. and within var. denote between-instance and within-instance variance.

Experimental setup		Stability metrics					
TSC	HPA	ICC	Between var.	Within var.	Direction flip	Criterion flip	Criterion entropy
–	–	0.8890	0.0119	0.00149	0.1333	0.1958	0.1727
–	✓	0.8718	0.0113	0.00166	0.1212	0.1939	0.1708
✓	–	0.8613	0.0141	0.00228	0.1333	0.1883	0.1664
✓	✓	0.8714	0.0177	0.00261	0.1212	0.2030	0.1797

4.5 STABILITY ANALYSIS

To assess the stability of our evaluation framework, we perform two additional scoring runs on the test set, yielding three independent runs in total. We compare four configurations: ours, w.o. HPA, w.o. TSC, and w.o. HPA&TSC. Across the four compared configurations, we evaluate stability using three groups of metrics: 1) consistency metrics, including Intraclass Correlation Coefficient (ICC), between-instance, and within-instance variance; 2) categorical stability, measured by direction flip rate and criterion flip rate; and 3) criterion entropy, which quantifies the dispersion of criterion-level scores across runs. For flip-rate metrics, we discretize evaluation outcomes into three directions (generated-preferred, tied, and reference-preferred), using a threshold of 0 for criterion-level score differences and 0.5 for instance-level relative scores. The criterion flip rate is computed as the unweighted average over individual criteria, whereas the direction flip rate is computed from the instance-level relative scores defined in Eq. (4).

As shown in Table 5, all four configurations exhibit relatively high reproducibility, with ICC values ranging from 0.8613 to 0.8890 and within-instance variance remaining substantially smaller than between-instance variance. The direction flip rates are also relatively similar across configurations, ranging from 0.1212 to 0.1333. Notably, no single configuration consistently achieves the best value across all stability metrics. For example, the configuration without TSC or HPA achieves the highest ICC (0.8890), whereas configurations with HPA obtain a slightly lower direction flip rate (0.1212). These results suggest that the evaluation framework is generally reproducible under repeated scoring, while the effects of individual components on aggregate stability remain relatively small. A more detailed type-level stability analysis is provided in Appendix J.

5 CONCLUSION

In this work, we introduce SCIDIAGBENCH, a type-specific benchmark for human-aligned evaluation of scientific diagram generation. SCIDIAGBENCH provides a systematically categorized dataset of 165 diagrams spanning 11 diagram types and 10 scientific domains, together with an evaluation framework that combines type-specific criteria with preference-guided calibration to better reflect human judgments. Extensive experiments on representative agents demonstrate the importance of accounting for diagram-type differences and human evaluation preferences.

AI USE STATEMENT

We used generative AI tools to aid and polish manuscript writing, including improving grammar, wording, and readability. The authors reviewed and revised the AI-assisted text for accuracy and consistency with the reported methods and results.

We also used generative models to produce synthetic scientific diagrams for the experiments, including the generated-diagram component of SCIDIAGBENCH-VanillaPref, as described in Section 4.1. These synthetic outputs are distinct from the original human-created reference diagrams collected from scholarly sources. The preference labels used to evaluate human alignment were provided by human annotators, not generated by AI (Appendix E).

As described in Section 3.3 and illustrated in Figure 1, AI was also used within the evaluation framework. Specifically, Gemini 3.1 Pro generated type-specific evaluation criteria and associated weights, and distilled human preferences elicited on the development set into unified preference statements for criterion-level injection during Human Preference Alignment. The authors carefully reviewed the generated criteria for quality and relevance and the distilled preference statements for fidelity to the underlying human judgments before incorporating them into the scoring instructions. Thus, AI assisted in formulating the scoring guidance rather than replacing the human judgments on which preference alignment was based. VLM judges subsequently performed reference-based scoring, with the models and evaluation procedures documented in the paper. We have reviewed the AI-assisted work and take full responsibility for the final content of this paper, including its text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Alibaba Cloud. Qwen3.5-Plus. https://www.alibabacloud.com/help/en/model-studio/text-generation-model/, 2026a. Accessed: 2026-08-19.

Alibaba Cloud. Qwen-Image-2.0-Pro. https://www.alibabacloud.com/help/en/model-studio/qwen-image-api, 2026b. Accessed: 2026-08-19.

O. Anton et al. INTENTAS – an entanglement-enhanced atomic sensor for microgravity. *EPJ Quantum Technology*, 12:26, 2025. doi: 10.1140/epjqt/s40507-025-00330-9. URL https://doi.org/10.1140/epjqt/s40507-025-00330-9.

Jonas Belouadi, Eddy Ilg, Margret Keuper, Hideki Tanaka, Masao Utiyama, Raj Dabre, Steffen Eger, and Simone Paolo Ponzetto. TikZero: Zero-shot text-guided graphics program synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17793–17806, 2025.

ByteDance Seed Team. Seed2.0 model card: Towards intelligence frontier for real-world complexity. https://github.com/ByteDance-Seed/Seed2.0, 2026a. Accessed: 2026-06-27.

ByteDance Seed Team. Seedream 5. https://seed.bytedance.com/tech/seedream5, 2026b. Accessed: 2026-06-27.

A. Chalouan, A. J. Gil, A. Chabli, K. Bargach, H. Liemlahi, K. El Kadiri, V. Tendero-Salmeron, and J. Galindo-Zaldivar. cgps record of active extension in moroccan meseta and shortening in atlasic chains under the eurasia-nubia convergence. *Sensors*, 23:4846, 2023. doi: 10.3390/s23104846. URL https://doi.org/10.3390/s23104846.

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

Yifan Chang, Jiaxin Ai, Jianwen Sun, Yuandong Pu, Siqi Luo, Liangliang Zhao, Yuchen Ren, Minghao Liu, Yunfei Yu, Yu Qiao, Kaipeng Zhang, and Yihao Liu. Faithful, enriched, and precise: Benchmarking natural-science illustration generation by T2I models. *arXiv preprint arXiv:2606.05949*, 2026.

Marius-F. Danca. Mandelbrot set as a particular julia set of fractional order, equipotential lines and external rays of mandelbrot and julia sets of fractional order. *Fractal and Fractional*, 8(1):69, 2024. doi: 10.3390/fractalfract8010069. URL https://doi.org/10.3390/fractalfract8010069.

Anurag Dhote, Mohammed Javed, and David S Doermann. A survey on figure classification techniques in scientific documents. *arXiv preprint arXiv:2307.05694*, 2023.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Licheng Zhang, Xiaorui Wang, and Zhendong Mao. DeepResearch Bench: A comprehensive benchmark for deep research agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hQ0K2Hhq7H.

Jiaxin Ge, Zora Zhiruo Wang, Xuhui Zhou, Yi-Hao Peng, Sanjay Subramanian, Qinyue Tan, Maarten Sap, Alane Suhr, Daniel Fried, Graham Neubig, and Trevor Darrell. AutoPresent: Designing structured visuals from scratch. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2902–2911, 2025.

Google. Gemini 3 Pro Image. https://ai.google.dev/gemini-api/docs/models/gemini-3-pro-image, 2025. Accessed: 2026-06-27.

Google. Gemini 3.1 Pro. https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro, 2026. Accessed: 2026-06-27.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.

Siyuan Huang, Yutong Gao, Juyang Bai, Yifan Zhou, Zi Yin, Xinxin Liu, Rama Chellappa, Chun Pong Lau, Sayan Nag, Cheng Peng, and Shraman Pramanick. SciFig: Towards automating scientific figure generation. *arXiv preprint arXiv:2601.04390*, 2026.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9307–9315, 2024.

