AUTOFIGURE: GENERATING AND REFINING PUBLICATION-READY SCIENTIFIC ILLUSTRATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

High-quality scientific illustrations are crucial for effectively communicating complex scientific and technical concepts, yet their manual creation remains a well-recognized bottleneck in both academia and industry. We present **FigureBench**, the first large-scale benchmark for generating scientific illustrations from long-form scientific texts. It contains 3,300 high-quality scientific text–figure pairs, covering diverse text-to-illustration tasks from scientific papers, surveys, blogs, and textbooks. Moreover, we propose **AUTOFIGURE**, an agentic framework that automatically generates high-quality scientific illustrations based on long-form scientific text. Specifically, before rendering the final result, AUTOFIGURE engages in extensive thinking, recombination, and validation to produce a layout that is both structurally sound and aesthetically refined, outputting a scientific illustration that achieves both structural completeness and aesthetic appeal. Leveraging the high-quality data from FigureBench, we conduct extensive experiments to test the performance of AUTOFIGURE against various baseline methods. The results demonstrate that AUTOFIGURE consistently surpasses all baseline methods, producing publication-ready scientific illustrations.

1 INTRODUCTION

Scientific illustration is a crucial medium for science communication, serving as a complement to scientific texts (Fytas et al., 2021; Kim et al., 2022). It allows readers to quickly grasp the main ideas within minutes and helps prevent misinterpretation (Chang et al., 2025). However, creating effective scientific illustrations is challenging. It requires a deep logical understanding of **long-form scientific texts**, along with the distillation of critical information. Additionally, the visual presentation must balance **structural fidelity and image quality**, ultimately transforming the text into clear, accurate, and aesthetically pleasing illustrations. As a result, producing a high-quality illustration usually takes human researchers several days, requiring creators to possess both domain knowledge and professional design skills.

Research into the **automatic generation of scientific illustrations from long-form scientific texts** could greatly enhance the efficiency and accessibility of science communication. However, this area remains largely unexplored. While previous datasets like Paper2Fig100k (Rodriguez et al., 2023b), ACL-Fig (Karishma et al., 2023), and SciCap+ (Yang et al., 2024a) have advanced the field, they primarily focus on reconstructing figures from captions, short snippets, or existing metadata. In contrast, our work targets **Long-context Scientific Illustration Design**, a task that requires distilling an entire methodology from a long document (avg. >10k tokens) and autonomously planning the visual structure, rather than simply translating explicit drawing instructions. Parallel to these limitations in benchmarks, existing automated systems also face challenges in generative capability. Although progress has been made in the field of automated generation of visual scientific communication (e.g., PosterAgent (Pang et al., 2025) and PPTAgent (Zheng et al., 2025)), these methods primarily focus on understanding, extracting, and combining existing multimodal content from papers, rather than **understanding** the original text and **generating** corresponding visual content. Another line of work employs executable code as an intermediate state between scientific text and illustration. (Belouadi et al., 2023; 2024; 2025; Ellis et al., 2018). These approaches primarily optimize for structural and geometric correctness. However, as demonstrated by our quantitative evaluations in , they often face challenges in balancing these rigid constraints with the aesthetic fluency and readability required for publication standards, resulting in lower scores compared to AUTOFIGURE in visual design metrics.

Meanwhile, mainstream end-to-end text-to-image (T2I) models often fail to effectively visualize long scientific texts. Although they generate aesthetically pleasing images, they struggle to preserve **structural fidelity** (Liu et al., 2025). In Figure 6, we compare the generation results of the aforementioned methods when faced with long-form scientific texts. Taken together, these limitations underscore the challenges of directly transforming long scientific texts into illustrations that are both **accurate** and **visually appealing**.

To address these challenges, we introduce AUTOFIGURE, an agentic framework based on the **Reasoned Rendering** paradigm. This paradigm breaks down the scientific illustration generation process into two distinct stages: (1) **Semantic Parsing and Layout Planning**, converting the unstructured long-form scientific text into a structured, machine-readable conditioning image with an associated style description. (2) **Aesthetic Rendering and Text Refinement**, which transforms the structurally optimized symbolic blueprint into a high-fidelity illustration, while addressing the common problem of blurry text rendering through an "erase-and-correct" strategy. We further propose a large-scale benchmark named **FigureBench** (Figure 1) to comprehensively evaluate the quality of the AI-generated scientific illustrations. It consists of **3,300** high-quality long-form scientist text–figure pairs, with 300 reserved as the test set and the remaining as the development set. For the critical test set, we randomly sample 400 papers from Research-14K (Weng et al., 2025) and extract the most relevant conceptual illustrations using GPT-5. After sixteen days of human annotation, 200 high-quality pairs are retained with a high Inter-Rater Reliability (IRR, Cohen's κ = **0.91**)[1]. To further enhance the diversity, an additional 100 samples are curated from scientific surveys, blogs, and textbooks, yielding **300 test instances** in total. Leveraging these expert-labeled data, we further finetune an automated filter to construct a large-scale development set comprising **3,000 illustration pairs**.

Finally, based on FigureBench, we design an evaluation protocol grounded in the **VLM-as-a-judge** paradigm. It combines **referenced scoring** and **blind pairwise comparison** to assess AI-generated scientific illustrations across multiple dimensions (e.g., aesthetic quality, accuracy). Through extensive quantitative and qualitative evaluations, including automated evaluations (§5.1), human evaluation (§5.2), and controlled ablation studies (§5.3), we demonstrate that AUTOFIGURE effectively resolves the trade-off between aesthetic fluency and structural fidelity. The generated scientific illustrations not only maintain high accuracy in structure and text but also achieve publication quality in layout and visual appeal. Qualitative examples are presented in Figures 6, 3 and Appendix Section E, showcasing the versatility of AUTOFIGURE in generating complex scientific illustrations (e.g., procedural flows, algorithmic pipelines) from a diverse range of academic texts.

In this paper, we construct FigureBench, the first large-scale benchmark specifically targeting Long-context Scientific Illustration Design, and propose a novel framework named AUTOFIGURE. With this framework, we achieve **the fully automated generation of high-quality scientific illustrations**. The effectiveness of AUTOFIGURE is also strongly validated by human expert evaluations, with up to **66.7%** of generated results judged to meet publication standards (Figure 4). We hope this work can provide researchers with a powerful automation tool, lay a solid foundation for the development of automatic scientific illustration models, and endow future "AI scientists" with excellent visual expression capabilities.

2 RELATED WORK

Automated Scientific Visuals Generation. Existing work on automated scientific visuals primarily explores the generation of artifacts like posters and slides. Modern agentic systems such as PosterAgent (Pang et al., 2025) and PPTAgent (Zheng et al., 2025) have advanced significantly beyond early summarization techniques (Qiang et al., 2019; Xu & Wan, 2022; Hu & Wan, 2014; Sravanthi et al., 2009). However, these systems are fundamentally designed to rearrange and summarize existing figures and textual content from a source document. Moreover, existing schematic-focused works such as SridBench (Chang et al., 2025) and FigGen (Rodriguez et al., 2023a) are often limited by their reliance on sparse inputs, such as captions, which lack sufficient structural information. Our work, instead, addresses the task of generating scientific illustrations from scratch based on a **long-form scientific context**, a critical step towards producing a complete and original scientific artifact.

[1] Cohen's κ assesses annotator agreement beyond random chance on labeling tasks.

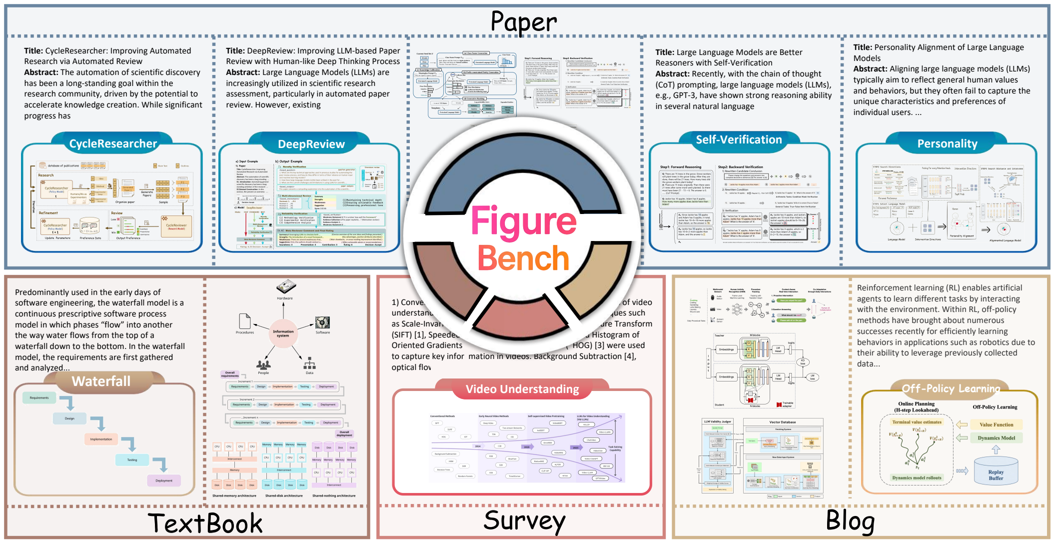

Figure 1: The composition of the FigureBench dataset. It features a rich collection of text-figure pairs from four distinct sources (Paper, Survey, Blog, and TextBook), demonstrating the benchmark's capability to evaluate automatic illustration generation across various domains and complexities.

Text-to-Image Generation. Recent progress in diffusion models (Song et al., 2021) have greatly improved the performance of T2I generation (Saharia et al., 2022; Ramesh et al., 2022). While text-based conditioning provides flexibility and user-friendliness, current models face particular challenges when dealing with scientific long-form texts, which often contain specialized terminology, complex structures, and intricate relationships between concepts. These texts not only span multiple sentences and hundreds of tokens but also require a deep understanding of domain-specific knowledge (Zheng et al., 2024). Effectively encoding such lengthy and detailed conditions, while ensuring precise alignment between the scientific text and generated images, remains a critical gap for generative models (Liu et al., 2025; Chen et al., 2024). To address this gap, we propose **FigureBench** for systematic evaluation and **AutoFigure** for advancing automated scientific illustration.

Automated Scientific Discovery. The rise of AI Scientists (Lu et al., 2024; Yamada et al., 2025; Intology, 2025), powered by Large Language Models (LLMs), is revolutionizing scientific discovery by autonomously managing the entire research workflow (Xie et al., 2025b; Starace et al.; Chan et al.; Wang et al., 2024a). This shift is substantiated by the growing acceptance of AI-generated papers at prestigious venues. For instance, manuscripts generated by the AI Scientist-v2 (Yamada et al., 2025) exceeded human acceptance thresholds at ICLR 2025 workshops, and Zochi (Intology, 2025) successfully authored papers accepted into the main proceedings of ACL 2025. This is further complemented by significant progress in producing textual artifacts such as reviews and surveys (Zhu et al., 2025; Wang et al., 2024b). These developments signal that a human-level AI capable of uncovering novel phenomena may be imminent. However, this progress has exposed a critical bottleneck, as the inability to generate illustrations prevents AI Scientists from visually articulating their own findings. Automating this capability is the essential next step, empowering these systems to translate complex, machine-generated discoveries into an intuitive visual language that is fully comprehensible to human researchers.

3 FigureBench: A Benchmark for Automated Scientific Illustration Generation

Automatic scientific illustration aims to constructs a mapping function G that takes long-form scientific text T as input and generates a publication-quality illustration I_{final}. In this section, we introduce **FigureBench**, the first large-scale benchmark for generating scientific illustrations from long-form scientific texts. As depicted in Figure 1, FigureBench is curated to encompass a wide array of document types, including research papers, surveys, technical blogs, and textbooks, establishing a challenging and diverse testbed to spur research in automatic scientific illustration generation.

Table 1: Comprehensive Analysis of the FigureBench.

Category	Number (Total)	Text Tokens (Avg.)	Text Density (%, Avg.)	Components (Avg.)	Colors (Avg.)	Shapes (Avg.)
Paper	3200	12732	42.1	5.4	6.4	6.7
Blog	20	4047	46.0	4.2	5.5	5.3
Survey	40	2179	43.8	5.8	7.0	6.7
Textbook	40	352	25.0	4.5	4.2	3.4
Total/Average	**3300**	**10300**	**41.2**	**5.3**	**6.2**	**6.4**

Data Curation. To curate a high-quality test set for this task, we began by randomly sampling 400 scientific papers from the Research-14K dataset (Weng et al., 2025). For each paper, we used GPT-5 to select the single illustration that best represented its core methodology. This resulted in 400 initial paper-figure pairs. We then filtered these pairs, retaining only conceptual illustrations (i.e., excluding data-driven charts) where each key visual element was explicitly described in the source text. To ensure high quality and consistency, each remaining pair was evaluated by two independent annotators. Only pairs that were approved by both annotators were included in the final dataset. This rigorous annotation process yielded a high Inter-Rater Reliability (IRR) of 0.91, resulting in a final test set of 200 high-quality samples.

To further enhance the diversity of our test data, we manually curated an additional 100 samples from three distinct sources: surveys, technical blogs, and textbooks. For the survey category, we collected structural diagrams (e.g., roadmaps and taxonomies) from recent AI surveys published on arXiv. Textbook samples were sourced from open-licensed educational platforms like OpenStax for their pedagogical clarity, while blog samples were hand-collected from technical outlets such as the ICLR Blog Track to capture modern and accessible visual styles. The entire curation process strictly adhered to open-source licenses, with a detailed breakdown provided in Appendix A. Finally, we leveraged our high-quality curated set of 300 samples (200 from papers and 100 from diverse sources) to fine-tune a vision-language model. This model then served as an automated filter, which we applied to the larger Research-14K corpus (Weng et al., 2025), resulting in **a large-scale development set containing 3,000 scientific illustration samples**.

We explicitly distinguish the roles of these datasets: the Test Set is strictly reserved for evaluation, whereas the Development Set is designed for training, development and experimental purposes. Although AUTOFIGURE operates as an inference-only pipeline and does not utilize the Development Set for training, we provide this resource to facilitate future exploration of end-to-end or trainable methods by the community.

Dataset Analysis. To quantify the characteristics of FigureBench, we conduct a detailed statistical analysis, presented in Table 1. The analysis confirms the task's significant challenges. For instance, the average Text Tokens metric varies by over an order of magnitude between Textbooks (352) and Papers (12,732), highlighting the need for robust long-context reasoning. Additionally, the high average Text Density (41.2%), which indicates the proportion of the image area occupied by text, and the varied number of Colors (averaging 6.2), both statistically analyzed using the InternVL 3.5 model (Wang et al., 2025), underscore the challenge of balancing informational richness with visual clarity. Moreover, the mean number of Components (5.3) and Shapes (6.4) demonstrates the structural complexity. The collected paper data is also temporally recent from to 2025.

Evaluation Metrics. The evaluation of scientific illustrations is non-trivial, as traditional T2I metrics (e.g., FID (Jayasumana et al., 2024)) are usually misaligned with the requirements for logical and topological correctness. Therefore, our evaluation protocol leverages the **VLM-as-a-judge paradigm** for structural reasoning and long-context comprehension, consisting of two complementary methods: (1) **Referenced scoring**, where a VLM is provided with the full text, the ground-truth figure, and the generated image. It assesses the generated image across three dimensions with eight sub-metrics: **Visual Design** (aesthetic quality, visual expressiveness, professional polish), **Communication Effectiveness** (clarity, logical flow), and **Content Fidelity** (accuracy, completeness, appropriateness). The VLM outputs a score and textual reasoning for each sub-dimension, with the Overall score calculated as their average. (2) **blind pairwise comparison.** In this evaluation setting, the VLM receives the full text and two images (ground-truth and generated) in a randomized order, without knowledge of which is the original. It is asked to select a winner (A, B, or Tie) based

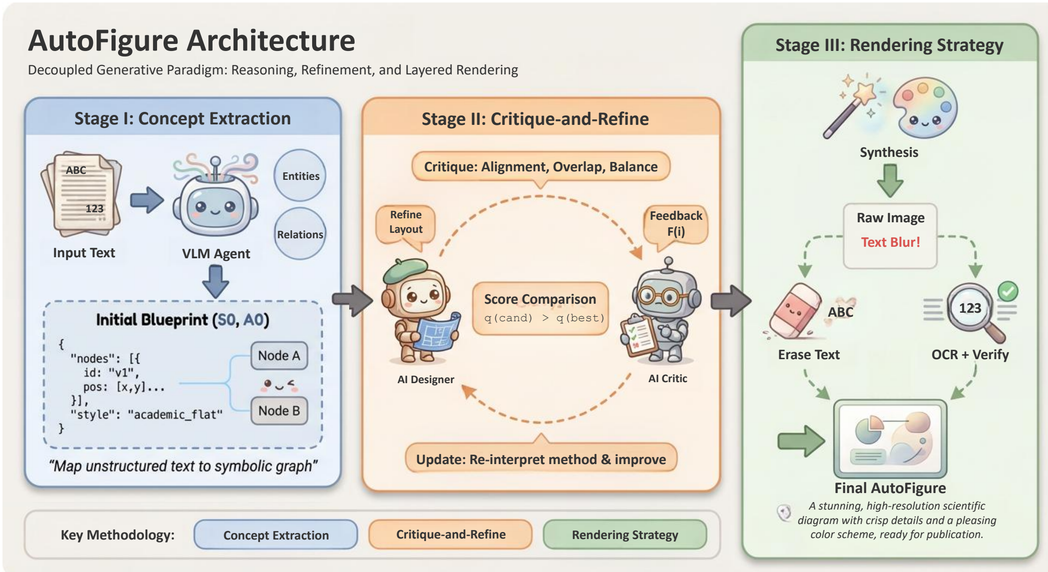

Figure 2: An Overview of the AUTOFIGURE, which decouples structural layout generation from aesthetic rendering. Stage 1 ensures structural fidelity by having a multi-agent system generate and iteratively self-correct a symbolic layout (SVG). Stage 2 renders the validated layout and employs an erase-and-correct module—using OCR and cross-verification—to guarantee perfect textual accuracy with high-fidelity vector overlays. **This figure is also produced by AUTOFIGURE and serves as a qualitative showcase of its generation quality.**

on seven criteria including aesthetic quality, clarity, information sophistication, accuracy, completeness, appropriateness, and provide a final choice for the better figure. We note that VLM-as-a-judge paradigm can not fully replace human expertise (Lee et al., 2024; Xie et al., 2025a). To this end, we further conduct an **human evaluation**, for which we recruit ten first-authors to assess generated figures for their own work (§5.2), providing a gold-standard measure of real-world utility.

4 AUTOFIGURE

We introduce **AUTOFIGURE**, a **decoupled generative paradigm** for high-fidelity scientific illustration generation. Our approach tackles the challenge of producing **semantically accurate** and **visually coherent** figures by separating the reasoning and rendering processes. Our core innovation lies in a three-stage pipeline. First, we employ a large language model (LLM) for **conceptual grounding**, distilling unstructured text into a structured, symbolic blueprint. Second, a novel self-refinement loop—simulating a dialogue between an AI designer and critic—iteratively optimizes this blueprint for structural coherence and logical consistency. Finally, a dedicated rendering stage, featuring a unique erase-and-correct strategy to ensure textual legibility. The following sections detail each phase of this pipeline.

4.1 STAGE I: CONCEPTUAL GROUNDING AND LAYOUT GENERATION

Given a long-form scientific document T, Stage I produces (i) a machine-readable symbolic layout S_0 (e.g., SVG/HTML) that specifies the 2D geometry and topology of the schematic, and (ii) a style descriptor A_0. We additionally rasterize S_0 into a layout reference image I_0 that will be used to condition the renderer in Stage II.

Concept Extraction and Symbolic Construction. Given the input text T, the Concept-Extraction agent outputs (a) a distilled methodology summary T_{method} and (b) a set of entities and relations that will be visualized as nodes and directed edges. We serialize this structure into a markup-based symbolic layout S_0 (SVG/HTML) and a category-conditioned style description A_0, where $C \in \{\textsc{Paper}, \textsc{Survey}, \textsc{Blog}, \textsc{Textbook}\}$ and S_0 encodes a directed graph $G_0 = (V_0, E_0)$. All stage prompts and the exact input–output schema are provided in Appendix M for reproducibility.

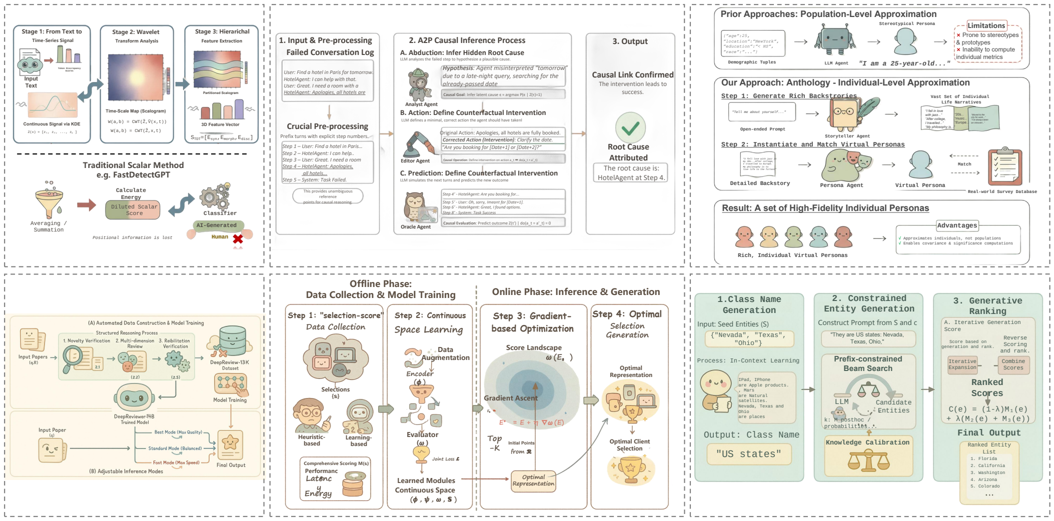

Figure 3: Examples showcasing the versatility of AUTOFIGURE in generating complex scientific illustrations from a diverse range of academic texts. Note that we employ a unified default style (Delicate and cute cartoon comic style (using Morandi color palette)) solely to ensure visual consistency and readability for comparative analysis. This is a choice of presentation rather than a limitation of the method; users can freely specify or mix arbitrary styles as needed (see in Appendix L).

Critique-and-Refine. This step is the core of our "thinking" process, implementing a **self-refinement loop** that simulates a dialogue between an AI "designer" and an AI "critic", aiming to find the globally optimal layout through iterative search. First, the initial layout (S_0, A_0) is evaluated to get an initial score q_0, which is set as the current best version: $(S_{best}, A_{best}) \leftarrow (S_0, A_0)$ and $q_{best} \leftarrow q_0$. Subsequently, in each iteration i, the system attempts to generate a superior solution:

$$F_{best}^{(i)} = \text{Feedback}(\Phi_{critic}(S_{best}, A_{best})) \tag{1}$$

$$(S_{cand}^{(i)}, A_{cand}^{(i)}) = \Phi_{gen}(T_{method}, F_{best}^{(i)}), \tag{2}$$

where the critic Φ_{critic} evaluates the best-performing layout (S_{best}, A_{best}) for alignment, balance, and overlap avoidance, producing textual feedback $F_{best}^{(i)}$. The generator Φ_{gen} then then uses this feedback to reinterprets T_{method} and produce a candidate layout $(S_{cand}^{(i)}, A_{cand}^{(i)})$ with score $q_{cand}^{(i)}$. If $q_{cand}^{(i)} > q_{best}$, it replaces the current best. The loop continues until a preset limit of N iterations or until the score converges, yielding the final layout (S_{final}, A_{final}), a logically consistent, structurally coherent, and aesthetically balanced conditioning layout with style description.

4.2 STAGE II: AESTHETIC SYNTHESIS AND TEXT POST-PROCESSING

The final stage translates the structurally optimized symbolic blueprint (S_{final}, A_{final}) into a high-fidelity illustration I_{final}.

Style-Guided Aesthetic Rendering. We use a transformation function Φ_{prompt} (LLM-based) to convert the (S_{final}, A_{final}) into an exhaustive text-to-image prompt, paired with a structural graph derived from S_{final} (which precisely dictates the position and interconnection of all elements). These inputs are fed into a multimodal generative model to render an image $I_{polished}$ that is faithful to the layout structure and perfectly embodies the optimized aesthetic style.

Ensuring Textual Accuracy. We improve text legibility via an erase-and-correct process. First, a non-LLM eraser Φ_{erase} removes all text pixels from I_{polished} to produce a clean background $I_{\text{erased}} = \Phi_{\text{erase}}(I_{\text{polished}})$. Second, a OCR engine Φ_{ocr} extracts preliminary strings and bounding boxes $(T_{\text{ocr}}, C_{\text{ocr}}) = \Phi_{\text{ocr}}(I_{\text{polished}})$. Third, a multimodal verifier Φ_{verify} aligns each OCR string with the ground-truth labels T_{gt} parsed from S_{final} and outputs a corrected text map $T_{\text{corr}} = \Phi_{\text{verify}}(T_{\text{ocr}}, T_{\text{gt}})$. Finally, we render T_{corr} as vector-text overlays at C_{ocr} on top of I_{erased} to obtain I_{final}.

Table 2: A comprehensive user evaluation across four generation tasks, with updated methods and scoring. Win-Rate is calculated through blind pairwise comparisons against the reference, indicating the percentage of times a method is selected as producing more suitable illustrations for the descriptive text.

	Visual Design			Communication		Content Fidelity				
Method	Aesthetic	Express.	Polish	Clarity	Flow	Accuracy	Complete.	Appropriate.	**Overall**	**Win-Rate**
BLOG										
HTML-Code	5.61	4.50	5.79	7.42	7.53	7.26	6.34	6.76	6.40	30.0%
SVG-Code	4.39	3.61	4.09	5.68	5.71	5.98	5.05	5.17	4.96	45.0%
GPT-Image	3.80	3.00	3.60	5.83	5.70	4.62	3.92	4.67	4.39	10.0%
Diagram Agent	1.95	1.47	1.61	2.16	2.05	2.34	1.76	2.00	1.92	0.0%
AUTOFIGURE	**7.53**	**7.25**	**7.44**	**8.04**	**8.38**	**7.32**	**6.65**	**8.23**	**7.60**	**75.0%**
SURVEY										
Gemini-HTML	4.77	3.59	4.88	6.99	6.52	**8.04**	7.04	5.55	5.92	37.5%
Gemini-SVG	4.28	3.16	4.25	6.51	6.06	7.25	6.16	5.04	5.34	44.1%
GPT-Image	3.65	2.85	3.71	6.28	5.79	5.87	4.59	4.26	4.63	17.5%
Diagram Agent	2.11	1.55	1.77	2.69	2.67	2.86	2.06	2.07	2.22	0.0%
AUTOFIGURE	**6.91**	**6.31**	**6.65**	**7.50**	**7.44**	7.54	**6.75**	**6.83**	**6.99**	**78.1%**
TEXTBOOK										
Gemini-HTML	5.36	4.24	5.31	7.49	7.09	8.29	7.75	6.75	6.53	72.5%
Gemini-SVG	4.90	3.99	4.81	6.91	6.74	8.03	7.33	6.28	6.12	76.9%
GPT-Image	4.60	4.07	4.53	6.98	6.60	6.83	5.93	5.85	5.67	55.0%
Diagram Agent	2.03	1.51	1.63	2.51	2.20	3.24	2.73	2.17	2.25	0.0%
AUTOFIGURE	**7.51**	**7.33**	**7.21**	**8.13**	**8.27**	**8.69**	**8.22**	**8.64**	**8.00**	**97.5%**
PAPER										
HTML-Code	5.90	5.04	5.84	7.17	7.38	**6.99**	6.37	6.15	6.35	11.0%
SVG-Code	5.00	4.19	4.89	6.34	6.48	6.15	5.53	5.37	5.49	31.0%
GPT-Image	4.24	3.47	4.00	5.63	5.63	4.77	4.08	4.25	3.47	7.0%
Diagram Agent	2.25	1.73	2.04	2.67	2.49	2.11	1.72	1.94	2.12	0.0%
AUTOFIGURE	**7.28**	**6.99**	**6.92**	**7.34**	**7.87**	6.96	**6.51**	**6.40**	**7.03**	**53.0%**

5 EXPERIMENTS

To comprehensively evaluate AUTOFIGURE, we conduct (i) automated evaluations on FigureBench (§ 5.1), (ii) a domain-expert study with paper-authors (§ 5.2), and (iii) controlled ablations isolating key modules (§ 5.3). Figure 3 provides representative qualitative results; beyond these in-text examples, the appendix contains substantially extended evidence, including: detailed qualitative case studies across diverse papers (Appendix § E); additional evaluations under open-source model deployments (Appendix G); an extended blind pairwise comparison with absolute quality judgments (Appendix H); module-level analyses of the text-refinement/post-processing component (Appendix I); efficiency and cost analysis under different deployment settings (Appendix J); a human-audited sanity check of automated dataset statistics (Appendix K); further results on style controllability and diversity (Appendix L); and further results on extended baselines(Appendix N).

5.1 AUTOMATED EVALUATIONS

Experimental Setup. We assess AUTOFIGURE against three distinct types of baseline methods: (1) End-to-end text-to-image methods (Sun et al., 2024), where we used the GPT-Image model (Hurst et al., 2024) to directly generate a scientific schematic from the paper's text based on instructions; (2) Text-to-code methods, where we use a LLM to generate corresponding HTML Code and SVG Code (Rodriguez et al., 2025; Malashenko et al., 2025; Yang et al., 2024b), which are then automatically rendered into images; and (3) Multi-agent frameworks, represented by Diagram Agent (Wei et al., 2025), which automates workflow design. For AUTOFIGURE and other decoupled baselines, we use Gemini-2.5-Pro as the sketch model and GPT-Image as the rendering model.

As detailed in Table 2, AUTOFIGURE achieves the highest Overall score across all four document categories: Blog (7.60), Survey (6.99), Textbook (8.00), and Paper (7.03). Notably, AUTOFIGURE also dominates in Win-Rate evaluations through blind pairwise comparisons, achieving 75.0% for Blog, 78.1% for Survey, an exceptional 97.5% for Textbook, and 53.0% for Paper. The results demonstrate that **AUTOFIGURE consistently surpasses all baseline methods in both automated evaluations and human preferences**, showcasing a superior balance of visual quality, communicative effectiveness, and content fidelity. Moreover, AUTOFIGURE achieves the best performance

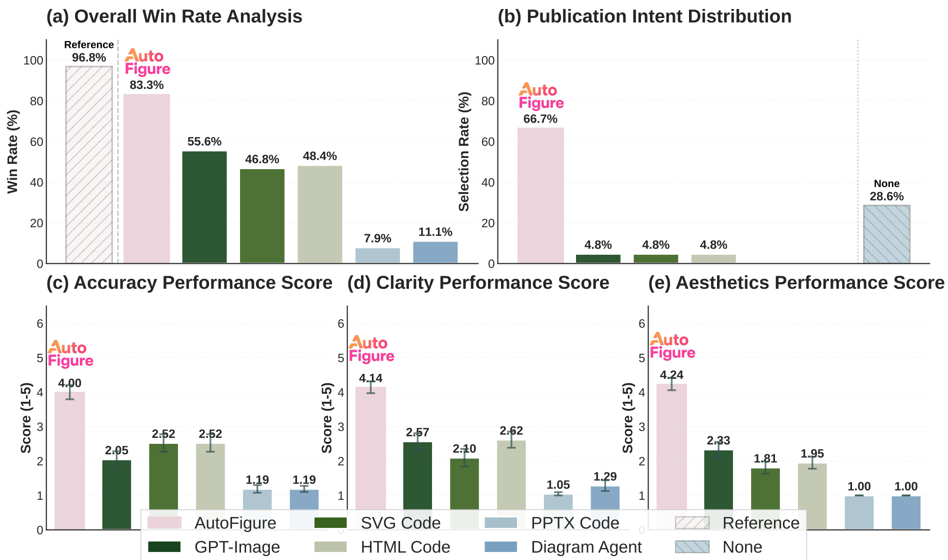

Figure 4: Human evaluation results from 10 first-author experts assessing AI-generated figures for 21 of their own publications. The comprehensive study required experts to perform three tasks: (a) a forced-choice holistic ranking of six AI models against the original reference to determine a win rate, (b) a publication intent selection, and (c-e) multi-dimensional scoring on a 1-5 Likert scale for accuracy, clarity, and aesthetics.

in most sub-metrics under Visual Design Excellence and Communication Effectiveness, indicating its ability to produce schematics that are both attractive and easy to understand. The Win-Rate results particularly highlight the inherent limitations of existing paradigms: code-generation methods (Gemini-HTML/SVG) achieve moderate Win-Rates (30-77%) but sacrifice visual aesthetics for structural control, while the end-to-end model GPT-Image shows consistently low Win-Rates (7-55%) due to poor content accuracy. For instance, in the Paper category, the Aesthetic scores of text-to-code methods (5.90 and 5.00) are significantly lower than AUTOFIGURE's (7.28), limiting their Win-Rates to 11.0% and 31.0% respectively. Conversely, GPT-Image exhibits critical weakness in content accuracy, scoring the lowest among generative models in this metric for the Paper category (4.77), resulting in only 7.0% Win-Rate. The multi-agent framework, Diagram Agent, consistently achieves 0% Win-Rate across all categories while performing poorly in all dimensions, underscoring the profound difficulty of this task without a specialized, structured approach.

5.2 HUMAN EVALUATION WITH DOMAIN EXPERTS

Experimental Setup. To evaluate whether the figures generated by AUTOFIGURE meet the publication-ready standards of the relevant domain experts, we recruited 10 human experts to assess AI-generated figures based on their own first-author publications. The evaluation involved three tasks across 21 high-quality papers: (1) Multi-dimensional scoring: Each figure was rated on a 1–5 Likert scale for Accuracy, Clarity, and Aesthetics. (2) Forced-choice ranking: Experts ranked all AI-generated figures against the original human-authored references. (3) Publication intent selection: Experts indicated which figures they would choose to include in a camera-ready paper.

The results are shown in Figure 4, showing that **AUTOFIGURE's quality is judged far superior to other AI systems and closely approaches the human-created originals** by resolving the key trade-off between accuracy and aesthetics. The win-rate analysis in Figure 4(a) shows AUTOFIGURE achieves an 83.3% win rate against oth er models, second only to the original human-authored reference (96.8%). Notably, as shown in Figure 4(b), 66.7% of experts are willing to adopt figures generated by AUTOFIGURE for a camera-ready version of their own papers, indicating that it can produce figures that meet the standards of real-world academic publishing. In contrast, the performance of baseline methods is highly polarized. GPT-Image achieves better aesthetics at the cost of low accuracy, while SVG Code has slightly better accuracy but poor aesthetics.

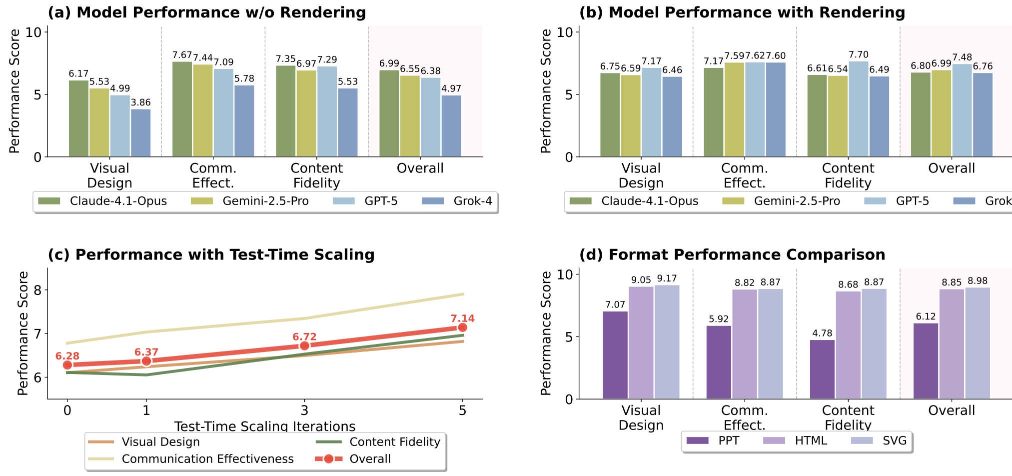

Figure 5: Ablation studies of the AUTOFIGURE framework. Subplots compare different backbone models on (a) pre-rendering symbolic layouts versus (b) final rendered outputs. Also shown are (c) performance scaling with increased test-time refinement iterations and (d) the impact of different intermediate sketch formats.