K. V. Jobin, A. Mondal, and C. V. Jawahar. DocFigure: A dataset for scientific document figure classification. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, pp. 74–79. IEEE, 2019.

Z. Karishma, S. Rohatgi, K. S. Puranik, J. Wu, and C. L. Giles. ACL-Fig: A dataset for scientific figure classification. In *Proceedings of the Workshop on Scientific Document Understanding (SDU 2023)*, 2023.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1–2):81–93, 1938. doi: 10.1093/biomet/30.1-2.81.

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

Xiao-Ming Li, Liang Chang, Jian Zhou, Jili Sun, Xingqiang Lu, Xingbo Han, Jianping Wang, Hongliang Lu, Xing Li, Ning Wang, Yujia Qiu, and Yirong Wu. Haishao-1 satellite: Low-inclination orbit spaceborne synthetic aperture radar. *The Innovation*, 6(9):100949, 2025. doi: 10.1016/j.xinn.2025.100949. URL https://doi.org/10.1016/j.xinn.2025.100949.

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

Chumeng Liang and Jiaxuan You. DiagramEval: Evaluating LLM-generated diagrams via graphs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 12678–12690, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.640. URL https://aclanthology.org/2025.emnlp-main.640/.

Ruo-Ying Liu, Bing-Zhi Li, Ying-Jin Yuan, and Zhi-Hua Liu. Multiscale metabolic engineering in biological lignin valorization. *The Innovation*, 6(11):100993, 2025a. doi: 10.1016/j.xinn.2025.100993. URL https://doi.org/10.1016/j.xinn.2025.100993.

Z. Liu, Z. Guo, Z. Su, L. Huang, Y. Yang, Z. Han, Z. Pan, and W. Zhang. Paper2Any: Turn paper/text/topic into editable research figures, technical route diagrams, and presentation slides. GitHub repository, 2025b. URL https://github.com/OpenDCAI/Paper2Any. Accessed: 2026-08-26.

Yuxuan Luo, Yuhui Yuan, Junwen Chen, Haonan Cai, Ziyi Yue, Yuwei Yang, Fatima Zohra Daha, Ji Li, and Zhouhui Lian. MMMG: A massive, multidisciplinary, multi-tier generation benchmark for text-to-image reasoning. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025) Datasets and Benchmarks Track*, 2025.

Ishani Mondal, Zongxia Li, Yufang Hou, Anandhavelu Natarajan, Aparna Garimella, and Jordan Boyd-Graber. SciDoc2Diagrammer-MAF: Towards generation of scientific diagrams from documents guided by multi-aspect feedback refinement. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13342–13375, Miami, Florida, USA, 2024. Association for Computational Linguistics.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, et al. MinerU2.5: A decoupled vision-language model for efficient high-resolution document parsing. *arXiv preprint arXiv:2509.22186*, 2025.

OpenAI. GPT-5.4. https://developers.openai.com/api/docs/models/gpt-5.4, 2026a. Accessed: 2026-06-27.

OpenAI. GPT Image 2: State-of-the-art image generation model. https://developers.openai.com/api/docs/models/gpt-image-2, 2026b. Accessed: 2026-06-27.

Wei Pang, Kevin Qinghong Lin, Xiangru Jian, Xi He, and Philip Torr. Paper2Poster: Towards multimodal poster automation from scientific papers. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025) Datasets and Benchmarks Track*, 2025.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

Kaiyue Sun, Rongyao Fang, Chengqi Duan, Xian Liu, Aoxue Li, and Xihui Liu. T2I-ReasonBench: Benchmarking reasoning-informed text-to-image generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 8919–8944, San Diego, California, United States, 2026. Association for Computational Linguistics.

Q. Wang et al. Untethered miniature robots for minimally invasive thrombus treatment: From bench to clinical trials. *The Innovation*, 6(6):100874, 2025. doi: 10.1016/j.xinn.2025.100874. URL https://doi.org/10.1016/j.xinn.2025.100874.

Ran Zhao, Sheng Jin, Size Wu, Kang Liao, Zerui Gong, Zujin Guo, Yang Xiao, and Wei Li. Knowledge visualization: A benchmark and method for knowledge-intensive text-to-image generation. *arXiv preprint arXiv:2604.22302*, 2026.

Hao Zheng, Xinyan Guan, Hao Kong, Wenkai Zhang, Jia Zheng, Weixiang Zhou, Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. PPTAgent: Generating and evaluating presentations beyond text-to-slides. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 14402–14418, Suzhou, China, 2025. Association for Computational Linguistics.

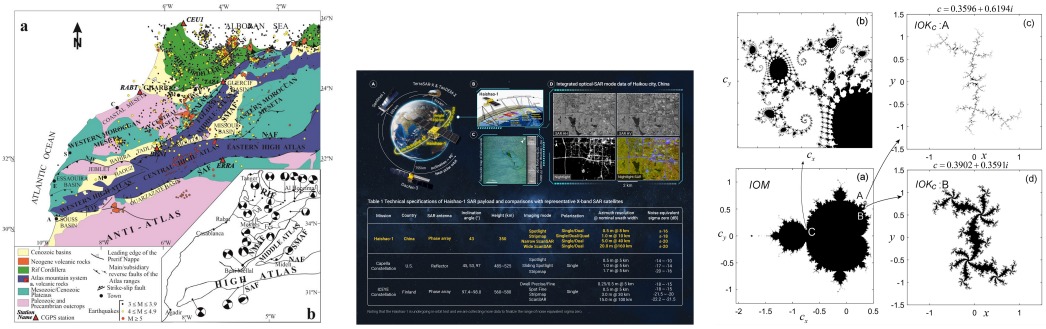
Dawei Zhu, Rui Meng, Yale Song, Xiyu Wei, Sujian Li, Tomas Pfister, and Jinsung Yoon. PaperBanana: Automating academic illustration for AI scientists. *arXiv preprint arXiv:2601.23265*, 2026a.

Minjun Zhu, Zhen Lin, Yixuan Weng, Panzhong Lu, Qiujie Xie, Yifan Wei, Sifan Liu, QiYao Sun, and Yue Zhang. Autofigure: Generating and refining publication-ready scientific illustrations. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL `https://openreview.net/forum?id=5N3z9JQJKq`.

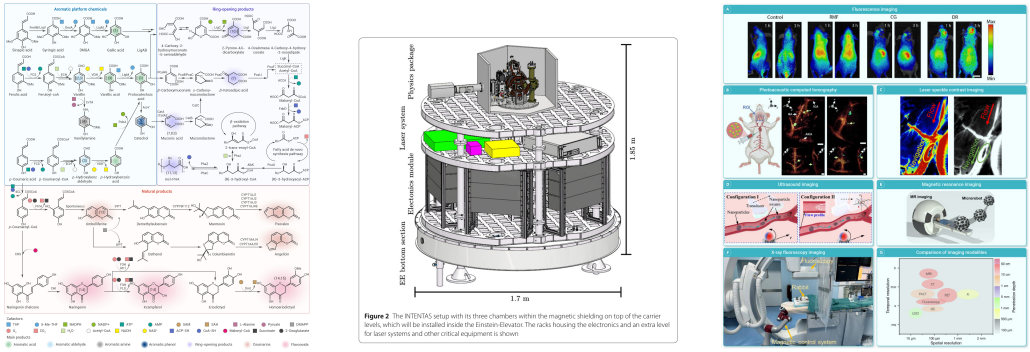
A Dataset Construction Details: Filtering and Composition

A.1 Representative Examples of Excluded Diagram Categories

The filtering step in Section 3 removes diagram categories whose correctness depends on exact numerical, geometric, chemical, spatial, or optical fidelity. Figure 3 combines the six representative examples supplied with the dataset materials. They make the filtering decision concrete but are not benchmark instances. The source images are only scaled to fit the composite layout; no panel is cropped, recolored, annotated, or otherwise altered.