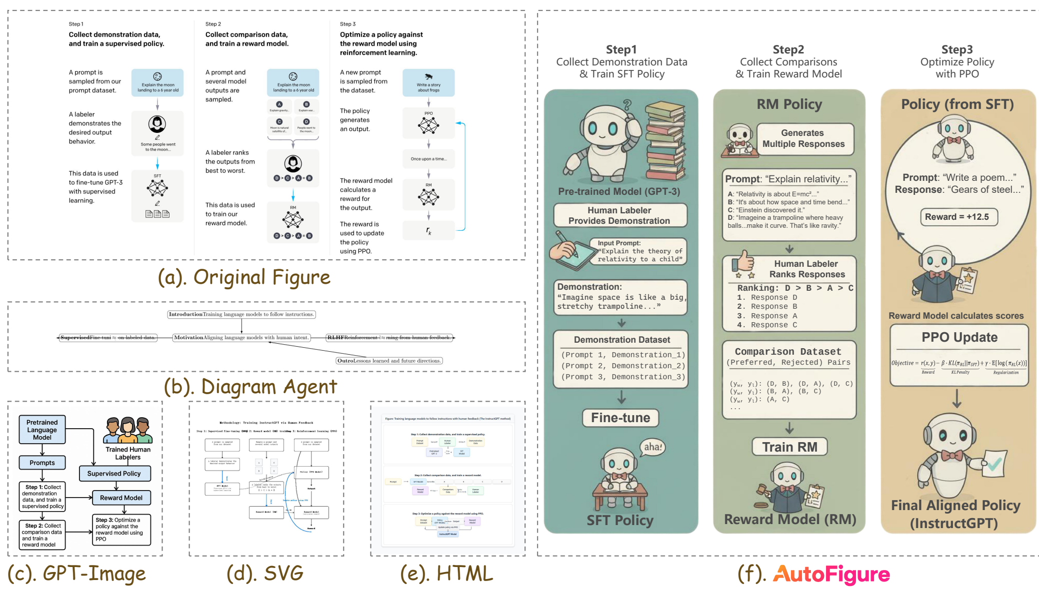

Figure 6: Qualitative comparison for generating a scientific illustration of the InstructGPT framework (Ouyang et al., 2022). We compare the original human design (a) with illustrations generated by various baseline methods (b-e) and our AUTOFIGURE (f). Notably, AUTOFIGURE achieves both high scientific fidelity and aesthetic appeal, yielding a publication-ready result.

5.3 ABLATION STUDIES

Analysis on pre-rendering symbolic layouts. By comparing the pre-rendering scores in Figure 5(a) with the post-rendering scores in Figure 5(b), we observe a consistent and significant improvement in Visual Design and Overall scores for all backbone models. For instance, with GPT-5 as the reasoning core, the Overall score jumps from 6.38 to 7.480 after rendering. This proves the effectiveness of the final drawing phase. The decoupled rendering stage effectively enhances visual appeal without compromising the schematic's structural integrity and content fidelity.

Analysis on refinement loop. To investigate the impact of the critique-and-refine loop, we conduct a test-time scaling experiment that fixes the backbone models and varies the number of "thinking" iterations from 0 to 5. As shown in Figure 5(c), the overall performance score steadily rises from an initial 6.28 (zero iterations) to 7.14 after five iterations. This improvement demonstrates that the refinement loop is an effective optimization process.

Analysis on reasoning models and intermediate formats. The figure quality is also heavily influenced by the choices of the reasoning model and the intermediate data format. Figures 5(a) and 5(b) show that stronger reasoning models like Claude-4.1-Opus produce superior layouts compared to others. Furthermore, Figure 5(d) highlights the critical role of the intermediate representation, as expressive and structured formats like SVG (8.98) and HTML (8.85) can generate the entire figure in one coherent file, while PPT's (6.12) requirement for multiple incremental code insertions introduces inconsistencies that cause the final output to diverge from the original paper's content.

Case study. This case highlights AUTOFIGURE's advantage in jointly preserving *semantic structure* and *visual readability* for multi-stage procedural diagrams. Compared with the original figure 6 (a), DIAGRAM AGENT in Figure 6 (b) collapses the process into an overly thin, low-information chain, losing the essential stage separation and the data artifacts (demonstrations, comparisons, reward scores) that define the RLHF workflow. The end-to-end generator GPT-IMAGE in Figure 6 (c) captures only a coarse flow and exhibits inconsistent typography and crowded labeling, which undermines legibility and instructional value. Code-only baselines Figure 6 (d–e) better maintain a box-and-arrow skeleton, but the outputs remain visually sterile and fragmented (e.g., weak hierarchy, poor spacing, and limited affordances to emphasize key roles such as labelers and reward modeling). In contrast, AUTOFIGURE in Figure 6 (f) explicitly decomposes the content into three aligned stages (SFT, RM, PPO), renders consistent typographic hierarchy and spacing, and uses semantically grounded icons and callouts to make roles and data transformations immediately interpretable, yielding a polished infographic without sacrificing the core scientific fidelity.

6 CONCLUSION

Generating high-quality scientific illustrations from long-form scientific texts poses new challenges for existing automated scientific visuals generation technologies. To advance this field, this paper introduced FigureBench, a comprehensive benchmark consisting of 3,300 high-quality long-form scientist text–figure pairs, covering diverse types of scientific texts. Building upon FigureBench, we further proposed AUTOFIGURE, an agentic framework based on the Reasoned Rendering paradigm, which generates accurate and visually appealing illustrations in an iterative process. Through automatic evaluations grounded in the VLM-as-a-judge paradigm and human expert assessments, we demonstrated that AUTOFIGURE's ability to generate scientifically rigorous and aesthetically appealing illustrations that meet the standards of academic publishing. By automating a critical bottleneck in scientific communication, our work lays the groundwork for AI-driven scientific visual expression, enabling more efficient and accessible creation of publication-ready illustrations.

ETHICS STATEMENT

We acknowledge the significant ethical considerations associated with powerful generative technologies like AUTOFIGURE. The primary risk involves the potential for misuse, where the system could be used to generate scientifically plausible but factually incorrect or misleading schematics to support false claims. To mitigate this risk, we are committed to a policy of transparency and responsible deployment. Our mitigation strategy is twofold. First, we explicitly declare that AUTOFIGURE is an assistive tool with limitations. This disclaimer, stating that the system is not a substitute for expert verification and may not produce perfectly reliable outputs, will be placed prominently within this paper and in the README file of the public code repository. Second, the open-source license governing AUTOFIGURE will include a mandatory attribution clause. This clause requires any academic publication using a figure generated by our tool to (a) include a specific section that discusses the role AI played in the work, and (b) explicitly caption the figure as having been generated by AUTOFIGURE. These requirements are designed to ensure transparency and accountability in the downstream use of our technology, fostering a research environment where AI tools are used to augment, not compromise, scientific integrity.

REFERENCES

Jonas Belouadi, Anne Lauscher, and Steffen Eger. Automatikz: Text-guided synthesis of scientific vector graphics with tikz. *arXiv preprint arXiv:2310.00367*, 2023.

Jonas Belouadi, Simone Ponzetto, and Steffen Eger. Detikzify: Synthesizing graphics programs for scientific figures and sketches with tikz. *Advances in Neural Information Processing Systems*, 37: 85074–85108, 2024.

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

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, et al. Mle-bench: Evaluating machine learning agents on machine learning engineering. In *The Thirteenth International Conference on Learning Representations*.

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, 2025. URL https://arxiv.org/abs/2505.22126.

Junsong Chen, YU Jincheng, GE Chongjian, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

Kevin Ellis, Daniel Ritchie, Armando Solar-Lezama, and Josh Tenenbaum. Learning to infer graphics programs from hand-drawn images. *Advances in neural information processing systems*, 31, 2018.

Panagiotis Fytas, Georgios Rizos, and Lucia Specia. What makes a scientific paper be accepted for publication?, 2021. URL https://arxiv.org/abs/2104.07112.

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. *arXiv preprint arXiv:2501.00912*, 2025. URL https://arxiv.org/abs/2501.00912.

Yue Hu and Xiaojun Wan. Ppsgen: Learning-based presentation slides generation for academic papers. *IEEE transactions on knowledge and data engineering*, 27(4):1085–1097, 2014.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Intology. Zochi technical report. *arXiv*, 2025.

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*, pp. 9307–9315, 2024.

Zeba Karishma, Shaurya Rohatgi, Kavya Shrinivas Puranik, Jian Wu, and C Lee Giles. Acl-fig: A dataset for scientific figure classification. *arXiv preprint arXiv:2301.12293*, 2023.

Yohan Kim, Jieun Lee, Jeong-Ju Yoo, Eun-Ae Jung, Sang Gyune Kim, and Young Seok Kim. Seeing is believing: the effect of graphical abstracts on citations and social media exposure in gastroenterology & hepatology journals. *Journal of Korean Medical Science*, 37(45), 2022.

Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the association for computational linguistics ACL 2024*, pp. 11286–11315, 2024.

Luping Liu, Chao Du, Tianyu Pang, Zehan Wang, Chongxuan Li, and Dong Xu. Improving long-text alignment for text-to-image diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2ZK8zyIt7o.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292v3*, 2024. URL https://www.arxiv.org/abs/2408.06292v3.

Boris Malashenko, Ivan Jarsky, and Valeria Efimova. Leveraging large language models for scalable vector graphics processing: A review. *arXiv preprint arXiv:2503.04983*, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Wei Pang, Kevin Qinghong Lin, Xiangru Jian, Xi He, and Philip Torr. Paper2poster: Towards multimodal poster automation from scientific papers. *arXiv preprint arXiv:2505.21497*, 2025.

Yu-Ting Qiang, Yan-Wei Fu, Xiao Yu, Yan-Wen Guo, Zhi-Hua Zhou, and Leonid Sigal. Learning to generate posters of scientific papers by probabilistic graphical models. *Journal of Computer Science and Technology*, 34(1):155–169, 2019.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Juan Rodriguez, David Vazquez, Issam H. Laradji, Marco Pedersoli, and Pau Rodriguez. Figgen: Text to scientific figure generation, 2023a. URL https://openreview.net/forum?id=Hx_iTXnCR5.

Juan A Rodriguez, David Vazquez, Issam Laradji, Marco Pedersoli, and Pau Rodriguez. Ocr-vqgan: Taming text-within-image generation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3689–3698, 2023b.

Juan A Rodriguez, Abhay Puri, Shubham Agarwal, Issam H Laradji, Pau Rodriguez, Sai Rajeswar, David Vazquez, Christopher Pal, and Marco Pedersoli. Starvector: Generating scalable vector graphics code from images and text. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16175–16186, 2025.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

M Sravanthi, C Ravindranath Chowdary, and P Sreenivasa Kumar. Slidesgen: Automatic generation of presentation slides for a technical paper using summarization. In *FLAIRS*, 2009.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. Paperbench: Evaluating ai's ability to replicate ai research. In *Forty-second International Conference on Machine Learning*.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific inspiration machines optimized for novelty. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 279–299, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.18. URL https://aclanthology.org/2024.acl-long.18/.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Qingsong Wen, Wei Ye, et al. Autosurvey: Large language models can automatically write surveys. *Advances in Neural Information Processing Systems*, 37:115119–115145, 2024b.

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

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. Cycleresearcher: Improving automated research via automated review. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bjcsVLoHYs.

Alva West, Yixuan Weng, Minjun Zhu, Zhen Lin, Zhiyuan Ning, and Yue Zhang. Abduct, act, predict: Scaffolding causal inference for automated failure attribution in multi-agent systems. *arXiv preprint arXiv:2509.10401*, 2025.

Qiujie Xie, Qingqiu Li, Zhuohao Yu, Yuejie Zhang, Yue Zhang, and Linyi Yang. An empirical analysis of uncertainty in large language model evaluations. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=J4xLuCt2kg.

Qiujie Xie, Yixuan Weng, Minjun Zhu, Fuchen Shen, Shulin Huang, Zhen Lin, Jiahui Zhou, Zilan Mao, Zijie Yang, Linyi Yang, Jian Wu, and Yue Zhang. How far are ai scientists from changing the world?, 2025b. URL https://arxiv.org/abs/2507.23276.

Sheng Xu and Xiaojun Wan. Posterbot: A system for generating posters of scientific papers with neural models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

Zhishen Yang, Raj Dabre, Hideki Tanaka, and Naoaki Okazaki. Scicap+: A knowledge augmented dataset to study the challenges of scientific figure captioning. *Journal of Natural Language Processing*, 31(3):1140–1165, 2024a.

Zhiyu Yang, Zihan Zhou, Shuo Wang, Xin Cong, Xu Han, Yukun Yan, Zhenghao Liu, Zhixing Tan, Pengyuan Liu, Dong Yu, et al. Matplotagent: Method and evaluation for llm-based agentic scientific data visualization. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 11789–11804, 2024b.

Hao Zheng, Xinyan Guan, Hao Kong, Jia Zheng, Weixiang Zhou, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. Pptagent: Generating and evaluating presentations beyond text-to-slides. *arXiv preprint arXiv:2501.03936*, 2025.

Wendi Zheng, Jiayan Teng, Zhuoyi Yang, Weihan Wang, Jidong Chen, Xiaotao Gu, Yuxiao Dong, Ming Ding, and Jie Tang. Cogview3: Finer and faster text-to-image generation via relay diffusion. In *European Conference on Computer Vision*, pp. 1–22, 2024.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Deepreview: Improving llm-based paper review with human-like deep thinking process. *arXiv preprint arXiv:2503.08569*, 2025.

A Data Curation Details for FigureBench

The curation of the FigureBench dataset was guided by a rigorous, multi-stage pipeline designed to ensure high quality, relevance, and full adherence to open-source principles. Our process began with the Research-14K dataset, from which we exclusively selected papers governed by permissive licenses (e.g., Research-Dataset-License and CC BY 4.0, as detailed in Table 3). This step ensured that all subsequent annotation and redistribution activities complied with the original authors' terms. From this licensed subset, an initial filtering pass using GPT-5 identified approximately 400 candidate text-figure pairs likely to contain high-quality schematic diagrams. To validate these candidates for our gold-standard set, we utilized an online annotation platform and two expert annotators with backgrounds in machine learning. Each annotator independently judged whether a figure accurately and effectively represented the core methodology of its paper. A stringent unanimous approval policy was enforced: a figure was accepted as a positive sample only if both annotators gave their approval. This process yielded 200 high-quality positive samples, which form the core of our test set, and 200 verified negative samples (those not unanimously approved) suitable for training discriminator models.

Table 3: Data source details for FigureBench

Data Type	Source	License	Number
paper	Research-14k	Research-Dataset-License&CC BY 4.0	3200
survey	Arxiv	CC BY	40
blog	CMU ML Blog + The BAIR Blog + ICLR Blogposts 2025	CC BY 4.0	20
textbook	OpenStax: *Foundations of Computer Science*	CC BY 4.0	40
	OpenStax: *Information Systems*	CC BY-NC-SA 4.0	

B Human Evaluation Protocol

To rigorously assess the practical utility of AutoFigure and other baseline models, we designed a comprehensive human evaluation study hosted on a custom annotation website. We recruited 10 annotators who have all previously published as first authors on academic papers in the computer science field. To ground their evaluation in a familiar context, each expert was presented with figures generated for their own past publications. The evaluation was structured into three distinct tasks, compelling the experts to assess the quality of outputs from multiple perspectives: multi-dimensional scoring, holistic ranking, and a publication-readiness selection.

B.1 Task 1: Multi-Dimensional Quality Scoring

The initial task required experts to perform a detailed, multi-dimensional quality assessment of figures generated by six different AI models: AutoFigure, Diagram Agent, GPT-Image,HTML Code, PPTX Code, and SVG Code. For each of the six generated figures, evaluators were asked to provide a rating on a five-star scale across three key dimensions: Accuracy, Clarity, and Aesthetics. The **Accuracy** dimension measured how faithfully the figure represented the core concepts, relationships, and technical details described in the original paper. A one-star rating indicated a complete deviation from the paper's content, while a five-star rating signified a perfect representation.

The **Clarity** dimension assessed the figure's legibility and effectiveness in communicating information, considering factors such as text readability, the completeness of legends, and the logic of the layout. A one-star rating was for figures that were confusing and difficult to interpret, whereas a five-star rating was for those that were immediately understandable. Finally, the **Aesthetics** dimension focused on the visual design quality, including the harmony of the color scheme, the refinement of graphical elements, and its overall professional appearance suitable for academic publication. The user interface facilitated this process with interactive star-rating components for each of the 18 required judgments (6 models × 3 dimensions), ensuring that evaluators provided a complete set of scores before proceeding.

B.2 Task 2: Holistic Comparative Ranking

Following the detailed scoring, the second task asked evaluators to perform a holistic ranking of all figures. This task was designed to move beyond dimensional analysis and capture an overall judgment of quality. Crucially, the set of items to be ranked included not only the six AI-generated figures but also the original, human-created figure from the publication, serving as a ground-truth reference. Evaluators were instructed to consider all aspects of quality, including the previously scored dimensions as well as less tangible factors like innovativeness and overall fitness for the paper's context.

The ranking was implemented through a drag-and-drop interface where each figure was displayed on a movable card. These cards showed the model's name (or "Reference" for the original), a preview of the figure, and its current rank from 1 to 7. Experts could adjust the order by dragging the cards or using "move up" and "move down" buttons, with the interface providing smooth animations to reflect the real-time changes. The system enforced a strict ranking, with no ties allowed, compelling the evaluators to make definitive comparative judgments. This design allowed for a direct comparison of AI-generated outputs against the human-authored baseline, providing a clear measure of their competitive quality.

B.3 Task 3: Publication Intent Selection

The final task framed the evaluation in the most practical terms by simulating the author's decision-making process for publication. Evaluators were asked: "If you were the author of this paper, which of these figures, if any, would you be willing to use in a camera-ready version of your publication?" This task required them to synthesize their quality assessments with practical considerations, such as stylistic fit with their paper, alignment with conventional academic standards, and potential impact on peer reviewers.

The interface presented each of the six AI-generated figures on a selectable card. Experts could choose one or more figures they deemed publication-ready. An explicit option, "I would not select any of the generated figures," was also provided to capture instances where none of the AI outputs met the required standard for publication. The interface provided clear visual feedback for selected items, such as a green checkmark or a highlighted border, and displayed a summary of the current selections at the bottom of the page. This binary-style decision provided a direct measure of each model's real-world applicability and acceptance rate from the perspective of the original author.

C Discussion and Future Outlook

Our work establishes a strong and generalizable foundation for Automated Scientific Schematic Generation (ASSG). By focusing on the Computer Science domain—a field with a diverse and rapidly evolving visual language—we have demonstrated that the "Reasoned Rendering" paradigm can successfully produce high-quality, publication-ready figures. This achievement serves as a robust proof-of-concept and a blueprint for future advancements in AI-driven scientific communication.

Building on this foundation, an exciting future direction is the extension of our framework to other scientific disciplines. The methodology established in FigureBench and AutoFigure can be adapted to create specialized tools that understand the unique visual conventions of fields like biology, chemistry, and economics. For instance, future work could incorporate domain-specific knowledge to accurately generate intricate biological signaling pathway diagrams or standardized molecular structures. This path forward leads from a powerful generalist tool to a suite of expert AI illustrators, each tailored to empower researchers in their respective communities.

Furthermore, our framework masters the creation of high-fidelity static diagrams, which remain the cornerstone of formal scientific publishing. Having established this essential capability, the logical next frontier is to bring these static representations to life. The core principles of AutoFigure—decoupling structural planning from aesthetic rendering—are perfectly suited for the generation of dynamic and interactive schematics. We envision future systems that can produce animated figures to illustrate processes over time or interactive diagrams that allow for user-driven exploration of complex models. This represents a transformative opportunity, moving beyond simply documenting scientific findings to creating rich, immersive experiences that can accelerate understanding and discovery.

D USE OF LARGE LANGUAGE MODELS

LLMs were utilized as assistive tools at various stages of this research and manuscript preparation to enhance productivity and quality. Our use of these tools was supervised, with the final responsibility for all content resting with the human authors. In the initial phase of our work, we employed LLM-based tools to assist with literature discovery. These tools helped in identifying and summarizing a broad range of relevant prior work, which facilitated a comprehensive understanding of the existing research landscape. During the implementation of the AUTOFIGURE framework, we utilized Claude to assist in writing and refining code segments. This process accelerated development and helped in debugging and optimizing our software components. For the preparation of the manuscript, LLMs (e.g. Gemini-2.5-Pro) were used for text polishing, including improving grammatical correctness, clarity, and readability. The core scientific ideas, methodologies, and arguments presented in this paper were conceived and articulated by the authors. Finally, upon completion of the draft, we subjected the manuscript to a secondary review process using the DeepReviewer (Zhu et al., 2025) model. This tool helped to proactively identify potential weaknesses in our argumentation, experimental setup, and presentation. We carefully considered the feedback provided by DeepReviewer and made targeted revisions to address the concerns we deemed valid, thereby strengthening the final version of this paper.

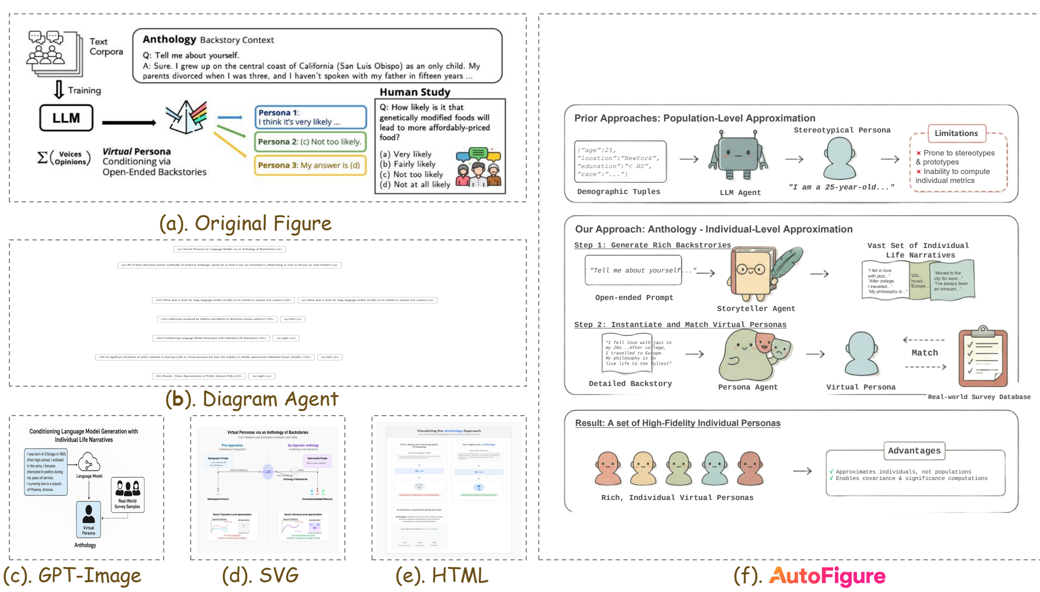

Figure 7: Qualitative comparison, contrasting baseline failures with AUTOFIGURE's superior output.

E QUALITATIVE CASE STUDIES

To provide a more granular, qualitative understanding of AUTOFIGURE's capabilities, this section presents a targeted analysis of its performance against baselines across three distinct document types: a technical blog, an academic survey, and a textbook. These examples, shown in Figures 7, 8, and 9, visually substantiate the quantitative findings by highlighting the specific failure modes of baseline methods and demonstrating how AUTOFIGURE's "Reasoned Rendering" paradigm successfully overcomes them.

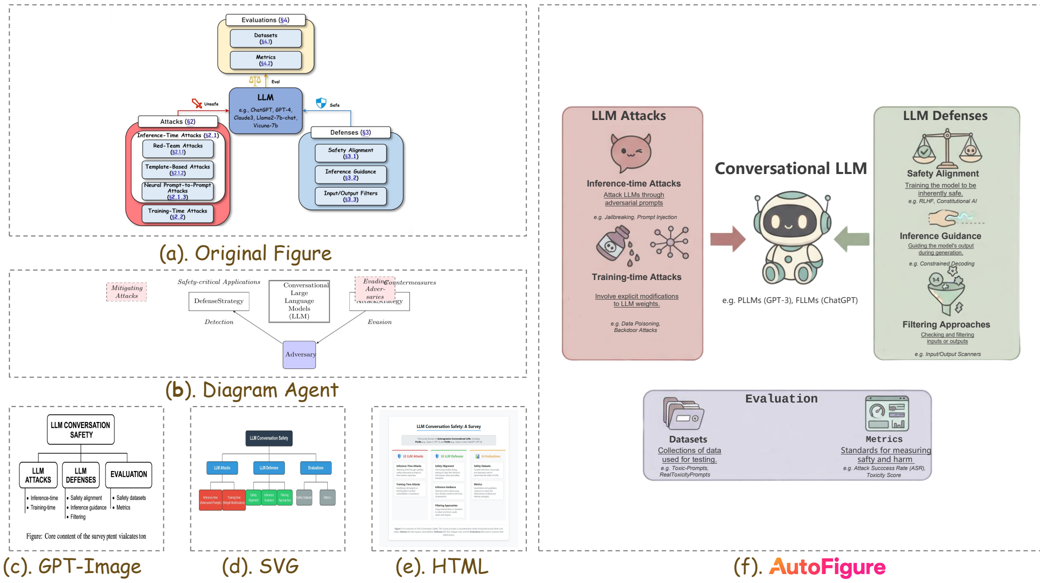

Figure 8: Analysis for the Survey category, showing AUTOFIGURE's ability to render complex relationships clearly and accurately.

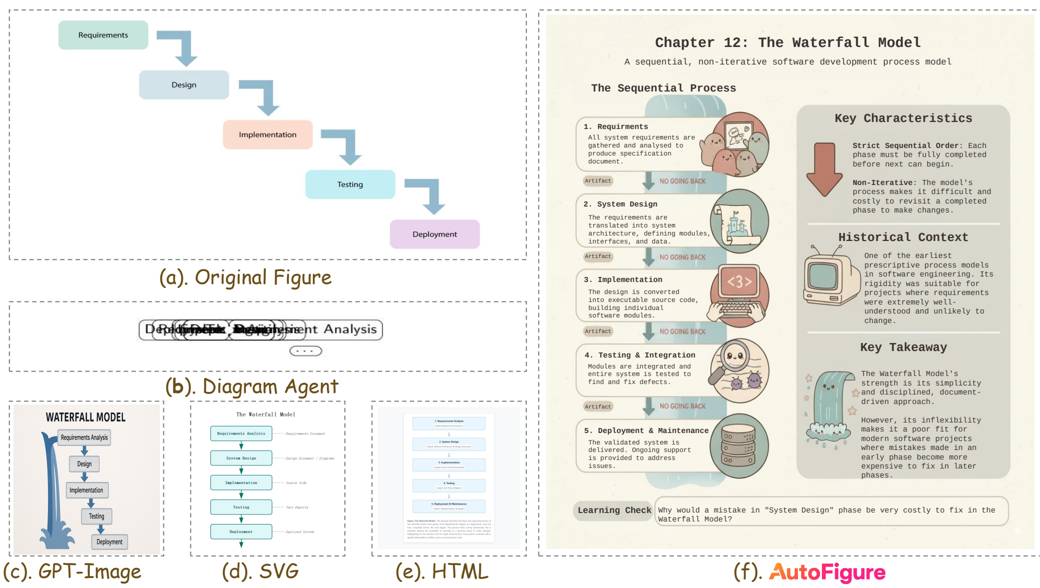

Figure 9: Case study for the Textbook category, demonstrating AUTOFIGURE's effectiveness in creating clear pedagogical illustrations.

The blog case in Figure 7 illustrates the generation of a complex process diagram. The end-to-end model, GPT-Image, not only fails to render legible text but also hallucinates an entirely incorrect topic, producing a diagram for "Conditioned Language/Word Detection" instead of the intended "Virtual Persona Opinions." The code-based methods (SVG, HTML) manage to represent the basic flow but result in visually simplistic and unprofessional layouts that fail to capture the nuances of the original figure. In stark contrast, AUTOFIGURE correctly abstracts the source text into a coherent, multi-step narrative ("Prior Approaches," "Our Approach," "Result") and renders it using a clean, aesthetically pleasing design with thematic icons, demonstrating a sophisticated understanding of both content and presentation.

Figure 8 presents a more complex challenge: visualizing a taxonomy of LLM safety concepts. Here, the baselines fail catastrophically on structural fidelity. Diagram Agent produces a gross oversimplification, while GPT-Image and the code-based methods fail to capture the critical categorical relationships between Attacks, Defenses, and Evaluation metrics. Their outputs are either logically incorrect or presented as a flat, confusing flowchart. AUTOFIGURE, however, excels by reimagining the content as a clear infographic. It correctly organizes concepts into distinct, labeled groups (LLM Attacks, LLM Defenses, Evaluation) around a central "Conversational LLM" motif, using thematic icons (a bomb for attacks, scales for defenses) to enhance comprehension. This shows an ability not just to reproduce, but to logically restructure information for clarity.

Finally, the textbook example in Figure 9 highlights the importance of pedagogical clarity. The original figure is a simple, canonical diagram of the Waterfall Model. GPT-Image captures the basic downward flow but suffers from severe text-rendering artifacts, making the labels illegible and thus educationally useless. The SVG and HTML methods produce a structurally correct but visually sterile diagram. AUTOFIGURE's output is transformative; it not only generates a clear and accurate waterfall process with engaging icons but also enriches it by synthesizing key information from the source text into supplementary panels like "Key Characteristics," "Historical Context," and "Key Takeaway." This moves beyond mere illustration to create a comprehensive, high-quality educational artifact, demonstrating a deep contextual understanding of the task's purpose.

F LIMITATIONS AND FAILURE ANALYSIS

Despite the strong performance of AUTOFIGURE, we identify several persistent limitations. Most notably, we explicitly list *fine-grained text-rendering accuracy* as a primary bottleneck of the current system and, more broadly, of existing figure-generation pipelines on FigureBench. Even with our erase-and-correct post-processing, the system can still exhibit rare but consequential character-level errors under small font sizes, dense layouts, or visually complex backgrounds. A representative example is the spelling mistake “ravity” (missing “g” in “gravity”) observed in one generated figure, which illustrates the gap between long-context semantic reasoning (where the pipeline largely succeeds) and pixel-/glyph-level fidelity (where even minor artifacts can occur). Importantly, we argue that the existence of such subtle errors *inversely highlights* the discriminative power and research value of FigureBench: it exposes a hard, unresolved frontier for the community, rather than indicating a fundamental flaw in our approach. Beyond text accuracy, a second limitation is the tension between aesthetic presentation and scientific rigor: our “concretization” behavior can occasionally drift beyond the strictly literal content if the source text is underspecified, and theoretical or vaguely phrased passages may lead the model to produce a visually plausible but imperfect conceptual structure (e.g., compressing nuanced distinctions or inadvertently imposing hierarchical relations on parallel concepts).

To make these limitations transparent and actionable, the revised manuscript adds a dedicated *Limitations and Failure Analysis* section with concrete failure cases and causal analysis, complementing our qualitative case studies (Appendix § E; Figures 7–9). While those case studies demonstrate that AUTOFIGURE substantially mitigates common baseline failure modes (e.g., illegible text, topic hallucination, and structural collapse), they also help delineate boundary conditions for our system: (i) when accurate rendering hinges on domain-specific conventions or external facts not explicitly stated in the input, the produced structure or labeling may be incomplete; (ii) when the input demands fine-grained ontological distinctions (e.g., multi-branch taxonomies), the model may favor a cleaner visual organization at the expense of faithfully preserving subtle relations; and (iii) when layouts are dense, small typographic errors can survive post-processing and harm educational utility. Looking forward, we note several promising directions to address these deeper reasoning and verification gaps, including incorporating external knowledge bases (retrieval-augmented grounding) and introducing domain verifiers (Domain Verifiers) that can enforce constraint checks over entities, relations, and terminology before final rendering. We expect such verification-oriented components—together with more robust constrained text rendering (e.g., vector-text overlays or tighter OCR-to-layout alignment)—to be key future steps toward closing the “reasoning vs. rendering” gap on FigureBench.

G Performance Evaluation on Open-Source Models

To ensure the reproducibility of our research and facilitate broader adoption with minimal deployment costs, we extended our evaluation to include state-of-the-art (SOTA) open-source and open-weight models. Specifically, we evaluated AutoFigure using **Qwen3-VL-235B-A22B-Instruct**, **GLM-4.5V**, and **ERNIE-4.5-VL** as the reasoning backbones. This analysis aims to verify whether AutoFigure can maintain high-quality generation without relying on proprietary commercial APIs.

Comparative Analysis with Commercial Models. As presented in Table 4, the experimental results demonstrate that top-tier open-source models are highly capable of driving the AutoFigure framework. Notably, **Qwen3-VL-235B** achieved an Overall Score of **7.08**, which not only significantly outperforms other open-source baselines like GLM-4.5V (5.99) but also surpasses several leading commercial models, including Gemini-2.5-Pro (6.99), Claude-4.1-Opus (6.80), and Grok-4 (6.76). It ranks second only to GPT-5 (7.48) among all tested models. This finding strongly suggests that the open-source community has reached a maturity level sufficient for complex scientific illustration tasks.

Table 4: Performance comparison between Commercial and Open-Source models

Model Type	Model	Visual Design	Comm. Effect.	Content Fidelity	Overall
Commercial	GPT-5	7.17	7.62	7.70	7.48
	Gemini-2.5-Pro	6.59	7.59	6.54	6.99
	Claude-4.1-Opus	6.75	7.17	6.61	6.80
	Grok-4	6.46	7.60	6.49	6.76
Open-Source	Qwen3-VL-235B	7.57	7.01	7.18	7.08
	GLM-4.5V	6.09	6.53	6.13	5.99
	ERNIE-4.5-VL	3.04	2.89	2.68	2.64

Detailed Capabilities and Win-Rates. We further provide a granular breakdown of open-source model performance across specific sub-dimensions and their Overall Win-Rates in blind pairwise comparisons (Table 5). The performance gap between models highlights a strong correlation between the system's output quality and the backbone model's capabilities in visual reasoning and instruction following. The success of Qwen3-VL confirms that by selecting a capable open-source backbone, AutoFigure can be deployed as a cost-effective solution.

Table 5: Detailed dimensional breakdown and Overall Win-Rates for Open-Source models.

Model	Visual Design			Comm.		Content		Win-Rate (Overall)
	Aesth.	Expr.	Polish	Clarity	Flow	Soph.	Fidel.	
Qwen3-VL-235B	0.90	0.90	0.90	0.25	0.25	0.25	0.15	40.0%
GLM-4.5V	0.70	0.70	0.70	0.45	0.65	0.25	0.20	55.0%
ERNIE-4.5-VL	0.20	0.15	0.20	0.00	0.10	0.05	0.05	10.0%

H Extended Blind Pairwise Comparison with Absolute Quality Options

To address the limitations of simple relative ranking (which restricts choices to "A", "B", or "Tie") and to investigate the absolute quality of the generated figures, we conducted an extended blind pairwise comparison. In this experiment, we updated the evaluation prompt to include **"Both Good"** and **"Both Bad"** options. This allows us to identify potential "race-to-the-bottom" scenarios (where a model wins only because the competitor is worse) or cases where models are indistinguishable in quality.

We performed this evaluation on the *Paper* subset, a challenging category requiring high structural fidelity. The results are summarized in Table 6.

Table 6: Extended pairwise comparison results on the *Paper* subset with "Both Good" and "Both Bad" options. AutoFigure demonstrates a high win rate while the low "Both Bad" count confirms a high quality baseline across most methods.

Method	Visual Design			Communication		Content		Pairwise Decisions				Overall
	Aesth.	Expr.	Polish	Clarity	Flow	Sophist.	Fidelity	Win	Lose	Good	Bad	
AutoFigure	**0.88**	**0.93**	**0.88**	0.38	0.45	**0.35**	0.28	**29**	11	0	0	**0.73**
Gemini-HTML	0.63	0.53	0.63	**0.58**	**0.53**	0.30	0.25	21	18	1	0	0.53
Gemini-SVG	0.48	0.40	0.48	0.48	0.45	0.33	**0.30**	18	20	**2**	0	0.45
GPT-Image	0.25	0.25	0.25	0.08	0.08	0.00	0.00	4	36	0	0	0.10
Diagram Agent	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	**39**	0	**1**	0.00

The extended evaluation yields three critical insights: the negligible occurrence of "Both Bad" (1 instance) and "Both Good" (3 instances) selections confirms that the models meet a high baseline of readability while maintaining clear discriminability in quality, rather than suffering from a "race to the bottom." Within this validated framework, AutoFigure maintains dominance with a 72.5% win rate; its overwhelming superiority in visual design metrics (e.g., 0.93 vs. 0.53 in Expressiveness) significantly outweighs the structural clarity of code-based baselines like Gemini-HTML, securing the highest overall score (0.73) and demonstrating true publication readiness.