(a) Data-driven map **(b)** Astronomical sensing **(c)** Mathematical principle



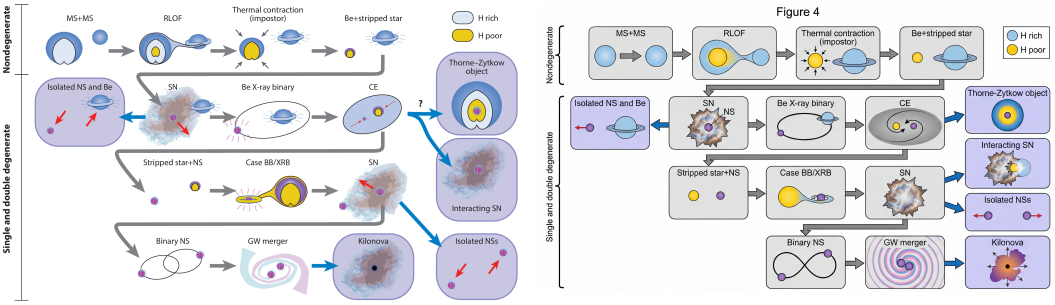
(d) Chemical structure and reaction **(e)** Mechanical CAD **(f)** Photos and imaging panels

Figure 3: Representative examples of categories excluded from end-to-end generation by agents.

Sources and permissions. Panel (a) is reproduced from Chalouan et al. (2023), DOI: 10.3390/s23104846, under the CC BY 4.0. Panel (b) is reproduced unchanged from Figure 1 of Li et al. (2025), DOI: 10.1016/j.xinn.2025.100949, under the CC BY-NC-ND 4.0. Panel (c) is reproduced from Figure 1 of Danca (2024), DOI: 10.3390/fractalfract8010069, under the CC BY 4.0. Panel (d) is reproduced from Figure 2 of Liu et al. (2025a), DOI: 10.1016/j.xinn.2025.100993, under the CC BY 4.0. Panel (e) is reproduced from Figure 2 of Anton et al. (2025), DOI: 10.1140/epjqt/s40507-025-00330-9, under the CC BY 4.0; the source article does not identify this panel as third-party material. Panel (f) is reproduced from Figure 4 of Wang et al. (2025), DOI: 10.1016/j.xinn.2025.100874, under the CC BY-NC-ND 4.0. Panels (b) and (f) must remain unmodified and are subject to the non-commercial and no-derivatives conditions of their license. Any future modification to a CC BY panel should be identified as an adaptation.

A.2 Potential Training-Data Contamination Filtering

To mitigate potential training-data contamination, we manually inspect the candidate diagrams and exclude those whose preliminary generation output reproduces the reference diagram with anomalously high similarity. Figure 4 illustrates a reference–generation pair used to exemplify this filtering criterion. Despite differences in rendering style, the Vanilla-generated diagram closely follows the reference's overall layout, labeled evolutionary stages, and branching structure. We treat this similarity as a precautionary filtering signal rather than conclusive evidence of training-data contamination.



(a) Reference diagram

(b) Vanilla-generated diagram

Figure 4: Example illustrating potential training-data contamination filtering. The reference diagram (a) and Vanilla-generated output (b) share a similar organization of evolutionary stages and branches, motivating precautionary exclusion during manual screening. Both images are shown without cropping or other content modifications.

A.3 Dataset Composition

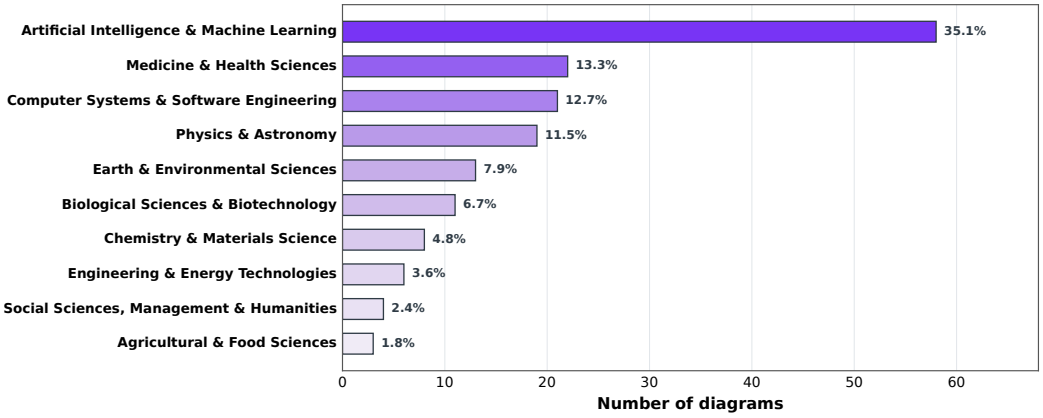


Figure 5: Domain distribution of all 165 diagrams. Bar lengths indicate counts; labels at bar ends indicate percentages.

B Detailed Definitions of the 11 Diagram Types

This appendix provides the detailed definition and characteristics of each of the 11 diagram types.

B.1 TIMELINE (TL)

Definition

A timeline is a visual schematic that arranges information along a chronological axis. It maps the temporal dimension through spatial positioning to illustrate the developmental trajectory, evolutionary trends, or historical progression of a subject.

Characteristics

It utilizes a temporal axis as the structural backbone of the diagram, annotated with explicit chronological markers (e.g., specific years such as 2018, 2023, and 2024, or precise months).
Specific information is conveyed through nodes. These nodes typically incorporate text labels, icons, or logos to signify particular events occurring or milestones achieved at distinct points in time.

B.2 LINEAGE DENDROGRAM (LD)

Definition

A lineage dendrogram is a visual representation illustrating genealogical relationships and evolutionary history. Utilizing a hierarchically distinct branching structure, it depicts how a subject (e.g., biological species, genetic sequences, technological models, or celestial bodies) diverges, proliferates, or iterates from a single common ancestor or root. Driven by the passage of time, environmental selection, or anthropogenic intervention, it delineates the comprehensive historical pathway culminating in the subject's current diversified state.

Characteristics

It features a root or origin, denoting the starting point or primordial state of the subject.
Branches (or edges) represent evolutionary pathways or genetic distances. The length of these lines frequently bears quantitative significance, indicating metrics such as the duration of time or the degree of divergence.

B.3 PHASED EVOLUTION (PE)

Definition

A phased evolution diagram visually represents systemic evolution by segmenting the process into distinct, discrete stages or phases. It primarily focuses on demonstrating qualitative transformations, paradigm shifts, or transitions in environmental states that occur during the subject's evolutionary trajectory. Each phase typically represents a relatively stable state or a specific operational mechanism, with logical progressions or cyclical relationships existing between successive phases.

Characteristics

The diagram is partitioned into distinct spatial blocks (e.g., panels, boxes, or tiers), with each block signifying a specific evolutionary stage.
Within each "stage block," the diagram typically details the micro-structures, chemical/physical reactions, system architectures, or key elements that are unique to that particular phase.

B.4 ALGORITHMIC FRAMEWORK (AF)

Definition

An algorithmic framework diagram is a visual schematic that delineates the overall architecture, core components, data flow, and execution logic of an algorithm through a system of graphical symbols. It translates abstract algorithmic models into intuitive, modular blueprints to depict a complete and cohesive framework.