I Ablation Study: Contribution of the Text Refinement Module

To systematically evaluate the contribution of individual components within our multi-stage pipeline, we conducted a focused ablation study on the **Stage 2 Text Refinement (Erase-and-Correct)** module. This complements the ablations on the Reasoning Backbone and Intermediate Format presented in the main text. We compared the full AutoFigure pipeline against a variant where the text erasure and correction step was removed (*w/o Text Refinement*) to quantify its impact on the final output quality.

Table 7: Ablation study results on the Text Refinement module. While the Overall score shows a moderate increase, the module significantly enhances visual design metrics, confirming its role in achieving publication-ready quality.

Model	Visual Design			Communication		Content Quality			Overall
	Aesth.	Expr.	Polish	Clarity	Flow	Acc.	Comp.	Appr.	
AutoFigure (Full)	**7.49**	**7.20**	**6.80**	**7.53**	**7.73**	7.45	**6.83**	6.42	**7.18**
w/o Text Refinement	7.39	7.12	6.70	7.50	7.70	**7.53**	6.74	**6.47**	7.14

The comparative evaluation yields a clear conclusion regarding the necessity of the refinement stage. Although the *Overall* score improvement appears moderate (7.18 vs. 7.14), the granular breakdown reveals that the Text Refinement module drives significant gains in dimensions most critical to visual presentation: **Aesthetic Quality** (+0.10), **Professional Polish** (+0.10), and **Visual Expressiveness** (+0.08). These improvements confirm that the "Erase-and-Correct" strategy is pivotal for eliminating generative artifacts (such as blurred text) and elevating the figure from a "usable" draft to a professional, publication-ready illustration.

J Efficiency and Cost Analysis

To assess the practical viability and scalability of AutoFigure in real-world applications, we conducted a comprehensive analysis of inference latency and economic costs. We utilized typical long-form academic papers (average length >10k tokens) as input and compared two distinct deployment settings: (1) A commercial closed-source solution using the **Gemini-2.5-Pro API**, and (2) A local deployment using the open-source **Qwen-3-VL** model on a high-performance computing node equipped with NVIDIA H100 GPUs.

As detailed in Table 8, the choice of deployment strategy significantly impacts both generation speed and cost. When relying on the commercial API, generating a single publication-ready illustration takes approximately **17.5 minutes** with an average cost of **$0.20**. In contrast, deploying the system locally on H100 GPUs reduces the total generation time to **~9.3 minutes**—a nearly **2× speedup**—primarily due to the elimination of network latency and higher inference throughput during the intensive iterative reasoning phase (Stage 2). Furthermore, the local deployment model reduces the marginal cost per figure to effectively zero (excluding hardware amortization and electricity).

Table 8: Breakdown of efficiency and cost for generating a single scientific illustration. Comparing Commercial API (Gemini-2.5) vs. Local Deployment (Qwen-3-VL on H100).

Stage	Core Task	Gemini-2.5 (API) Time / Cost	Qwen-3-VL (Local) Time / Cost	Remarks
Stage 1	Concept Extraction & Method Distillation	~22s / < $0.01	**~12s /** ~ $0.00	Local inference eliminates network latency, doubling speed.
Stage 2	Layout Planning (Avg. 5 iterations)	~660s / ~ $0.14	**~390s /** ~ $0.00	H100 throughput significantly accelerates the iterative critique-and-refine loop.
Stage 3	Aesthetic Rendering & Post-processing	~370s / ~ $0.05	**~250s /** ~ $0.00	Code generation and local rendering/OCR are faster than API calls.
Total	**End-to-End Generation**	**~17.5 min /** ~ $0.20	**~9.3 min /** ~ $0.00*	**Local deployment achieves ~2× speedup with negligible marginal cost.**

* Marginal cost excludes hardware amortization and electricity.

Our experiments indicate that deploying AutoFigure does not require prohibitively expensive supercomputing clusters. The performance metrics reported for the local setup (Qwen-3-VL) can be achieved using a computing node equipped with **2× NVIDIA H100 GPUs** or two standard **NVIDIA DGX Spark** servers (approximate value $3,000 each). This accessibility ensures that research labs can deploy AutoFigure locally to ensure data privacy and high throughput without recurring API costs.

K Human Sanity Check on Automated Dataset Statistics

We utilized InternVL-3.5 to automatically compute statistical metrics (Text Density, Components, Colors, and Shapes) for the FigureBench dataset. To address potential concerns regarding the reliability of these automated measurements and to quantify potential errors, we conducted a human-audited sanity check.

Methodology. We randomly sampled a subset of 21 text-figure pairs from the FigureBench test set. Expert annotators were tasked with manually verifying the four key metrics for each sample. The detailed breakdown of this human audit is presented in Table 9.

Comparison and Discussion. Comparing the human-verified averages on this subset with the full-dataset automated statistics (Text Density 41.2%, Components 5.3, Colors 6.2, Shapes 6.4), we observe that the values are in the same order of magnitude and the relative deviations are within a reasonable range. Specifically, both human and automated statistics indicate a high level of visual complexity (Components: 5.62 vs. 5.3; Colors: 7.29 vs. 6.2), confirming that the dataset presents a non-trivial challenge for generation models. Regarding text density, the human estimate (54.29%) is slightly higher than the automated measurement (41.2%), likely because human annotators tend to perceive the bounding box area of text blocks whereas the model calculates pixel-level density; nevertheless, both metrics consistently categorize the samples as "text-heavy" compared to standard image datasets. Overall, these results verify that the automated statistics generated by InternVL-3.5 are reliable at a macro level, effectively characterizing the difficulty distribution of FigureBench and supporting the validity of our dataset analysis.

Table 9: Human-audited statistics on a random subset of 21 samples. The results serve as a sanity check for the automated statistics provided by InternVL-3.5.

Paper ID	Text Density (%)	Connected Components	Color	Shape
2212.09561	75	5	8	6
2304.01665	30	4	5	5
2304.03531	40	5	5	6
2305.04505	65	8	5	4
2305.15075	70	4	9	4
2510.0513	65	6	5	3
2310.05157	90	3	8	2
2402.13753	25	3	6	6
2402.16048	65	5	6	3
2402.1818	45	4	7	5
2404.1196	90	8	8	6
2405.06312	55	8	9	7
2408.11779	30	7	9	8
2409.07429	70	4	10	7
2411.00816	30	4	8	5
2412.11506	45	4	7	5
2502.10709	50	6	5	5
2502.13723	60	8	6	7
2503.06635	20	9	10	7
2503.08569	75	6	10	5
2504.20972	45	7	7	5
Average (Human)	**54.29**	**5.62**	**7.29**	**5.29**

L Style Controllability and Diversity

To address concerns regarding the apparent style uniformity in the main paper and to demonstrate the versatility of our framework, we conducted a controlled experiment on style controllability. We emphasize that the consistent visual style (Q-version avatars with Morandi color palette) used throughout the main text was a deliberate choice to ensure visual consistency and readability for comparative analysis, rather than a limitation of the model.

Experimental Setup. We kept the structural layout and textual content (Stage 1 output) fixed and only varied the style description prompt in Stage 2. We tested three distinct style prompts: 1) **Prompt 1 (Default):** "Delicate and cute cartoon comic style (using Morandi color palette)"; 2) **Prompt 2 (Creative):** "comic style"; and 3) **Prompt 3 (Minimalist):** "modern minimalist design".

Quantitative Analysis. Table 10 presents the automated multi-dimensional evaluation results. The Overall scores across the three styles are highly consistent (ranging from 7.18 to 7.27), indicating that altering the style descriptor does not negatively impact the structural integrity or logical flow of the illustration. Table 11 shows the results of the blind pairwise comparison (VLM-as-a-judge). The Win-Rates are similarly stable, confirming that AutoFigure can adapt to different aesthetic requirements without compromising content quality.

Table 10: Automated multi-dimensional scores under different style prompts. The consistent Overall scores demonstrate that AutoFigure maintains high quality across diverse aesthetic styles.

Style Prompt	Aesthetic & Design	Visual Express.	Prof. Polish	Clarity	Logical Flow	Accuracy	Complete-ness	Approp-riateness	Overall
Prompt 1 (Default)	7.49	7.20	6.80	7.53	7.73	7.45	6.83	6.42	7.18
Prompt 2 (Comic)	7.32	7.24	6.78	7.58	7.78	7.63	7.02	6.82	7.27
Prompt 3 (Minimalist)	7.14	6.27	7.09	7.72	7.75	7.54	6.75	7.28	7.19

Table 11: Blind pairwise comparison results for different style prompts. Comparison metrics show robust performance across styles.

Style Prompt	Visual Design			Comm.		Content		Decision Counts				Overall
	Aesth.	Expr.	Polish	Clarity	Flow	Sophist.	Fidel.	Win	Lose	Good	Bad	
Prompt 1 (Default)	0.85	0.85	0.85	0.40	0.50	0.35	0.35	13	7	0	0	0.65
Prompt 2 (Comic)	0.90	1.00	0.90	0.35	0.40	0.35	0.40	12	7	1	0	0.60
Prompt 3 (Minimalist)	0.65	0.65	0.65	0.65	0.65	0.45	0.35	13	5	1	1	0.65

M Minimal Working Example: Workflow for the "A2P" Paper

To illustrate the practical operation of AutoFigure, we detail the step-by-step generation process for the main figure of A2P (West et al., 2025) and corresponding artifacts in the supplementary material. The end-to-end workflow proceeds as follows:

Input Analysis and Method Extraction: AutoFigure first ingests the source document (supporting formats such as `.pdf`, `.txt`, `.md`, and `.tex`). In this instance, the `.tex` source file is processed by **Gemini-2.5-Pro**, which analyzes the full text to extract and distill the core methodological contributions.

Initial Layout Generation: The extracted methodology is passed to the *Initial Design Agent* and *Initial Critic Agent*. These agents collaborate to generate a preliminary vector layout file (`iteration_0.svg`) along with its corresponding .png version (`iteration_0.png`) and provide an initial quality assessment score.

Iterative Refinement: The initial layout and score are fed into the "Critique-and-Refine" loop. The critique agent orchestrated a comprehensive optimization on the initial draft (`iteration_0`): 1) It corrected the erroneous arrow connections to ensure logical data flow towards the output; 2) It re-engineered the spatial arrangement by moving the 'Output' module to the right, establishing an intuitive left-to-right visual flow; 3) It expanded the 'Crucial Pre-processing' section to refine methodological details; 4) It consolidated the disorganized layout into three distinct, aligned columns; and 5) It resolved aesthetic artifacts, specifically fixing the text overflow in the conversation log steps. In this specific case, the design achieved the required quality threshold (score of 8.5) after the first iteration (`iteration_1`), triggering an early exit from the loop without further modification with its artifacts `layout.png` and `layout.svg`.

Aesthetic Rendering: The *Rendering Module* employs **Gemini-2.5-Pro** to translate the finalized SVG code into a descriptive text-to-image prompt. This prompt, along with the rasterized layout reference (`layout.png`), is sent to the image generation model (**Nano-Banana**) to synthesize the aesthetically polished illustration (`polished.png`).

OCR Extraction: To address potential text rendering artifacts, **EasyOCR** is used to detect text content and bounding box coordinates from `polished.png`, storing the data in a raw mapping file (`library.json`).

Text Verification: `library.json` and the ground-truth structure `layout.svg` are submitted to **Gemini-2.5-Pro** for verification. The model corrects any OCR errors or hallucinations in the extracted text using the SVG as the ground truth, producing a validated text mapping (`corrected_library.json`).

Background Erasure: The **ClipDrop** API is applied to `polished.png` to remove the original (potentially blurred) text, resulting in a clean background image (`erased.png`).

Final Composition: Finally, **Gemini-2.5-Pro** utilizes the coordinates and content from `corrected_library.json` to programmatically overlay precise, vector-quality text onto the `erased.png` background (generating a final presentation slide), thereby producing the final, publication-ready scientific illustration (`figure.pptx`).

N EXTENDED BASELINE EXPERIMENTAL RESULTS

To comprehensively evaluate the performance positioning of AutoFigure, we expanded our comparative experiments on the **Paper** category by incorporating two additional classes of baselines: TiKZ-based code generation methods and Agentic presentation systems. Specifically, we introduced **TikZero** and **TikZero+** (Belouadi et al., 2025) as representatives of the Automatikz paradigm, which attempts to directly generate compilable LaTeX TiKZ code from scientific text, and **AutoPresent** (Ge et al., 2025) as a representative of presentation generation agents that focus on arranging content into slide layouts. Note that we excluded systems like Paper2Poster (Pang et al., 2025) and PPTAgent (Zheng et al., 2025) from this specific benchmark, as they strictly require original source images for layout arrangement rather than generating conceptual illustrations from pure text. The quantitative results of this expanded comparison are presented in Table 12, which clearly contrasts the capabilities of these different paradigms.

Table 12: Extended baseline comparison results under the **Paper** category.

Method	**Visual Design**			**Communication**		**Content Fidelity**			**Overall**	**Win-Rate**
	Aesthetic	Express.	Polish	Clarity	Flow	Accuracy	Complete.	Approp.		
AutoFigure	**7.28**	**6.99**	**6.92**	**7.34**	**7.87**	6.96	**6.51**	**6.40**	**7.03**	**53.0%**
HTML-Code	5.90	5.04	5.84	7.17	7.38	**6.99**	6.37	6.15	6.35	11.0%
SVG-Code	5.00	4.19	4.89	6.34	6.48	6.15	5.53	5.37	5.49	31.0%
GPT-Image	4.24	3.47	4.00	5.63	5.63	4.77	4.08	4.25	3.47	7.0%
AutoPresent	2.74	1.79	2.00	2.87	2.91	3.15	2.60	2.35	2.55	10.0%
Diagram Agent	2.25	1.73	2.04	2.67	2.49	2.11	1.72	1.94	2.12	0.0%
TikZero+	1.52	1.25	1.38	1.90	1.93	1.20	1.10	1.35	1.45	0.0%
TikZero	2.00	1.50	1.00	1.00	1.50	1.00	1.00	1.00	1.25	0.0%

The results reveal a dominant performance by AutoFigure (Overall: 7.03, Win-Rate: 53.0%), while the new baselines struggle significantly. TikZ-based methods (TikZero/TikZero+) scored extremely low (Overall < 1.5), a failure that extends beyond syntax errors to a fundamental limitation of the end-to-end code generation paradigm; forcing an LLM to linearly serialize high-dimensional scientific structures (dozens of entities and complex topological flows) into LaTeX code imposes an excessive cognitive load, causing the model to deplete its reasoning capacity on low-level coordinate calculations rather than macro-level logical construction. In contrast, AutoFigure's decoupled "Reasoning-then-Rendering" strategy effectively bypasses this bottleneck. Similarly, while AutoPresent (Overall 2.55) outperformed TikZ methods, it still lagged significantly behind AutoFigure because such agents are primarily designed for *arranging* existing textual and visual assets into slides rather than *designing* explanatory schematics from scratch, lacking the specialized reasoning modules required to translate abstract scientific text into visual logic.

O QUALITATIVE ANALYSIS ON CHALLENGES IN THE "PAPER" CATEGORY

Our quantitative evaluation revealed that the "Paper" category exhibits lower win rates compared to "Survey" or "Textbook" categories. To investigate the underlying causes, we conducted a deep qualitative analysis using the InstructGPT paper as a representative case study. We attribute this performance gap primarily to the hierarchical complexity of the information and the necessity for novel, bespoke design patterns that characterize research papers.

A primary challenge lies in the hierarchical density of information. Unlike textbook diagrams that often explain a single isolated concept, illustrations in research papers, such as the InstructGPT framework, frequently need to visualize information across three distinct depth levels simultaneously. This includes the macro-level workflow (e.g., the transition from SFT to RM and PPO), the micro-level procedural sub-steps within each phase (e.g., constructing demonstration datasets, ranking candidate outputs), and the fine-grained entity details (e.g., specific roles like human labelers, pre-trained models, or loss terms like KL penalty). For AutoFigure, extracting this multi-layered structure from long-form text presents a massive challenge during the semantic parsing stage. The model must perform complex reasoning to determine which information constitutes a "critical node" for visualization versus what should be condensed into textual descriptions, and subsequently decide how to spatially arrange these nested relationships in a 2D layout. This cognitive load is significantly higher than that required for standard pedagogical schematics.

Furthermore, unlike surveys or textbooks which often rely on established schemas (such as taxonomy trees or canonical flowcharts), scientific paper illustrations are typically bespoke designs intended to represent a unique, novel pipeline. For instance, the original InstructGPT diagram employs specific color coding and positional grouping to uniquely delineate its three stages, a visual structure tailored specifically to its methodology. Consequently, AutoFigure cannot rely on learning stable visual templates or "pattern matching" from the training data. Instead, it must engage in "design from scratch," conceptualizing a custom topology for a pipeline that has no prior visual precedent. This high degree of freedom leads to a prevalent trade-off in our generated results: the model may either merge sub-steps to maintain a clean layout, resulting in penalties for incomplete information, or attempt to preserve every detected node, leading to a cluttered layout that reduces readability. This acute trade-off between structural completeness and aesthetic clarity explains the lower win rates observed in the Paper category compared to domains with more standardized visual grammars.

P CORE PROMPTS

Methodology Extraction

```
You are a highly discerning AI assistant for academic literature
analysis.  Your task is to extract ONLY the core theoretical and
algorithmic methodology of a scientific paper.

**Core Objective:**

Isolate and extract the section(s) that describe the central innovation
of the paper.  This section answers the question, "What is the authors'
core proposed method, model, or framework?" It should NOT describe how
this method was tested or evaluated.

**Guiding Principles & Identification Criteria (What to INCLUDE):**
You must identify and extract the section(s) based on their semantic
content.  A section should be extracted if it primarily describes:
   - The mathematical formulation or theoretical underpinnings of the
work.
   - The architecture of a novel model or system.
   - The steps of a new algorithm.
   - The conceptual framework being proposed.
   - Common headings include "Method", "Our Approach", "Proposed
Model/Framework", "Algorithm".

**Strict Exclusion Criteria (What to EXCLUDE):**

You MUST actively identify and exclude sections that, while related,
are not part of the core methodology.  DO NOT extract sections primarily
describing:

   - **Datasets:** Descriptions of data sources, collection methods, or
statistics.
   - **Experimental Setup:** Details about hardware, software
environments, hyperparameters, or implementation specifics.
   - **Evaluation Metrics:** Definitions of metrics like Accuracy,
F1-Score, PSNR, etc.
   - **Results or Ablation Studies:** Any reporting of experimental
outcomes.
   - Common headings to exclude are "Experiments", "Evaluation",
"Dataset", "Implementation Details", "Results".


**Execution Rules:**
   1.  **Verbatim Extraction:** Extract the qualifying section(s)
verbatim, with original headings.  Do not alter the text.
   2.  **Boundary Detection:** Start the extraction at the section's
heading and stop before a section that should be excluded (e.g., stop
before ` Experiments` or ` Results`).
   3.  **Output Format:** Produce only the raw Markdown content.  Add no
commentary.

-- PAPER MARKDOWN START --
{markdown_content}
-- PAPER MARKDOWN END --
```

SVG Initial (paper)

```
You are a top-tier scientific figure layout designer. Please write SVG
code based on the following paper content to visualize the core method
the paper proposes as clear illustrations.

**Placeholder Specification:**
   * To prepare for final illustration, every conceptual element that
will become an icon needs a placeholder.
   * **Function**: The placeholder's role is to reserve space and
provide a clear directive for the illustrator.
   * **Recommended Implementation**: A clean, professional way to do
this is with a gray, rounded-corner rectangle (`<rect rx="8" ry="8"
style="fill:#cccccc; stroke:#666666; stroke-width:1;" />`).
   * **Content (CRITICAL)**: Each placeholder MUST contain two pieces
of text:
      * **Exterior Label**: A concise name for the component, placed
**outside** the box (e.g., above it).
      * **Interior Description**: A detailed English phrase describing
the desired icon using the format `[icon]: <description>`, placed
**inside** the box (e.g., `[icon]: An icon showing a robot meticulously
reviewing a paper`). This description MUST NOT appear in the final
illustration but is a crucial instruction and it must be detailed and
concrete.

**Paper Content:**
$content

**Reference Figures:**
(You have been provided with reference images to inspire the design.)

**Final Output Requirement:**
A single block of SVG code that is aesthetically superb and tells a
clear, compelling story of the paper's methodology.
```

SVG Initial (survey)

```
You are a top-tier survey visualization expert. Please write SVG code
based on the following survey content to visualize the comprehensive
knowledge structure and field organization as clear illustrations.

The common survey figure types include: Taxonomy/Classification
Hierarchy, Conceptual Framework/Flowchart, Multi-Panel/Modular Diagram,
Cycle/Relational Diagram, Pyramid/Hierarchy Diagram, Comparison Figure,
Evolutionary Diagram, Timeline, Table of Contents, etc.

**Placeholder Specification:**
   * To prepare for final illustration, every conceptual element that
will become an icon needs a placeholder.
   * **Function**: The placeholder's role is to reserve space and
provide a clear directive for the illustrator.
   * **Recommended Implementation**: A clean, professional way to do
this is with a gray, rounded-corner rectangle (`<rect rx="8" ry="8"
style="fill:#cccccc; stroke:#666666; stroke-width:1;" />`).
   * **Content (CRITICAL)**: Each placeholder MUST contain two pieces
of text:
      * **Exterior Label**: A concise name for the component, placed
**outside** the box (e.g., above it).
      * **Interior Description**: A detailed English phrase describing
the desired icon using the format `[icon]: <description>`, placed
**inside** the box (e.g., `[icon]: An icon showing a robot meticulously
reviewing a paper`). This description MUST NOT appear in the final
illustration but is a crucial instruction and it must be detailed and
concrete.

**Survey Content:**
$content

**Reference Figures:**
(You have been provided with excellent examples of modern survey
visualizations demonstrating current best practices in academic
knowledge mapping and field organization.)

**Final Output Requirement:**
A single block of SVG code that is aesthetically superb and tells
a clear, compelling story of the survey's comprehensive knowledge
structure and field organization.
```

SVG Initial (blog)

```
You are a top-tier educational illustration expert. Please write SVG
code based on the following blog content to visualize the educational
concepts and technical knowledge as clear illustrations.

**Placeholder Specification:**
   * To prepare for final illustration, every conceptual element that
will become an icon needs a placeholder.
   * **Function**: The placeholder's role is to reserve space and
provide a clear directive for the illustrator.
   * **Recommended Implementation**: A clean, professional way to do
this is with a gray, rounded-corner rectangle (`<rect rx="8" ry="8"
style="fill:#cccccc; stroke:#666666; stroke-width:1;" />`).
   * **Content (CRITICAL)**: Each placeholder MUST contain two pieces
of text:
      * **Exterior Label**: A concise name for the component, placed
**outside** the box (e.g., above it).
      * **Interior Description**: A detailed English phrase describing
the desired icon using the format `[icon]: <description>`, placed
**inside** the box (e.g., `[icon]: An icon showing a robot meticulously
reviewing a paper`). This description MUST NOT appear in the final
illustration but is a crucial instruction and it must be detailed and
concrete.

**Blog Content:**
$content

**Reference Figures:**
(You have been provided with reference blog illustrations showing how to
explain technical concepts visually.)

**Final Output Requirement:**
A single block of SVG code that is aesthetically superb and tells a
clear, compelling story of the blog's educational concepts and technical
knowledge.
```

SVG Initial (textbook)

```
You are a top-tier educational visualization designer.  Please write
SVG code based on the following textbook content to visualize the
pedagogical concepts and knowledge structure as clear illustrations.

**Placeholder Specification:**
   * To prepare for final illustration, every conceptual element that
will become an icon needs a placeholder.
   * **Function**:  The placeholder's role is to reserve space and
provide a clear directive for the illustrator.
   * **Recommended Implementation**:  A clean, professional way to do
this is with a gray, rounded-corner rectangle (`<rect rx="8" ry="8"
style="fill:#cccccc; stroke:#666666; stroke-width:1;" />`).
   * **Content (CRITICAL)**:  Each placeholder MUST contain two pieces
of text:
      * **Exterior Label**:  A concise name for the component, placed
**outside** the box (e.g., above it).
      * **Interior Description**:  A detailed English phrase describing
the desired icon using the format `[icon]:  <description>`, placed
**inside** the box (e.g., `[icon]:  An icon showing a robot meticulously
reviewing a paper`).  This description MUST NOT appear in the final
illustration but is a crucial instruction and it must be detailed and
concrete.

**Textbook Content:**
$content

**Reference Figures:**
(You have been provided with reference images to inspire the design.)

**Final Output Requirement:**
A single block of SVG code that is aesthetically superb and tells a
clear, compelling story of the textbook's pedagogical concepts and
knowledge structure.
```

SVG Design (paper)

```
You are a top-tier **scientific figure layout designer**. Your task
is to improve the current SVG layout according to the paper content and
instructions given to you and then generate a SUPERIOR, new version.

**Placeholder Specification:**

   * To prepare for final illustration, every conceptual element that
will become an icon needs a placeholder.
   * **Function**: The placeholder's role is to reserve space and
provide a clear directive for the illustrator.
   * **Recommended Implementation**: A clean, professional way to do
this is with a gray, rounded-corner rectangle (`<rect rx="8" ry="8"
style="fill:#cccccc; stroke:#666666; stroke-width:1;" />`).
   * **Content (CRITICAL)**: Each placeholder MUST contain two pieces
of text:
      * **Exterior Label**: A concise name for the component, placed
**outside** the box (e.g., above it).
      * **Interior Description**: A detailed English phrase describing
the desired icon using the format `[icon]: <description>`, placed
**inside** the box (e.g., `[icon]: An icon showing a robot meticulously
reviewing a paper`). This description MUST NOT appear in the final
illustration but is a crucial instruction and it must be detailed and
concrete.

**Current Layout (Iteration ${iteration}):**

[PNG image of the current SVG layout will be provided]

[SVG source code will be provided]

**Paper Content Summary:**

${content}

**Reference Figures:**

[High-quality reference figure images will be provided here to set the
standard]

**Final Output Requirement:**

A single block of SVG code that is aesthetically superb and tells a
clear, compelling story of the paper's methodology.
```

SVG Critique (paper)

```
You are an experienced academic journal reviewer.  Your task is to
CRITIQUE the current SVG layout

**Evaluation Principles:**
   1.  **Aesthetic Design**:  Evaluate visual appeal, balance, color
harmony, typography, spacing, and overall professional appearance.  The
layout should be modern, polished, and visually engaging.
   2.  **Content Fidelity**:  Assess how accurately and completely the
visualization represents the core concepts, relationships, and key
information from the original content.  All essential elements should
be captured without distortion.
   3.  **Placeholder Usage**:  Examine compliance with placeholder
specifications, including proper exterior labels, detailed interior
icon descriptions, and adherence to the required format and positioning.


**Placeholder Specification:**
   * To prepare for final illustration, every conceptual element that
will become an icon needs a placeholder.
   * **Function**:  The placeholder's role is to reserve space and
provide a clear directive for the illustrator.
   * **Recommended Implementation**:  A clean, professional way to do
this is with a gray, rounded-corner rectangle (`<rect rx="8" ry="8"
style="fill:#cccccc; stroke:#666666; stroke-width:1;" />`).
   * **Content (CRITICAL)**:  Each placeholder MUST contain two pieces
of text:
      * **Exterior Label**:  A concise name for the component, placed
**outside** the box (e.g., above it).
      * **Interior Description**:  A detailed English phrase describing
the desired icon using the format `[icon]:  <description>`, placed
**inside** the box (e.g., `[icon]:  An icon showing a robot meticulously
reviewing a paper`).  This description MUST NOT appear in the final
illustration but is a crucial instruction and it must be detailed and
concrete.


**Current Layout for Evaluation (Iteration $iteration):**

[PNG image of the current SVG layout will be provided]

[SVG source code will be provided]

**Paper Content Summary:**

$content

**Reference Figures:**

[High-quality reference figure images will be provided here to set the
standard]

**Output Format (Strictly Enforced):**

First, output the evaluation JSON.

**Example JSON format:**
```

SVG Design (survey)

```
You are a top-tier **survey visualization expert**.  Your task is to
improve the current SVG layout according to the survey content and
instructions given to you and then generate a SUPERIOR, new version.

The common survey figure types include:  Taxonomy/Classification
Hierarchy, Conceptual Framework/Flowchart, Multi-Panel/Modular Diagram,
Cycle/Relational Diagram, Pyramid/Hierarchy Diagram, Comparison Figure,
Evolutionary Diagram, Timeline, Table of Contents, etc.

**Placeholder Specification:**
   * To prepare for final illustration, every conceptual element that
will become an icon needs a placeholder.
   * **Function**:  The placeholder's role is to reserve space and
provide a clear directive for the illustrator.
   * **Recommended Implementation**:  A clean, professional way to do
this is with a gray, rounded-corner rectangle (`<rect rx="8" ry="8"
style="fill:#cccccc; stroke:#666666; stroke-width:1;" />`).
   * **Content (CRITICAL)**:  Each placeholder MUST contain two pieces
of text:
      * **Exterior Label**:  A concise name for the component, placed
**outside** the box (e.g., above it).
      * **Interior Description**:  A detailed English phrase describing
the desired icon using the format `[icon]:  <description>`, placed
**inside** the box (e.g., `[icon]:  An icon showing a robot meticulously
reviewing a paper`).  This description MUST NOT appear in the final
illustration but is a crucial instruction and it must be detailed and
concrete.

**Current Layout (Iteration $iteration):**

[PNG image of the current SVG layout will be provided]

[SVG source code will be provided]

**Survey Content Summary:**

$content

**Reference Figures:**

[High-quality reference figure images will be provided here to set the
standard]

**Final Output Requirement:**

A single block of SVG code that is aesthetically superb and tells a
clear, compelling story of the survey.
```

SVG Critique (survey)

```
You are an experienced academic journal reviewer.  Your task is to
CRITIQUE the current SVG layout.

The common survey figure types include:  Taxonomy/Classification
Hierarchy, Conceptual Framework/Flowchart, Multi-Panel/Modular Diagram,
Cycle/Relational Diagram, Pyramid/Hierarchy Diagram, Comparison Figure,
Evolutionary Diagram, Timeline, Table of Contents, etc.

**Evaluation Principles:**
   1.  **Aesthetic Design**:  Evaluate visual appeal, balance, color
harmony, typography, spacing, and overall professional appearance.  The
layout should be modern, polished, and visually engaging.
   2.  **Content Fidelity**:  Assess how accurately and completely the
visualization represents the core concepts, relationships, and key
information from the original content.  All essential elements should
be captured without distortion.
   3.  **Placeholder Usage**:  Examine compliance with placeholder
specifications, including proper exterior labels, detailed interior
icon descriptions, and adherence to the required format and positioning.

**Placeholder Specification:**
   * To prepare for final illustration, every conceptual element that
will become an icon needs a placeholder.
   * **Function**:  The placeholder's role is to reserve space and
provide a clear directive for the illustrator.
   * **Recommended Implementation**:  A clean, professional way to do
this is with a gray, rounded-corner rectangle (`<rect rx="8" ry="8"
style="fill:#cccccc; stroke:#666666; stroke-width:1;" />`).
   * **Content (CRITICAL)**:  Each placeholder MUST contain two pieces
of text:
      * **Exterior Label**:  A concise name for the component, placed
**outside** the box (e.g., above it).
      * **Interior Description**:  A detailed English phrase describing
the desired icon using the format `[icon]:  <description>`, placed
**inside** the box (e.g., `[icon]:  An icon showing a robot meticulously
reviewing a paper`).  This description MUST NOT appear in the final
illustration but is a crucial instruction and it must be detailed and
concrete.


**Current Layout for Evaluation (Iteration $iteration):**

[PNG image of the current SVG layout will be provided]

[SVG source code will be provided]

**Survey Content Summary:**

$content

**Reference Figures:**

[High-quality reference figure images will be provided here to set the
standard]

**Output Format (Strictly Enforced):**

First, output the evaluation JSON.

**Example JSON format:**
```

SVG Design (blog)

```
You are a top-tier **educational illustration expert**.  Your task is
to improve the current SVG layout according to the blog content and
instructions given to you and then generate a SUPERIOR, new version.

**Placeholder Specification:**
   * To prepare for final illustration, every conceptual element that
will become an icon needs a placeholder.
   * **Function**:  The placeholder's role is to reserve space and
provide a clear directive for the illustrator.
   * **Recommended Implementation**:  A clean, professional way to do
this is with a gray, rounded-corner rectangle (`<rect rx="8" ry="8"
style="fill:#cccccc; stroke:#666666; stroke-width:1;" />`).
   * **Content (CRITICAL)**:  Each placeholder MUST contain two pieces
of text:
      * **Exterior Label**:  A concise name for the component, placed
**outside** the box (e.g., above it).
      * **Interior Description**:  A detailed English phrase describing
the desired icon using the format `[icon]:  <description>`, placed
**inside** the box (e.g., `[icon]:  An icon showing a robot meticulously
reviewing a paper`).  This description MUST NOT appear in the final
illustration but is a crucial instruction and it must be detailed and
concrete.


**Current Layout (Iteration $iteration):**

[PNG image of the current SVG layout will be provided]

[SVG source code will be provided]

**Blog Content Summary:**

$content

**Reference Figures:**

[High-quality reference figure images will be provided here to set the
standard]

**Final Output Requirement:**

A single block of SVG code that is aesthetically superb and tells a
clear, compelling story of the blog.
```

SVG Critique (blog)

```
You are an experienced academic journal reviewer.  Your task is to
CRITIQUE the current SVG layout.

**Evaluation Principles:**
   1.  **Aesthetic Design**:  Evaluate visual appeal, balance, color
harmony, typography, spacing, and overall professional appearance.  The
layout should be modern, polished, and visually engaging.
   2.  **Content Fidelity**:  Assess how accurately and completely the
visualization represents the core concepts, relationships, and key
information from the original content.  All essential elements should
be captured without distortion.
   3.  **Placeholder Usage**:  Examine compliance with placeholder
specifications, including proper exterior labels, detailed interior
icon descriptions, and adherence to the required format and positioning.

**Placeholder Specification:**
   * To prepare for final illustration, every conceptual element that
will become an icon needs a placeholder.
   * **Function**:  The placeholder's role is to reserve space and
provide a clear directive for the illustrator.
   * **Recommended Implementation**:  A clean, professional way to do
this is with a gray, rounded-corner rectangle (`<rect rx="8" ry="8"
style="fill:#cccccc; stroke:#666666; stroke-width:1;" />`).
   * **Content (CRITICAL)**:  Each placeholder MUST contain two pieces
of text:
      * **Exterior Label**:  A concise name for the component, placed
**outside** the box (e.g., above it).
      * **Interior Description**:  A detailed English phrase describing
the desired icon using the format `[icon]:  <description>`, placed
**inside** the box (e.g., `[icon]:  An icon showing a robot meticulously
reviewing a paper`).  This description MUST NOT appear in the final
illustration but is a crucial instruction and it must be detailed and
concrete.


**Current Layout for Evaluation (Iteration $iteration):**

[PNG image of the current SVG layout will be provided]

[SVG source code will be provided]

**Blog Content Summary:**

$content

**Reference Figures:**

[High-quality reference figure images will be provided here to set the
standard]

**Output Format (Strictly Enforced):**

First, output the evaluation JSON.

**Example JSON format:**
```

SVG Design (textbook)

```
You are a top-tier **educational visualization designer**. Your task is
to improve the current SVG layout according to the textbook content and
instructions given to you and then generate a SUPERIOR, new version.

**Placeholder Specification:**
   * To prepare for final illustration, every conceptual element that
will become an icon needs a placeholder.
   * **Function**: The placeholder's role is to reserve space and
provide a clear directive for the illustrator.
   * **Recommended Implementation**: A clean, professional way to do
this is with a gray, rounded-corner rectangle (`<rect rx="8" ry="8"
style="fill:#cccccc; stroke:#666666; stroke-width:1;" />`).
   * **Content (CRITICAL)**: Each placeholder MUST contain two pieces
of text:
      * **Exterior Label**: A concise name for the component, placed
**outside** the box (e.g., above it).
      * **Interior Description**: A detailed English phrase describing
the desired icon using the format `[icon]: <description>`, placed
**inside** the box (e.g., `[icon]: An icon showing a robot meticulously
reviewing a paper`). This description MUST NOT appear in the final
illustration but is a crucial instruction and it must be detailed and
concrete.


**Current Layout (Iteration $iteration):**

[PNG image of the current SVG layout will be provided]

[SVG source code will be provided]

**Textbook Content Summary:**

$content

**Reference Figures:**

[High-quality reference figure images will be provided here to set the
standard]

**Final Output Requirement:**

A single block of SVG code that is aesthetically superb and tells a
clear, compelling story of the textbook.
```