Characteristics

It demonstrates how an entire system accomplishes complex tasks by orchestrating multiple components. Furthermore, it illustrates the internal structure of the abstract system, encompassing inter-module interactions, engagement with the external environment, and multi-stage process control.

B.5 COMPUTER SYSTEM ARCHITECTURE (CSA)

Definition

A computer system architecture diagram is a visual representation used to abstract and delineate the hardware-software organizational structure, logical component partitioning, and the interaction interfaces and communication mechanisms within a computing system. Utilizing graphical language, it demarcates system boundaries, elucidates the dependencies between different tiers, and illustrates the routing paths of data and control signals across heterogeneous components (e.g., CPUs, GPUs, FPGAs, OS kernels, and user-space applications).

Characteristics

The system is explicitly partitioned into distinct logical tiers (e.g., hardware layer, driver layer, kernel layer, application layer) or functional modules.

B.6 CONCEPTUAL FRAMEWORK (CF)

Definition

A conceptual framework diagram is a visual representation that leverages specific visual elements to organize the relationships among the internal components of a research subject in a hierarchical, parallel, or progressive manner. This intuitively illustrates the architecture of a specific research domain or technological trajectory. By mapping the core operational mechanisms, logical workflows, or classification systems through graphical language or theoretical models, it facilitates a rapid comprehension of the ”comprehensive overview” or ”underlying logic” of the research direction.

Characteristics

It presents a holistic landscape of a specific research domain or technological architecture, typically characterized by explicit hierarchical or block-based demarcations.

B.7 TECHNICAL ROADMAP (TR)

Definition

A technical roadmap is a visual representation of a highly standardized, actionable, and reproducible protocol engineered to achieve specific scientific, engineering, or clinical objectives.

Characteristics

Highly Ordered Visual Layout: It demonstrates a strong sense of visual order, utilizing spatial segmentation to articulate temporal and task-based relationships.
Phased Progression: It typically incorporates explicit phase or stage markers that divide prolonged procedures, thereby facilitating macroscopic progress management.
Containerization: It employs panels or state-machine concepts to aggregate and categorize parallel processes or attribute-related tasks, effectively mitigating visual clutter.
Directional Flow: The primary directional flow adheres to conventional "left-to-right" or "top-to-bottom" reading patterns, signifying chronological progression or the primary task trajectory. Inverse flow is strictly restricted to denoting "feedback," "error reporting," or "correction."
Non-Linear Mechanisms: It may deviate from a singular linear process by incorporating core mechanisms to simulate authentic scientific or engineering decision-making processes, including Branching (path selection predicated on "If-Then" logic, explicitly delineating differentiated execution protocols under varying conditions), Convergence (funnel-based screening predicated on "Filter" logic, illustrating the hierarchical selection process from vast candidate pools to core targets), and Iteration (corrective circuits predicated on "While-Loop" logic, permitting self-correction and feedback via retrograde arrows).
Multimodal Integration: It may transcend abstract block diagrams by integrating concrete data and states, including State Visualization (clearly illustrating morphological or qualitative transformations occurring within the processed subjects, e.g., data, samples, models, throughout the procedure) and Heterogeneous Data Integration (appending specific reference standards, e.g., anatomical diagrams, code snippets, genetic locus maps, adjacent to the workflow, serving as intuitive benchmarks for the execution of corresponding steps).

B.8 Graphical Abstract (GA)

Definition

A graphical abstract is a single-panel illustration that provides a structured, intuitive summarization of a scientific paper's core content through highly condensed visual language.

Characteristics

Serving as the visual counterpart to the textual abstract, it is designed to enable readers to rapidly grasp the research background, underlying mechanisms, principal findings, and conclusions. It functions independently of the main text; readers can comprehend the complete scientific narrative conveyed by the diagram relying solely on its intrinsic elements and annotations, without needing to consult the manuscript for legend explanations.

B.9 Experimental Setup Diagram (ESD)

Definition

An experimental setup diagram is a visual and abstract representation of a hardware system constructed to validate specific scientific hypotheses or to acquire data.

Characteristics

Topology and Flow: It depicts the physical composition and spatial layout of experimental equipment, while explicitly defining the connection topology (physical, optical, or electrical circuits) among various components, as well as the flow logic of energy or information.
Balanced Abstraction: It oscillates between realism and abstraction by employing primary morphological features or strictly standardized symbols to represent devices, streamlining extraneous details to accentuate structural relationships.

B.10 Illustrative Case (IC)

Definition

An illustrative case is a visual representation that concretely demonstrates the actual execution process of an algorithm, the detailed reasoning logic of a model, or specific flaws within a system. It achieves this by utilizing concrete, real-world data instances (such as input texts, code snippets, API call parameters, or dialogue contents).

Characteristics

It employs concrete, real-world data instances (Concrete Examples) as its core medium of representation.

B.11 Mechanism Diagram (MD)

Definition

A mechanism diagram is a visual representation designed to unveil the specific operational principles underlying a particular phenomenon, function, or event from a microscopic or concrete perspective.

Characteristics

It primarily addresses the question "How does this specific event occur? (Principle)" rather than "What kind of system is this? (Structure)", placing a primary emphasis on Process and Function.
By employing representative elements (e.g., molecules, particles, data streams, force fields), it demonstrates how entities engage in physical, chemical, or informational interactions, explicitly illustrating the outcomes when different elements converge.

C Evaluation Dimensions and Criteria

Here we provide the detailed definitions and evaluation criteria for the hierarchical evaluation dimensions: Structure, Content, and Aesthetics.

C.1 Structure

Definition

Structure refers to the foundational architectural framework and topological organization of visual elements within a scientific diagram. It constitutes the visual "skeleton" of the diagram, encompassing the spatial layout, logical flow, and hierarchical relationships that govern how information is arranged, grouped, and interconnected.

Evaluation Criteria

Visual Organizational Layout

Positive Indicators (Strengths):

- Clear Reading Path and Logical Flow: Possesses distinct starting and ending points, ensuring a seamless and intuitive progression of the primary workflow.
- Effective Modularization and Grouping: Employs bounding boxes strategically to cluster elements belonging to the same logical tier, thereby reducing cognitive load.

- Explicit Hierarchical Relationships: Maintains a clear distinction between the overarching architecture and localized details, rendering parent-child hierarchical dependencies immediately comprehensible.
- Clean Topological Routing: Features clear arrow directionality and meticulously avoids unnecessary crossing of connecting lines. When depicting complex feedback loops, it utilizes appropriate curves or detours to preserve overall visual neatness.
- Even Spatial Distribution: Achieves a harmonious and well-proportioned distribution of the overall structure across the canvas (via either symmetry or dynamic equilibrium). Maintains a relatively consistent density of information nodes across different regions, complemented by an optimal utilization of white space.

Negative Indicators (Weaknesses):

- Structural Imbalance and Misplaced Emphasis: Marginalizes core processes to the periphery while centralizing secondary workflows; alternatively, exhibits disproportionate spatial scaling among peer-level modules, thereby misleading readers regarding their relative importance.
- Absence of Visual Partitioning: Conflates modules of disparate stages or functions without employing adequate spacing or boundary frames for physical isolation. This results in element overcrowding and potential visual occlusion.
- Haphazard Layout and Illogical Arrangement: Exhibits an arbitrary positioning of elements where variations in spatial density lack any underlying structural or semantic rationale. Rather than using intentional clustering or spacing to convey meaning, the layout features random areas of congestion and unjustified empty spaces, resulting in a disorganized visual narrative.
- Spatial Detachment and Referential Ambiguity: Positions logically related components—such as text labels, annotations, or legends—at excessive distances from their corresponding graphical objects. This physical separation disrupts visual association, thereby generating referential ambiguity and confusing the reader.