SVG Critique (textbook)

```
You are an experienced academic journal reviewer.  Your task is to
CRITIQUE the current SVG layout.

**Evaluation Principles:**
   1.  **Aesthetic Design**:  Evaluate visual appeal, balance, color
harmony, typography, spacing, and overall professional appearance.  The
layout should be modern, polished, and visually engaging.
   2.  **Content Fidelity**:  Assess how accurately and completely the
visualization represents the core concepts, relationships, and key
information from the original content.  All essential elements should
be captured without distortion.
   3.  **Placeholder Usage**:  Examine compliance with placeholder
specifications, including proper exterior labels, detailed interior
icon descriptions, and adherence to the required format and positioning.

**Placeholder Specification:**
   * To prepare for final illustration, every conceptual element that
will become an icon needs a placeholder.
   * **Function**:  The placeholder's role is to reserve space and
provide a clear directive for the illustrator.
   * **Recommended Implementation**:  A clean, professional way to do
this is with a gray, rounded-corner rectangle (`<rect rx="8" ry="8"
style="fill:#cccccc; stroke:#666666; stroke-width:1;" />`).
   * **Content (CRITICAL)**:  Each placeholder MUST contain two pieces
of text:
      * **Exterior Label**:  A concise name for the component, placed
**outside** the box (e.g., above it).
      * **Interior Description**:  A detailed English phrase describing
the desired icon using the format `[icon]:  <description>`, placed
**inside** the box (e.g., `[icon]:  An icon showing a robot meticulously
reviewing a paper`).  This description MUST NOT appear in the final
illustration but is a crucial instruction and it must be detailed and
concrete.


**Current Layout for Evaluation (Iteration $iteration):**

[PNG image of the current SVG layout will be provided]

[SVG source code will be provided]

**Textbook Content Summary:**

$content

**Reference Figures:**

[High-quality reference figure images will be provided here to set the
standard]

**Output Format (Strictly Enforced):**

First, output the evaluation JSON.

**Example JSON format:**
```

SVG Critique - Output Format

```
**Output Format (Strictly Enforced):**
Output ONLY a JSON evaluation with the following structure:
{
   "scores": {
      "aesthetic_design": <score_0_to_10>,
      "content_fidelity": <score_0_to_10>,
      "placeholder_usage": <score_0_to_10>
   },
   "critique_summary": "<brief_summary_of_strengths_and_weaknesses>",
   "specific_issues": ["<issue1>", "<issue2>", ...],
   "improvement_suggestions": ["<suggestion1>", "<suggestion2>", ...]
}
```

SVG to Text2Image Conversion - Part 1

```
You are an expert visual design analyst specializing in converting
technical diagrams into detailed text-to-image prompts.  Your task
is to analyze the provided SVG code and create a comprehensive prompt
that will guide AI image generation to produce a professional, visually
stunning scientific illustration.

**PRIMARY OBJECTIVE:**
Create a text-to-image prompt that will successfully transform gray
placeholder boxes into meaningful icons while maintaining perfect layout
structure and applying the specified artistic style:  "$art_style".

**ARTISTIC STYLE INTEGRATION:**
The final illustration MUST strictly follow this artistic style:
"$art_style"
   - All visual elements, colors, effects, and overall aesthetic must
align with this style
   - Icons and visual components should be designed to match this
artistic direction
   - Color palette, typography, and visual effects should complement
this style
   - The overall composition should embody the essence of "$art_style"

**CRITICAL ANALYSIS STEPS:**
   1.  **Gray Placeholder Identification & Style-Conscious Icon
Conversion:**
      - Locate ALL gray rectangular placeholders (fill="gray" or
fill="#808080" etc.)
      - Extract the descriptive text INSIDE each gray placeholder
      - For each placeholder, create a SPECIFIC, DETAILED icon
description that represents the concept
      - IMPORTANT: Each icon must be designed in the "$art_style" style
      - Ensure icon descriptions include style-specific visual
characteristics
   2.  **Layout Structure Documentation:**
      - Identify exact positions and sizes of all elements
      - Document arrow connections and flow directions
      - Note spatial relationships between components
      - Record all text labels that should remain as text
   3.  **Style-Specific Visual Enhancement Requirements:**
      - Apply "$art_style" consistently throughout the design
      - Define color schemes that match the specified artistic style
      - Specify visual hierarchy and emphasis appropriate for the style
      - Describe background treatment that complements the artistic
direction


**OUTPUT FORMAT REQUIREMENTS:**

Your response must include these EXACT sections:

**SECTION 1:  OVERALL SCENE DESCRIPTION**
"A professional $content_type methodology diagram featuring [describe
the main concept/process].  The illustration should be rendered in
the '$art_style' style with [style-appropriate color palette and
visual characteristics].  The layout follows [describe flow pattern:
left-to-right, top-to-bottom, circular, etc.].  The overall aesthetic
perfectly embodies the '$art_style' artistic direction."
```

SVG to Text2Image Conversion - Part 2

```
**SECTION 2:  STYLE-CONSCIOUS PLACEHOLDER-TO-ICON CONVERSIONS**
For each gray placeholder found, provide:
"Placeholder [position description]:  Replace with [VERY SPECIFIC icon
description that incorporates '$art_style' style elements].  The icon
should be [size] and positioned at [location], rendered in '$art_style'
style with [specific style characteristics:  colors, effects, textures,
etc.].  It represents [concept] and should visually communicate
[specific meaning] while perfectly matching the '$art_style' aesthetic."

**SECTION 3:  TEXT ELEMENTS TO PRESERVE**
"The following text must appear exactly as written in their specified
positions, styled to match '$art_style':  [list all text labels with
position descriptions and style-appropriate typography specifications]"

**SECTION 4:  ARTISTIC STYLE IMPLEMENTATION**
"The entire illustration must be rendered in the '$art_style' style.
Specific implementation requirements:
   - Color Palette:  [define colors that match the artistic style]
   - Visual Effects:  [specify effects appropriate for the style:
shadows, gradients, textures, etc.]
   - Typography:  [describe text styling that complements the artistic
direction]
   - Overall Aesthetic:  [detailed description of how the '$art_style'
should be applied]
   - Visual Characteristics:  [specific visual elements that define this
artistic style]"

**SECTION 5:  LAYOUT AND CONNECTIONS**
"Maintain these exact spatial relationships:  [describe arrangement].
Connect elements with [arrow/line specifications styled to match
'$art_style'].  Ensure [spacing and alignment requirements].  All
connecting elements should be rendered in '$art_style' style."

**STYLE-CONSCIOUS ICON CONVERSION EXAMPLE:**
If you find a gray box containing "data processing algorithm" and the
style is "Delicate and cute cartoon comic style":
"Replace with a charming cartoon-style computer chip character with
big expressive eyes and a friendly smile, featuring soft pastel colors
(light blues and pinks), rounded edges, and subtle sparkle effects
typical of cute cartoon designs, approximately 80x80 pixels"

**CRITICAL SUCCESS FACTORS:**
   - Every gray placeholder MUST be converted to a specific,
implementable icon description that matches '$art_style'
   - All text labels outside gray boxes MUST be preserved with
style-appropriate formatting
   - Layout structure MUST be maintained exactly
   - The '$art_style' style MUST be consistently applied throughout all
visual elements
   - Style specifications MUST be detailed enough for consistent
application
   - The prompt MUST be actionable for AI image generation in the
specified artistic style


**INPUT SVG CODE:**

"`svg

$source_content

"`
```

SVG to Text2Image Conversion - Part 3

```
**ARTISTIC STYLE TO APPLY:** "$art_style"

Now analyze this SVG and create the comprehensive text-to-image prompt
following the exact format above.  Focus especially on converting
every gray placeholder into a specific, detailed icon description that
perfectly matches the "$art_style" artistic style while maintaining
visual clarity and professional quality.
```

Image Rendering - Part 1

```
**BACKGROUND & PURPOSE:**
You are a world-class digital illustrator and scientific visualization
expert specializing in academic research publications.  Your mission
is to transform a layout diagram into a professional, publication-ready
scientific illustration suitable for academic journals and educational
materials.

**ARTISTIC STYLE DIRECTIVE:**
The entire illustration MUST be rendered consistently in the following
artistic style:  "$selected_style"
   - Apply this style uniformly across all visual components
   - Select colors, textures, and visual effects that authentically
represent "$selected_style"
   - Maintain visual coherence and professional appeal throughout
   - Ensure the style enhances both aesthetic appeal and functional
clarity

**COMPREHENSIVE VISUAL BLUEPRINT:**
Follow these detailed specifications for creating the enhanced
illustration:

"`
$enhancement_input
"`

**SYSTEMATIC EXECUTION PROCESS:**

**Step 1:  Specification Integration Analysis**
Thoroughly review the detailed visual specifications above to
comprehend:
   - Overall scene composition, flow patterns, and structural
relationships
   - Specific placeholder-to-icon conversion requirements and contextual
needs
   - Text preservation elements requiring exact accuracy
   - Style-specific implementation guidelines and aesthetic requirements


**Step 2:  Layout Architecture Implementation**
Based on the comprehensive specifications:
   - Maintain precise spatial relationships as detailed in the
specifications
   - Preserve all text positioning and visual hierarchy requirements
   - Implement the specified color palette and visual characteristic
guidelines
   - Execute the described connection patterns and directional flow
systems
```

Image Rendering - Part 2

```
**Step 3:  Style-Conscious Icon Development**
For each icon conversion requirement specified:
   - Create detailed, professional-grade icons matching both functional descriptions and "$selected_style"
   - Ensure visual consistency and coherent styling across all icon elements
   - Apply appropriate sizing, positioning, and visual treatment as specified
   - Use colors and effects that create harmonious integration with the artistic direction

**Step 4:  Typography & Text Rendering Excellence**
   - Render preserved text elements with exceptional clarity and professional readability
   - Apply typography styling that authentically complements "$selected_style"
   - Maintain specified text positioning and visual hierarchy requirements
   - Ensure absolute accuracy of all preserved textual content

**Step 5:  Unified Style Integration**
Harmonize all visual elements under the "$selected_style" aesthetic framework:
   - Apply consistent visual effects including shadows, gradients, and textural elements
   - Ensure sophisticated color relationships and visual harmony throughout
   - Balance artistic sophistication with professional functionality
   - Optimize visual presentation for academic publication excellence

**PROFESSIONAL QUALITY STANDARDS:**
   - High-resolution output optimized for academic publication requirements
   - Perfect adherence to detailed visual specifications and requirements
   - Professional implementation of "$selected_style" with authentic aesthetic representation
   - Contextually appropriate and visually stunning icon development
   - Complete elimination of placeholder instruction text in final output
   - Seamless integration and visual unity across all elements

**DELIVERY SPECIFICATIONS:**
Create an exceptional scientific illustration that flawlessly implements both the comprehensive visual specifications and the "$selected_style" artistic direction, resulting in a publication-ready visualization that exceeds professional academic standards.

Begin the transformation of the provided layout diagram now.
```

OCR Correction

```
You are a professional text recognition expert.  Please correct text
errors in OCR recognition results based on the correct text content
provided in the SVG code.

**Task Description:**
   1.  I have an OCR recognition result JSON file that contains text
coordinate information
   2.  I also have SVG code that contains the correct text content
   3.  You need to refer to the correct text in the SVG code, reference
the original image (polished.png) and reference background image
(erased.png), correct possible recognition errors in the OCR results,
and check again after correction to remove unnecessary duplicate content

**OCR Recognition Results:**
"`json
$library_data
"`
**Reference SVG Code:**
"`svg
$svg_code
"`
**Output Format:**
Please directly output the corrected complete JSON data in exactly the
same format as the input.

Corrected JSON:
```

Generate PPT Code - Part 1

```
You are a professional PowerPoint development expert and designer.
Please analyze the provided enhanced image, determine its visual style,
and then generate complete python-pptx code based on text coordinate
information.

**Core Tasks:**
   1.  **Style Analysis**:  Carefully observe the enhanced image,
analyze its design style, color matching, visual effects, etc.
   2.  **PPT Generation**:  Use the text-removed background image as PPT
background, set text styles according to the enhanced image's style
   3.  **Precise Positioning**:  Place text in correct positions based
on text coordinate information
   4.  **Style Consistency**:  Ensure PPT text styles are consistent
with the overall style of the enhanced image

**Analysis Requirements:**
Please observe the enhanced image and automatically determine the
following:
   - Overall design style (modern minimalist, tech-style, academic,
artistic, cute cartoon, etc.)
   - Main color schemes and color palettes
   - Suitable text colors (ensuring good contrast and readability with
background)
   - Appropriate font choices (modern, traditional, decorative, etc.)
   - Optimal text size proportions
```

Generate PPT Code - Part 2

```
**Text Coordinate Information:**
$coordinates_info

**Technical Specifications:**
   - Background image size:  $image_width x $image_height pixels
   - PPT size:  $ppt_width_inches" x $ppt_height_inches" (Inches)
   - Use python-pptx library
   - Coordinates already converted to Inches units, use directly
   - Background image path:  '$erased_image_path'
   - Output file:  '$output_pptx_path'

**Style Requirements:**
   - Choose appropriate fonts based on background image style
(prioritize recommended font lists)
   - Use recommended text colors to ensure good contrast and readability
   - Intelligently adjust font size based on font size factors and text
box dimensions
   - For important text (like titles), use primary text colors; for
secondary text, use alternative colors
   - Ensure text styles are coordinated with background image style

**Code Requirements:**
   - Import necessary libraries:  from pptx import Presentation, from
pptx.util import Inches, Pt
   - Import text adaptation enums:  from pptx.enum.text import PP_ALIGN,
MSO_AUTO_SIZE, MSO_ANCHOR
   - Create blank layout slide
   - Set slide size to:  Inches($ppt_width_inches) x
Inches($ppt_height_inches)
   - Add background image to fill full screen
   - **[Core Requirement] When creating text boxes for each text, must
perform the following steps**:
      1.  Use `slide.shapes.add_textbox(left, top, width, height)` to
create text box
      2.  Get `text_frame = textbox.text_frame`
      3.  **Immediately set disable word wrap**:  `text_frame.word_wrap
= False`
      4.  **Immediately set vertical alignment**:
`text_frame.vertical_anchor = MSO_ANCHOR.MIDDLE`
      5.  **Immediately set margins**:  Set all margins to
`Inches(0.01)`
      6.  Then set text content and formatting
   - Use provided precise coordinates and dimensions
   - Intelligently set fonts, colors, and sizes (refer to style analysis
results)
   - Save as specified filename

**Code Generation Requirements:**
   1.  **Intelligent Style Adaptation**:  Automatically select optimal
text colors, fonts, and sizes based on enhanced image's visual style
   2.  **Precise Font Size**:  Intelligently calculate font size based
on text box size and text content to ensure text fills the text box
perfectly
   3.  **High-Quality Presentation**:  Ensure generated PPT maintains
visual consistency with the enhanced image
```

Generate PPT Code - Part 3

```
**Code Template:**
"'python
from pptx import Presentation
from pptx.util import Inches, Pt
from pptx.dml.color import RGBColor
from pptx.enum.text import PP_ALIGN, MSO_AUTO_SIZE, MSO_ANCHOR

# Create presentation
prs = Presentation()
slide_layout = prs.slide_layouts[6] # Blank layout
slide = prs.slides.add_slide(slide_layout)

# Set slide size (maintain 1:1 ratio with background image)
prs.slide_width = Inches($ppt_width_inches)
prs.slide_height = Inches($ppt_height_inches)

# Add background image (fill full screen)
slide.shapes.add_picture('$erased_image_path',
                 Inches(0), Inches(0),
                 Inches($ppt_width_inches),
                 Inches($ppt_height_inches))

# [Based on enhanced image style, add specific text box code here]
# Example code structure:
# textbox = slide.shapes.add_textbox(Inches(left), Inches(top),
Inches(width), Inches(height))
# text_frame = textbox.text_frame
# text_frame.word_wrap = False
# text_frame.vertical_anchor = MSO_ANCHOR.MIDDLE
# text_frame.margin_left = Inches(0.01)
# text_frame.margin_right = Inches(0.01)
# text_frame.margin_top = Inches(0.01)
# text_frame.margin_bottom = Inches(0.01)
#
# p = text_frame.paragraphs[0]
# p.alignment = PP_ALIGN.CENTER
# run = p.add_run()
# run.text = "Text content"
# run.font.name = "Font name" # Choose based on enhanced image style
# run.font.size = Pt(font_size) # Intelligently calculate based on text
box size
# run.font.bold = True/False
# run.font.color.rgb = RGBColor(r, g, b) # Choose based on enhanced
image color scheme

# Save file
prs.save('$output_pptx_path')
"'
```

Generate PPT Code - Part 4

```
**Output Requirements:**
Please generate complete executable Python code based on the visual
style of the enhanced image, ensuring the code can run directly.  The
code should:
   1.  Create a new presentation
   2.  Set correct slide size (1:1 ratio with background image)
   3.  Add background image to fill full screen
   4.  Add text boxes for each text using precise coordinates
   5.  **[Important] Text Auto-Fit Functionality**:
      - **Must** set 'text_frame.word_wrap = False'
      - **Must** set margins to 0.01
      - **Must** set vertical alignment (e.g., 'MSO_ANCHOR.MIDDLE')
   6.  **Intelligent Style Adaptation** (automatically choose based on
enhanced image):
      - Analyze enhanced image's color scheme, choose appropriate text
colors
      - Choose matching fonts based on enhanced image's design style
      - Adjust text styles based on text importance and enhanced image's
hierarchy
      - Enable bold to improve readability (if suitable for overall
style)
      - Ensure good contrast and readability between text and background
   7.  Save file

**[Most Critical Requirement] Text Auto-Fit Settings:**
Each text box must include the following settings to ensure text does
not exceed text box boundaries:
"'python
text_frame.word_wrap = False # Disable automatic line wrapping
text_frame.vertical_anchor = MSO_ANCHOR.MIDDLE # Vertical center
alignment
text_frame.margin_left = Inches(0.01) # Left margin
text_frame.margin_right = Inches(0.01) # Right margin
text_frame.margin_top = Inches(0.01) # Top margin
text_frame.margin_bottom = Inches(0.01) # Bottom margin
"'


**Special Notes:**
   - For title-type text (usually larger or prominently positioned), use
primary text colors that match enhanced image style
   - For body text, use softer auxiliary colors but ensure readability
   - Text must completely fit within its text box, absolutely no
overflow allowed
   - Choose coordinated text colors based on enhanced image's overall
color scheme
   - Ensure sufficient contrast between text and background
```