Logical Accuracy

Positive Indicators (Strengths):

- Structural Abstraction and In-depth Detail Mining: Effectively distills lengthy and complex text to construct a lucid core visual skeleton while fully extracting and integrating high-value details from the paper (e.g., specific intermediate steps, conditional variables, and micro-mechanisms). Without compromising the readability of the overarching framework, it highly encourages the generation of information-rich, fine-grained diagrams to maximize independent explanatory power.
- Precise and Profound Logical Mapping: Ensures the topological structure remains strictly faithful to the original paper's intent. Accurately reconstructs deep-level associations within the text—encompassing not merely basic causal or temporal relationships, but also precisely capturing and visualizing complex feedback loops, conditional branches, and nested hierarchies. This transforms exhaustive textual narratives into a multi-layered, visually articulate logical network.

Negative Indicators (Weaknesses):

- Detail Overload and Structural Imbalance: Mechanically enumerates redundant nodes or secondary information without adequately distinguishing primary from subordinate elements (i.e., lacking effective information hierarchy management). This induces severe visual clutter and obfuscates the core logical trajectory, thereby negating the diagram's fundamental purpose of elucidating complex information.
- Insufficient Granularity and Explanatory Deficit: Exhibits an overly rudimentary structure confined to superficial macroscopic overviews. Omits pivotal transitional steps, core mechanistic branches, or crucial underlying details explicitly articulated

in the paper, resulting in vacuous content and incomplete workflows that fail to convey the comprehensive analytical depth of the research.

- Structural Hallucination and Logical Misalignment: Presents visual topological relationships that fundamentally conflict with the factual assertions or logical deductions of the paper. Alternatively, it erroneously conflates and associates information nodes that lack any inherent logical intersection during the detail extraction process.

C.2 CONTENT

Definition

Content refers to the concrete, discrete constituent elements utilized to populate the structural framework of a scientific diagram. It represents the specific, localized building blocks—encompassing text labels, scientific nomenclature, alphanumeric data, standard symbols, and individual graphical motifs (such as specific molecular models, cell depictions, or distinct physical components)—that inherently carry factual, domain-specific meaning independent of their spatial arrangement.

Evaluation Criteria

Visual Element Representation

Positive Indicators (Strengths):

- Precise Semantic Representation: The selected graphical elements, icons, and symbols intuitively and accurately convey their intended scientific meanings through attributes such as color and shape. The element design is highly distilled, entirely devoid of extraneous ornamentation or non-functional details that might impede comprehension.
- Adequate Legend Provision: Clear, legible, and appropriately positioned explanatory legends are provided for key abstract elements, abbreviations, non-standard symbols, as well as for color schemes and visual encodings used to differentiate experimental groups or logical classifications. This ensures the self-explanatory capacity of the content.
- High-Fidelity Micro-Rendering: Individual visual components (e.g., cellular structures, laboratory apparatuses, molecular models) exhibit sharp edges and accurate proportions, remaining completely free from generative artifacts, blurring, or the 'melting' effects typical of large models. Text and numerical data are rendered in crisp, vectorized, or high-resolution typography, devoid of typographical overlapping or garbled characters.

Negative Indicators (Weaknesses):

- Redundancy and Semantic Duplication: Elements are subjected to unnecessary repetitive rendering; alternatively, disparate visual entities are employed to represent a singular scientific concept, engendering cognitive ambiguity and misinterpretation.
- Generative Morphological Distortion: Localized graphical elements manifest severe morphological anomalies (e.g., erroneous physical linkages in biological structures, logically flawed chemical bond angles) or characteristic AI-generated distortions, thereby undermining the scientific rigor and validity of the content.
- Information Deficit and Illegibility: Suboptimal generation quality results in critical text, formula labels, or minute symbols becoming blurred, fragmented, or rendered as unidentifiable garble. Furthermore, the absence of requisite legend guidance for complex systems precludes readers from successfully decoding visual symbols into their underlying scientific concepts.

Content Accuracy

Positive Indicators (Strengths):

- Precise and Comprehensive Informational Articulation: Predicated on the exact and unambiguous utilization of technical terminology, data labels, and standardized symbols, the representation thoroughly extracts and delineates intricate details such as specific parameters, experimental conditions, or physical/chemical mechanisms outlined in the paper. The informational elements within the diagram are not solely factually accurate but also manifest a high degree of scientific explanatory capacity.
- Terminological Consistency: Abbreviations and variable nomenclatures deployed within the visual elements are strictly consistent with the textual discourse of the paper.
- Rigorous Logical Correspondence: The transformation and mapping of informational elements across disparate procedural stages exhibit strict logical coherence, remaining entirely devoid of informational discontinuities or omissions.
- High-Value Informational Richness and Depth: The model demonstrates a profound capacity for mining and synthesizing macroscopic information from the paper (encompassing textual narratives, tabular data, and formulaic derivations), effectively translating tacit knowledge or auxiliary data (e.g., specific experimental environments, dosages, temporal intervals) into substantive visual content. Provided that the augmentation of detail does not devolve into excessive clutter that obscures the core scientific value, representations characterized by greater informational density and finer granularity warrant highly positive appraisal.

Negative Indicators (Weaknesses):

- Insufficiency of Critical Details: There is a failure to effectively extract the requisite parameters, conditions, or mechanistic nuances from the paper that underpin the core logic of the visual representation. As a result, the information remains thin and only lists superficial macroscopic concepts.
- Factual, Logical, and Terminological Inaccuracies: The informational elements conveyed contradict the established parameters of the paper or recognized scientific facts. This encompasses profound factual and logical fallacies (e.g., erroneously depicted physical/biological structures, visual representations of data points that contradict the paper's conclusions), as well as surface-level representational inconsistencies and ambiguities. Examples include the utilization of discrepant variable nomenclatures, imprecise or ambiguous technical terminology, erroneous mathematical subscripts or symbols, and the introduction of undefined non-standard abbreviations.
- Informational Noise and Irrelevant Elements: Introduced textual descriptions, graphical motifs, or specific data points bear no relevance to the localized thematic context; alternatively, the unrestrained over-accumulation of trivial details engenders a diffusion of visual focus, causing the valuable core content to be submerged beneath extraneous informational noise.

C.3 AESTHETICS

Definition

Aesthetics refers to the stylistic articulation, visual presentation, and overall graphic design quality of a scientific diagram. It constitutes the visual 'presentation layer' superimposed upon the structural framework and content elements, encompassing deliberate choices in color palettes (including harmony and contrast), typographic styling (font types and legibility), line weights, gradients, textural details, and holistic visual polish.

Evaluation Criteria

Stylistic Execution and Visual Polish

Positive Indicators (Strengths):

- Harmonious and Functional Color Palettes: The application of color is both visually unified and functionally deliberate. Colors are appropriately utilized to clearly delineate distinct data modules, guide the viewer's visual flow, or emphasize critical scientific conclusions. The color selection aligns with professional scientific publishing standards and ensures good contrast and accessibility.
- Coherent Visual Language: The overall stylistic execution demonstrates a high degree of consistency. The diagram maintains a unified visual language (e.g., strictly adhering to either a 2D flat design or a consistent 3D rendering style) across all independent content elements. Graphical properties—such as stroke weights, corner radii, arrow typologies, and shading techniques—are standardized throughout, preventing any discordant patchwork effect.
- Typographic Legibility: Typography is scaled appropriately, and font properties (e.g., font weights, italics) are judiciously applied to establish a clear visual hierarchy and ensure optimal readability without straining the viewer's eyes.

Negative Indicators (Weaknesses):

- Inharmonious or Chaotic Color Schemes: Exhibits extensive use of highly saturated, fluorescent, or visually clashing color palettes. A distinct lack of color logic is evident (e.g., utilizing dark backgrounds that inadvertently obscure foreground text and reduce contrast).
- Fragmented or Collage-like Style: Displays a discordant integration of source materials or AI-generated segments, such as the jarring juxtaposition of flat (2D) and volumetric (3D) graphical styles within the same workflow, resulting in a compromised and fractured visual consistency.
- Visual Noise and Unnecessary Decoration: Introduces superficial, non-functional decorative elements (e.g., superfluous drop shadows, exaggerated 3D lighting/glow effects, or overly complex background textures) that obscure primary data and distract readers from the core scientific message, thereby diminishing the academic rigor and seriousness of the diagram.
- Disproportionate Scaling and Distortion: Shows inappropriate aspect ratios, wherein critical graphical elements, icons, or typographic components are erroneously stretched, compressed, or otherwise subjected to spatial distortion to fit a layout, sacrificing aesthetic proportion.

D TYPE-SPECIFIC EVALUATION WEIGHTS

Table 6 reports the top-level dimension weights $W_{d,\tau}$ used in Equation 3. As described in Section 3.3, these weights are generated by Gemini 3.1 Pro and averaged over five independent trials, with the three weights summing to one for each diagram type. Technical Roadmap assigns a higher weight to Structure (0.50) than to Content (0.35), whereas Illustrative Case places greater emphasis on Content (0.50) than on Structure (0.35).

Table 6: Type-specific top-level dimension weights $W_{d,\tau}$. Each column sums to one. Type abbreviations follow Table 1.

Dimension	TL	LD	PE	AF	CSA	CF	TR	GA	ESD	IC	MD
Structure	0.45	0.46	0.45	0.50	0.50	0.52	0.50	0.41	0.46	0.35	0.45
Content	0.40	0.39	0.40	0.35	0.35	0.33	0.35	0.40	0.38	0.50	0.40
Aesthetics	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.19	0.16	0.15	0.15

E HUMAN ANNOTATION PROTOCOL

This appendix specifies the protocol used to obtain the human preference labels reported in Section 4.1. The annotation task compared a generated diagram with its human-created reference for the same paper and target caption. Annotators were shown the pair in a web interface and were not given any automatic evaluation score or method identity. The interface used for this task is shown in Figure 6.

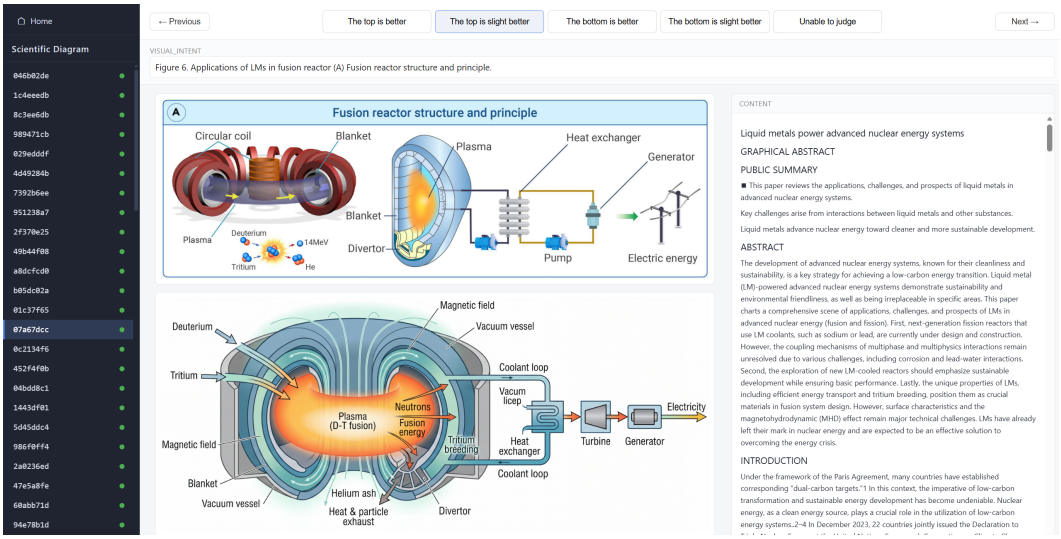


Figure 6: Human annotation interface used to compare a generated diagram with its reference. The screenshot is reproduced unchanged and includes Figure 6A from Lin Zhang, Chang Deng, Xu Ji, and Xiaojing Liu, "Liquid metals power advanced nuclear energy systems," *The Innovation* 6, 100959 (2025), DOI: 10.1016/j.xinn.2025.100959. The source figure is reproduced under the CC BY-NC-ND 4.0 license; no changes were made.

Five-point preference scale. Annotators selected one of five ordered responses: "top is better", "top is slightly better", "unable to judge", "bottom is slightly better", or "bottom is better". The two "top" responses correspond to a preference for the generated diagram when it was displayed in the top position; the two "bottom" responses correspond to a preference for the reference when it was displayed in the bottom position. The display order was anonymized, and labels were mapped back to generated-versus-reference identities before aggregation. We encode the responses as $+2$, $+1$, 0, -1, and -2, respectively, after this identity mapping; the sign denotes direction and the magnitude denotes preference strength.

Independent annotation and strict aggregation. Each of the 110 Vanilla–reference pairs was labeled independently by three annotators. For each instance i, let $z_{i,r} \in \{-2, -1, 0, +1, +2\}$ be annotator r's identity-mapped response. We first assign a directional label $d_{i,r} = \text{sign}(z_{i,r})$: generated-preferred, reference-preferred, or unable to judge. An instance enters the high-agreement set only when at least two annotators provide directionally consistent judgments, and no annotator provides a conflicting directional judgment. Thus, instances lacking at least two consistent directions or containing directional conflicts are excluded. This rule excludes five direction-conflict instances and retains 105 instances, comprising 35 generated-preferred and 70 reference-preferred cases.

For every retained instance, the human preference-strength label is the median of the three signed ordinal responses,

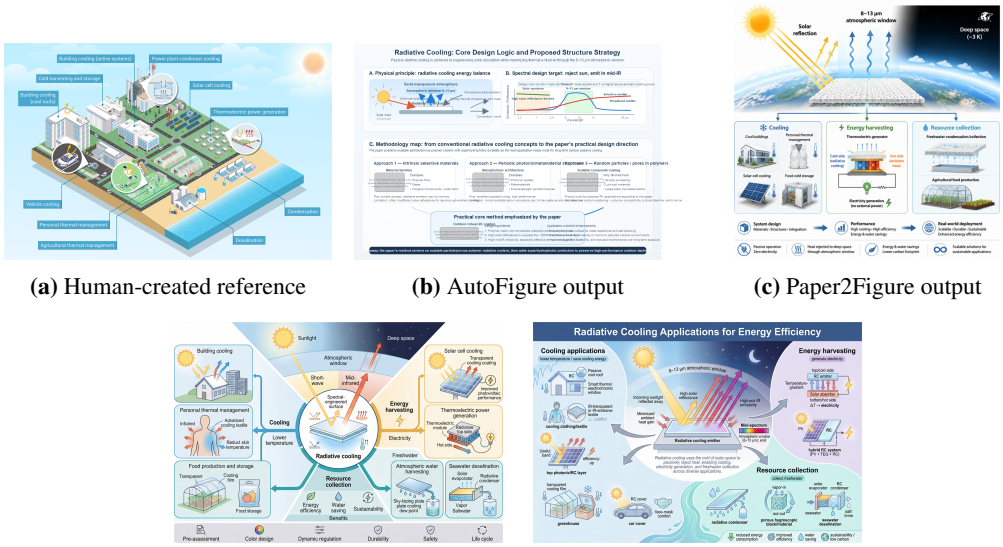
$$h_i = \text{median}(z_{i,1}, z_{i,2}, z_{i,3}), \qquad i \in \mathcal{H}, \tag{5}$$

where $\mathcal{H}$ is the 105-instance high-agreement set. The median preserves the ordinal nature of the scale and prevents one annotator's extreme strength judgment from determining the instance label. Directional analyses use $\text{sign}(h_i)$, whereas rank-correlation analyses use h_i itself. No automatic score is used in either filtering or aggregation.

Agreement statistics. Fleiss' κ is computed on the three-way directional labels over all 110 instances. Ordinal Krippendorff's α is computed from the five-level responses using their ordinal distance. Confidence intervals are obtained by resampling instances, preserving the three annotations within each instance; the resulting estimates are Fleiss' $\kappa = 0.730$ (95% CI $[0.630, 0.821]$) and ordinal Krippendorff's $\alpha = 0.799$ (95% CI $[0.744, 0.843]$).

F Qualitative Case Study

Figure 7 presents one example instance from the Illustrative Case category.



(a) Human-created reference **(b)** AutoFigure output **(c)** Paper2Figure output

(d) Vanilla output **(e)** PaperBanana output

Figure 7: Top row: (a) the original human-created reference, (b) AutoFigure output, and (c) Paper2Figure output. Bottom row: (d) Vanilla output and (e) PaperBanana output.

G Per-Category Dimension-Level Results

Table 7: Per-category dimension-level results for PaperBanana and Vanilla. Structure and Content rows are the weighted top-level scores; VOL, LA, VER, and CA are their corresponding sub-dimensions.

Method	Dimension	Diagram type											Overall
		TL	LD	PE	AF	CSA	CF	TR	GA	ESD	IC	MD	
PaperBanana	Structure (weighted)	0.5379	0.4843	0.5797	0.5632	0.5091	0.5834	0.5243	0.6204	0.4572	0.5037	0.4953	0.5326
	VOL	0.5302	0.5086	0.5738	0.5697	0.5076	0.5786	0.5295	0.6186	0.4833	0.5361	0.5176	0.5412
	LA	0.5459	0.4659	0.5893	0.5553	0.5105	0.5896	0.5174	0.6228	0.4355	0.4764	0.4766	0.5259
	Content (weighted)	0.5002	0.4827	0.6195	0.5574	0.4810	0.5557	0.5432	0.6090	0.4478	0.4586	0.4914	0.5224
	VER	0.5313	0.5168	0.6306	0.5683	0.4821	0.5371	0.5577	0.5858	0.4685	0.4779	0.4935	0.5318
	CA	0.4803	0.4529	0.6146	0.5482	0.4798	0.5686	0.5338	0.6311	0.4323	0.4435	0.4885	0.5158
	Aesthetics	0.5762	0.5366	0.6020	0.6003	0.4674	0.5564	0.5597	0.6470	0.5058	0.5192	0.5281	0.5544
	Overall	0.5272	0.4917	0.5975	0.5665	0.4926	0.5703	0.5361	0.6195	0.4610	0.4824	0.4985	0.5312
Vanilla	Structure (weighted)	0.4515	0.3957	0.5464	0.4531	0.4980	0.5912	0.5003	0.6198	0.5071	0.3675	0.4694	0.4909
	VOL	0.4602	0.4171	0.5382	0.4684	0.5022	0.5979	0.5358	0.6193	0.5428	0.4122	0.4973	0.5083
	LA	0.4418	0.3794	0.5574	0.4362	0.4942	0.5848	0.4654	0.6188	0.4758	0.3301	0.4475	0.4756
	Content (weighted)	0.3970	0.4158	0.5772	0.4565	0.4860	0.5687	0.4976	0.6087	0.5133	0.3566	0.4351	0.4829
	VER	0.4433	0.4683	0.5851	0.4859	0.4733	0.5448	0.5494	0.5876	0.5465	0.4235	0.4548	0.5057
	CA	0.3631	0.3648	0.5736	0.4359	0.4947	0.5851	0.4538	0.6282	0.4910	0.3050	0.4202	0.4650
	Aesthetics	0.5123	0.4652	0.5847	0.5368	0.5040	0.5833	0.5638	0.6547	0.6068	0.4484	0.4725	0.5393
	Overall	0.4406	0.4168	0.5626	0.4670	0.4947	0.5820	0.5094	0.6228	0.5258	0.3740	0.4566	0.4957

Table 7 reports the complete per-category breakdown for PaperBanana and Vanilla. In addition to the weighted Structure and Content scores, we report their Visual Organizational Layout (VOL),

Logical Accuracy (LA), Visual Element Representation (VER), and Content Accuracy (CA) sub-dimensions, together with Aesthetics and the overall score. All entries are relative scores against the human-created reference, so 0.5 denotes human parity. Category abbreviations follow Table 1.

Analysis. The breakdown clarifies why the aggregate advantage of the more elaborate agent is not uniform. PaperBanana outperforms Vanilla on seven of the eleven categories, but Vanilla remains better on Computer System Architecture (CSA), Conceptual Framework (CF), Graphical Abstract (GA), and Experimental Setup Diagram (ESD). The ESD gap is especially pronounced: PaperBanana's overall score is 0.4610 versus 0.5258 for Vanilla, and it trails Vanilla in every reported dimension, including Structure (0.4572 vs. 0.5071), Content (0.4478 vs. 0.5133), and Aesthetics (0.5058 vs. 0.6068). In particular, Content Accuracy is 0.4323 for PaperBanana compared with 0.4910 for Vanilla. CSA shows a milder version of the same failure mode: PaperBanana has slightly stronger Structure (0.5091 vs. 0.4980), but lower Content (0.4810 vs. 0.4860), Content Accuracy (0.4798 vs. 0.4947), and Aesthetics (0.4674 vs. 0.5040), leaving the overall scores nearly tied (0.4926 vs. 0.4947). The fact that Content Accuracy is lower for PaperBanana on both ESD and CSA is consistent with the possibility that explicit intermediate content planning can discard details or impose an overly rigid structure on categories that benefit from holistic generation. This table-level evidence is diagnostic rather than causal, and does not by itself identify which planning component is responsible.

The difficult categories also reveal a ceiling shared by current systems. PaperBanana improves over Vanilla on Lineage Dendrogram (0.4917 vs. 0.4168), Illustrative Case (0.4824 vs. 0.3740), and Mechanism Diagram (0.4985 vs. 0.4566), yet remains below the 0.5 human-parity reference in all three. Their accuracy sub-dimensions are particularly limiting: PaperBanana reaches only 0.4529/0.4659 (CA/LA) on Lineage Dendrogram, 0.4435/0.4764 on Illustrative Case, and 0.4885/0.4766 on Mechanism Diagram. Thus, planning improves relative performance over the minimal baseline, but does not reliably solve precise topology, concrete entity depiction, or causal-relation fidelity. Across these categories, the recurring gap between visual organization or element representation and the corresponding accuracy score suggests that diagrams can look plausible while still communicating incomplete or incorrect scientific relations.

H Underlying-Model Sensitivity

We assess backbone sensitivity within PaperBanana by replacing either its VLM or its Multimodal Generative Model while holding the other component fixed. The VLM variants are Gemini 3.1 Pro (Google, 2026), Qwen3.5-Plus (Alibaba Cloud, 2026a), and Seed 2.0 Pro (ByteDance Seed Team, 2026a); the multimodal generative model variants are GPT-Image-2 (OpenAI, 2026b), Qwen-Image-2.0-Pro (Alibaba Cloud, 2026b), and Seedream 5 (ByteDance Seed Team, 2026b).

Table 8: PaperBanana results under different underlying models. The default configuration pairs GPT-5.4 with Gemini 3 Pro Image; each variant replaces one component while holding the other fixed. Type abbreviations follow Table 1.

Model	TL	LD	PE	AF	CSA	CF	TR	GA	ESD	IC	MD	Overall
Default	0.5272	0.4917	**0.5975**	**0.5665**	0.4926	0.5703	**0.5361**	**0.6195**	0.4610	**0.4824**	**0.4985**	**0.5312**
VLM variants												
Gemini 3.1 Pro	0.4439	0.3824	0.5617	0.5380	**0.4992**	**0.5835**	0.4835	0.5699	**0.5200**	0.4751	0.4204	0.4980
Qwen3.5-Plus	0.4723	0.4103	0.5582	0.4905	0.4895	0.5721	0.4773	0.5850	0.4465	0.4750	0.4681	0.4950
Seed 2.0 Pro	**0.5277**	**0.5108**	0.5758	0.4620	0.4848	0.5614	0.4597	0.5808	0.4371	0.4455	0.4624	0.5007
Multimodal Generative Model variants												
GPT-Image-2	0.4627	0.4914	0.5932	0.5153	0.4845	0.5524	0.5082	0.6140	0.4493	0.4787	0.4318	0.5074
Qwen-Image-2.0-Pro	0.2626	0.2334	0.4801	0.2739	0.3443	0.3525	0.3546	0.3939	0.3378	0.3071	0.2729	0.3285
Seedream 5	0.2835	0.2773	0.4055	0.2563	0.2808	0.3144	0.2854	0.2874	0.2458	0.2411	0.2290	0.2824

VLM Comparison. With Gemini 3 Pro Image fixed, overall scores span a relatively narrow interval from 0.4950 to 0.5312. GPT-5.4 attains the highest overall score and leads on six of the 11 diagram types, while Gemini 3.1 Pro and Seed 2.0 Pro achieve the best performance on the remaining five types.

Multimodal Generative Model Comparison. With GPT-5.4 fixed, Gemini 3 Pro Image obtains the highest score on every diagram type and overall. GPT-Image-2 is comparatively close in aggregate (0.5074 vs. 0.5312), whereas Qwen-Image-2.0-Pro and Seedream 5 yield substantially lower overall scores of 0.3285 and 0.2824.

I Ablation Results for GPT-5.4 Judge

Table 9 presents the ablation results for the evaluation framework using GPT-5.4 as the judge VLM (the results for Gemini 3.1 Pro are provided in Section 4.4). While GPT-5.4 exhibits similar positive trends for the directional agreement rate (DAR) point estimates when the proposed mechanisms are applied, its overall alignment remains low due to its inherent calibration defects discussed in Section 4.3.

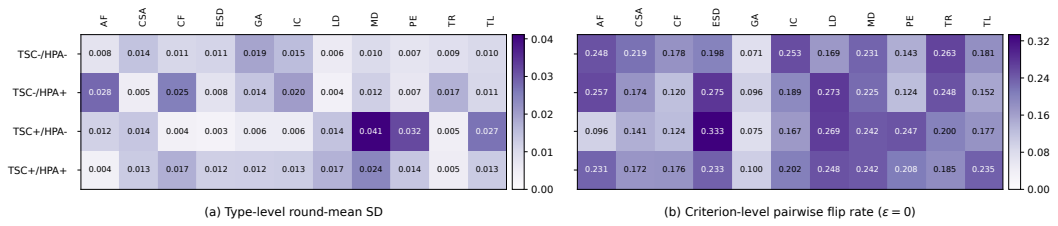
Table 9: Human-alignment results for the 2×2 evaluation-framework analysis on the 105 high-agreement instances using GPT-5.4. Gen. and Ref. recall correspond to human preferences for the generated and reference diagrams, respectively. DAR is the primary endpoint.

Experimental setup		Directional Agreement Rate			Strength Alignment	
TSC	HPA	Overall	Gen.-pref.	Ref.-pref.	Kendall's τ_b	Spearman's ρ
–	–	0.4952	0.7143	0.3857	0.0989	0.1292
–	✓	0.5143	**0.8286**	0.3571	**0.2680**	**0.3444**
✓	–	0.5143	0.7143	0.4143	0.0625	0.0710
✓	✓	**0.5714**	0.7143	**0.5000**	0.2277	0.2934

J Type-Level Stability Analysis

This appendix complements the aggregate stability results in Section 4.5 with a breakdown across diagram types. We use the same three independent scoring runs on SciDiagBench-Test and the same four configurations: ours, w.o. HPA, w.o. TSC, and w.o. HPA&TSC.

Figure 8 reports the standard deviation of type-level mean scores across the three runs and the criterion-level pairwise flip rate for each of the 11 diagram types. As shown in Figure 8, configurations using only one of TSC or HPA exhibit several pronounced type-specific fluctuations. For example, the TSC-only configuration shows relatively large score variations on Mechanism Diagram, Phased Evolution, and Timeline, whereas the HPA-only configuration exhibits a larger fluctuation on Algorithmic Framework. Experimental Setup Diagram also shows relatively high criterion-level flip rates under several configurations. In comparison, integrating TSC and HPA reduces some of these extreme type-specific variations and yields a more balanced stability profile across diagram types. These observations suggest that, although the complete framework does not uniformly improve every aggregate stability metric, combining TSC and HPA can help mitigate pronounced instability on particular diagram types. This may reflect the complementary roles of type-specific criteria and preference-based alignment in constraining the variability of the judge model across heterogeneous scientific diagrams.



(a) Type-level round-mean SD (b) Criterion-level pairwise flip rate ($\varepsilon = 0$)

Figure 8: Stability across diagram types. The left panel reports the type-level standard deviation (SD) across three runs, while the right panel reports criterion-level pairwise flip rates with $\varepsilon = 0$.

K LIMITATIONS

Limited dataset size. The benchmark contains 165 high-quality instances. Although these instances cover 11 diagram types and 10 scientific domains, reflecting the diversity of scientific diagrams encountered in practice, the dataset size is currently insufficient to support large-scale training or fine-tuning of models. A natural direction for future work is to use the current high-quality instances as seeds for automated expansion, followed by systematic filtering and human quality control. Such a process could increase coverage while preserving the structural and semantic standards of the benchmark.

Limited coverage of generation improvement. This work primarily studies how to evaluate scientific diagram generation. We include a Vanilla agent as a baseline to provide a representative generation system for the evaluation experiments, but do not present a comprehensive method for improving the generation process itself. Designing agents that can better plan, render, and verify diagrams remains an important direction for future research.

Residual reference-preference bias in VLM evaluation. Table 4 shows that the largest source of VLM evaluation error is the misclassification of reference-preferred instances. In particular, among instances that human annotators judge to be better represented by the reference image, the VLM sometimes assigns preference to the generated image: even with both TSC and HPA, the DAR is 0.8143 for the Ref.-pref. stratum, corresponding to an error rate of 18.57%, compared with a DAR of 0.9429 and an error rate of 5.71% for the Gen.-pref. stratum. Our evaluation framework suppresses this bias, but does not eliminate it completely.