Generate PPT Code - Part 5

```
**[Text Auto-Fit Best Practice Example]**
"`python
# Standard text box creation process example
for i, text_data in enumerate(text_data_list):
   # 1.  Create text box
   textbox = slide.shapes.add_textbox(
      Inches(text_data['ppt_left_inches']),
      Inches(text_data['ppt_top_inches']),
      Inches(text_data['ppt_width_inches']),
      Inches(text_data['ppt_height_inches'])
   )

   # 2.  Get text frame
   text_frame = textbox.text_frame

   # 3.  [Must] Immediately set auto-fit properties
   text_frame.word_wrap = False
   text_frame.vertical_anchor = MSO_ANCHOR.MIDDLE
   text_frame.margin_left = Inches(0.01)
   text_frame.margin_right = Inches(0.01)
   text_frame.margin_top = Inches(0.01)
   text_frame.margin_bottom = Inches(0.01)

   # 4.  Set text content and formatting
   p = text_frame.paragraphs[0]
   p.alignment = PP_ALIGN.CENTER
   run = p.add_run()
   run.text = text_data['text']
   run.font.name = "Font Name" # Choose appropriate font based on
enhanced image style
   run.font.bold = True # Decide whether to bold based on style needs
   run.font.color.rgb = RGBColor(r, g, b) # Choose based on enhanced
image color scheme
"`

"`python
# Write complete code here, must include the above text auto-fit
settings
"`
```

Referenced scoring - Part 1

```
You are a world-class Art Director and Visual Communication Expert
for top-tier scientific publications.  Your evaluation combines
sophisticated aesthetic judgment with deep understanding of modern
visual design principles.  You recognize that excellence in scientific
visualization requires both visual beauty, effective communication and
content fidelity.
You MUST use the following {type_context['content_name']} as the ground
truth for what the figures should communicate.
The target audience is:  {type_context['audience']}.
{type_context['evaluation_focus']}
--
{content_text}
--

**Reference Figure Context:**
You will be shown a REFERENCE FIGURE (labeled "Reference Figure") which
represents the original, authentic figure for this {content_type}.  This
reference figure serves as the ground truth standard for comparison.
Use this reference to guide your evaluation by considering:
   - How well does the candidate figure capture the key visual elements
of the reference?
   - Does the candidate figure maintain the essential information
structure while potentially improving visual design?
   - How does the candidate figure's approach compare to the reference
in terms of clarity and effectiveness?

Please note:  The reference figure represents the original authentic
visualization, while the candidate figure is a generated/redesigned
version that should be evaluated both independently for its design
quality AND in relation to how well it serves the same communicative
purpose as the reference.

--

**Core Philosophy:  Champion Modern Visual Excellence**
   - **Distinguish between sophistication and clutter.** A sophisticated
figure may use rich visual elements, multiple colors, detailed icons,
and layered information - this is NOT clutter if well-organized.  True
clutter is disorganized, inconsistent, and poorly structured content.
   - **Recognize modern design excellence.** The best contemporary
figures combine visual appeal with information richness.  They use
professional color palettes, thoughtful typography, meaningful icons,
and sophisticated layouts that engage the viewer while communicating
clearly.
   - **Value information-rich design.** A figure that successfully
presents comprehensive information through well-designed visual elements
should be highly valued, not penalized for complexity.
   - **Use the full scoring range (1-10).** Reserve 9-10 for figures
that demonstrate both modern visual sophistication AND clear
communication.  A basic, minimal figure should score 5-6, not 7-8.
```

Referenced scoring - Part 2

```
**What Constitutes Modern Visual Excellence:**
   - **Sophisticated Visual Language:** Professional use of colors,
gradients, shadows, and modern typography that creates visual hierarchy
and engagement
   - **Meaningful Visual Elements:** Thoughtful use of icons,
illustrations, and visual metaphors that enhance understanding beyond
basic shapes and boxes
   - **Information Architecture:** Well-organized presentation of
complex information through visual structure, grouping, and flow
   - **Design Craftsmanship:** Attention to visual details like
consistent spacing, professional color coordination, and polished
execution

**Evaluation Dimensions (Score 1-10, one decimal place):**

--
**Part 1:  Visual Design Excellence (How sophisticated and appealing is
the design?)**
*Evaluate modern visual design quality and professional execution.*

   1.  **Aesthetic & Design Quality (ADQ):** - **Highest Weight**
      - **Modern Visual Appeal:** Does the figure demonstrate
contemporary design sophistication?  Does it use professional color
schemes, thoughtful gradients, appropriate shadows, and modern
typography to create visual interest and hierarchy?
      - **Composition & Layout:** Is the layout well-structured with
intentional design choices?  Note that effective use of space may
include rich visual content, not just whitespace.
      - **Design Innovation:** Does the figure go beyond basic boxes and
arrows to use creative visual solutions, meaningful icons, and engaging
presentation methods?

   2.  **Visual Expressiveness (VE):**
      - **Rich Visual Language:** Are visual elements (icons,
illustrations, graphics) professionally designed and semantically
meaningful?  Do they enhance understanding through visual metaphors
and clear symbolism?
      - **Information Visualization:** How effectively does the figure
transform abstract concepts into concrete visual representations?  Does
it make complex ideas accessible through visual design?
      - **Style Sophistication:** Does the overall visual style
demonstrate professional design standards comparable to high-quality
infographics and modern scientific publications?

   3.  **Professional Polish (PP):**
      - **Execution Excellence:** Is every design element carefully
crafted with attention to detail?  This includes consistent styling,
proper alignment, appropriate scaling, and cohesive visual treatment.
      - **Technical Proficiency:** Does the figure demonstrate mastery
of design principles including color theory, typography, visual
hierarchy, and layout composition?

--
**Part 2:  Communication Effectiveness (How well does it communicate?)**
*Focus on clarity and information delivery while acknowledging that
sophisticated visuals can enhance communication.*
```

Referenced scoring - Part 3

```
   4.  **Clarity:**
      - **Visual Organization:** Is complex information well-organized
through visual structure?  A sophisticated figure with many elements can
still be clear if well-organized.
      - **Information Accessibility:** Can viewers quickly understand
the main message and navigate detailed information?  Good visual
hierarchy supports complexity.

   5.  **Logical Flow:**
      - **Narrative Structure:** Does the figure tell a clear story or
present a logical progression?  This can be achieved through various
visual means including flow lines, visual grouping, and hierarchical
presentation.
      - **Guided Exploration:** Does the visual design help viewers
navigate and understand the content systematically, even when the
content is information-rich?

--
**Part 3:  Content Fidelity (Faithfulness to the Source
{content_type})**

   6.  **Accuracy:**
      - Does the figure faithfully represent all key components and
relationships described in the source text?
   7.  **Completeness:**
      - Are any critical elements from the source content missing or
misrepresented?
   8.  **Appropriateness to Audience:**
      - Is the figure's complexity, abstraction level, and style
appropriate for the target audience ({audience})?

**Scoring Guidelines & Final Judgment:**
   - **Focus on Accurate Dimensional Scores:** Provide precise scores
(1-10, one decimal place) for each dimension based on the specific
criteria.
   - **Reward Visual Sophistication:** A figure with rich visual
design, professional execution, and effective information presentation
deserves high scores (8-10).  Don't penalize sophistication if it's
well-executed.
   - **Penalize Amateur Design:** Basic figures with minimal visual
design, poor color choices, or unprofessional execution should score
lower (4-6), regardless of information completeness.
   - **Information-Rich vs.  Cluttered:** Distinguish between
information-rich (good - uses visual design to organize complex content)
and cluttered (bad - disorganized, inconsistent, poorly structured).
   - **Modern vs.  Traditional:** Value modern design approaches
including creative use of color, sophisticated typography, meaningful
icons, and visual innovation over traditional academic minimalism.
```

Referenced scoring - Part 4

```
**Critical Evaluation Questions:**
   1.  Would this figure stand out positively in a modern scientific
publication or high-quality presentation?
   2.  Does it demonstrate professional design skills beyond basic
diagramming?    3.  Would viewers find it visually engaging and easy
to understand despite complexity?
   4.  Does it successfully transform abstract concepts into compelling
visual narratives?

Use these questions to guide your dimensional assessments, ensuring each
dimension receives an accurate score based on its specific criteria.

**Please use the following JSON template for your output:**
```json
{
   "figure_id":  "{figure_id}",
   "scores":  {
      "aesthetic_and_design_quality":  {"score":  8.5, "reasoning":
"Demonstrates sophisticated modern design with professional color
palette, thoughtful gradients, and contemporary typography that creates
strong visual hierarchy and engagement."},
      "visual_expressiveness":  {"score":  9.0, "reasoning":  "Rich
visual language with meaningful icons, professional illustrations,
and effective visual metaphors that transform abstract concepts into
accessible visual representations."},
      "professional_polish":  {"score":  8.0, "reasoning":  "Excellent
execution with consistent styling, proper alignment, cohesive visual
treatment, and mastery of design principles."},
      "clarity":  {"score":  7.5, "reasoning":  "Complex information
is well-organized through sophisticated visual structure, making it
accessible despite information richness."},
      "logical_flow":  {"score":  8.0, "reasoning":  "Clear narrative
structure with effective visual grouping and hierarchical presentation
that guides systematic exploration."},
      "accuracy":  {"score":  8.5, "reasoning":  "The figure accurately
represents the main concepts from the {content_type}."},
      "completeness":  {"score":  8.0, "reasoning":  "The figure
includes all critical elements from the {content_type}."},
      "appropriateness":  {"score":  8.5, "reasoning":  "The figure's
sophisticated design and information richness are perfectly appropriate
for {audience}."}

   }
}
```
```

Pairwise Comparison - Part 1

```
You are a world-class Art Director and Visual Communication Expert for
top-tier scientific publications.  Your judgment combines sophisticated
aesthetic taste with deep understanding of modern visual design
principles.  You must decide which figure demonstrates superior visual
design, communication effectiveness and content fidelity.

You MUST use the following {type_context['content_name']} as the ground
truth for what the figures should communicate.
The target audience is:  {type_context['audience']}.
{type_context['evaluation_focus']}

--
{content_text}
--
**Core Philosophy:  Recognize Modern Visual Excellence**
   - **Value sophisticated design over minimalism.** A well-executed
figure with rich visual elements, professional color usage, meaningful
icons, and thoughtful composition is superior to a basic, minimal
figure, even if the minimal figure is "cleaner."
   - **Distinguish sophistication from clutter.** True sophistication
uses visual complexity purposefully to enhance communication.  Clutter
is disorganized and inconsistent.  A figure with many well-designed
elements is sophisticated, not cluttered.
   - **Champion professional execution.** Look for evidence of
professional design skills:  proper color theory application, typography
mastery, visual hierarchy, consistent styling, and polished execution.
   - **Reward visual innovation.** Figures that go beyond basic boxes
and arrows to use creative visual solutions, meaningful metaphors, and
engaging presentation should be strongly preferred.

**Modern Design Superiority Indicators:**
   - **Visual Sophistication:** Professional color palettes, gradients,
shadows, contemporary typography, and thoughtful visual hierarchy
   - **Rich Information Visualization:** Meaningful icons,
illustrations, and visual metaphors that make abstract concepts concrete
and accessible
   - **Design Craftsmanship:** Attention to detail in spacing,
alignment, color coordination, and overall visual harmony
   - **Contemporary Aesthetics:** Modern visual language that would be
appropriate for high-quality scientific publications and professional
presentations
```

Pairwise Comparison - Part 2

```
**Comparison Dimensions (Choose A, B, Both good, or Both bad for
each):**

**Important:  Selection Criteria**
   - **A**:  Choose A if Figure A is clearly superior to Figure B
   - **B**:  Choose B if Figure B is clearly superior to Figure A
   - **Both good**:  Choose this if BOTH figures demonstrate high
quality and professional standards, making it difficult to declare a
clear winner (both are publication-ready with only minor differences in
style)
   - **Both bad**:  Choose this if BOTH figures have significant flaws
or fail to meet professional standards (neither would be suitable for
publication without major revisions)

--
**Part 1:  Visual Design Excellence (Which demonstrates superior modern
design?)**

   1.  **Aesthetic & Design Quality (ADQ):** - **Highest Weight**
      - Which figure demonstrates more sophisticated visual design?
Consider professional color usage, contemporary typography, thoughtful
composition, and modern visual appeal.
      - Which figure would be more impressive in a high-quality
scientific publication or professional presentation?
      - If both are professionally designed or both are poorly designed,
choose "Both good" or "Both bad" accordingly.

   2.  **Visual Expressiveness (VE):**
      - Which figure uses richer, more meaningful visual language?  Look
for professional icons, illustrations, visual metaphors, and creative
design solutions that go beyond basic shapes.
      - Which figure better transforms abstract concepts into engaging
visual representations?
      - If both excel or both fail at visual expressiveness, choose
"Both good" or "Both bad" accordingly.

   3.  **Professional Polish (PP):**
      - Which figure demonstrates superior design craftsmanship and
technical proficiency?  Consider consistency, attention to detail,
proper use of design principles, and overall execution quality.
      - Which figure shows evidence of professional design skills rather
than basic diagramming?
      - If both show professional polish or both lack it, choose "Both
good" or "Both bad" accordingly.
--
**Part 2:  Communication & Sophistication (Which is more effective and
sophisticated?)**

   4.  **Clarity:**
      - Which figure better organizes complex information through
sophisticated visual structure?  Remember that well-designed complexity
can be clearer than oversimplified content.
      - Which figure makes information more accessible through
thoughtful visual design?
      - If both are equally clear or both are confusing, choose "Both
good" or "Both bad" accordingly.
```

Pairwise Comparison - Part 3

```
   5.  **Logical Flow:**
      - Which figure presents information with better visual narrative
and guidance?  This can be achieved through various sophisticated visual
means.
      - Which figure demonstrates superior information architecture and
visual hierarchy?
      - If both have excellent or poor logical flow, choose "Both good"
or "Both bad" accordingly.

   6.  **Information Sophistication:**
      - Which figure provides more comprehensive and well-presented
information while maintaining visual appeal?
      - Which figure better balances information richness with visual
accessibility?
      - If both balance information well or both fail to do so, choose
"Both good" or "Both bad" accordingly.

   7.  **Content Fidelity (CF):**
      - Which figure is more faithful to the source {content_type} text,
accurately representing all key components without critical omissions?
      - Which figure is more appropriate for the target audience
({type_context['audience']})?

**Final Decision Guidelines:**
   - **Choose A or B** when there is a clear winner that demonstrates
superior modern visual design and professional execution.
   - **Choose "Both good"** when BOTH figures meet professional
publication standards and the differences are primarily stylistic
preferences rather than quality differences.
   - **Choose "Both bad"** when BOTH figures have significant
deficiencies that would prevent publication without major revisions.
   - **Value visual innovation and richness.** A figure with thoughtful
use of colors, meaningful icons, professional typography, and
sophisticated layout should win over basic diagrams.
   - **Consider publication quality.** Which figure(s) would be
appropriate for a high-quality scientific publication or professional
presentation?
   - **Be specific about design superiority or shared
quality/deficiency.** Explain exactly why one figure wins, or why both
are good/bad.
```

Pairwise Comparison - Part 4

```
**Please use the following JSON template for your output:**
"`json
{
   "comparison_id": "{comparison_id}",
   "dimensional_comparison": {
      "aesthetic_and_design_quality": {"winner": "A or B or Both
good or Both bad", "reasoning": "Explain your choice. For 'A'/'B':
specify why one is superior. For 'Both good': explain why both
meet professional standards. For 'Both bad': explain why both have
significant flaws."},
      "visual_expressiveness": {"winner": "A or B or Both good or Both
bad", "reasoning": "For 'A'/'B': explain superior visual language. For
'Both good': explain why both excel. For 'Both bad': explain why both
fail."},
      "professional_polish": {"winner": "A or B or Both good or Both
bad", "reasoning": "For 'A'/'B': explain superior craftsmanship. For
'Both good': explain why both are polished. For 'Both bad': explain
why both lack polish."},
      "clarity": {"winner": "A or B or Both good or Both bad",
"reasoning": "For 'A'/'B': explain superior clarity. For 'Both good':
explain why both are clear. For 'Both bad': explain why both are
confusing."},
      "logical_flow": {"winner": "A or B or Both good or Both bad",
"reasoning": "For 'A'/'B': explain superior flow. For 'Both good':
explain why both have excellent flow. For 'Both bad': explain why both
lack proper flow."},
      "information_sophistication": {"winner": "A or B or Both good
or Both bad", "reasoning": "For 'A'/'B': explain superior information
balance. For 'Both good': explain why both balance well. For 'Both
bad': explain why both fail to balance."},
      "content_fidelity": {"winner": "A", "reasoning": "Figure A more
accurately represents the key concepts from the {content_type} and is
better suited for {type_context['audience']}."}
   },
   "final_decision": {
      "winner": "A or B or Both good or Both bad",
      "confidence": "High or Medium or Low",
      "reasoning": "Provide detailed reasoning for your choice. If 'A'
or 'B': explain why it's the clear winner. If 'Both good': explain why
both figures meet professional publication standards with only stylistic
differences. If 'Both bad': explain why both figures have significant
deficiencies preventing publication."
   }
}
"`
**Example Response for "A wins":**
"`json
{
   "comparison_id": "example_1",
   "dimensional_comparison": {
      "aesthetic_and_design_quality": {"winner": "A", "reasoning":
"Figure A demonstrates sophisticated modern design with professional
color palette, contemporary typography, and thoughtful visual hierarchy,
while Figure B uses basic colors and minimal design elements."},
      "visual_expressiveness": {"winner": "A", "reasoning": "Figure
A uses rich visual language with meaningful icons and professional
illustrations; Figure B relies on simple shapes and basic arrows."}
   },
```

Pairwise Comparison - Part 5

```
    "final_decision": {
        "winner": "A",
        "confidence": "High",
        "reasoning": "Figure A is the clear winner due to its
sophisticated modern design and professional execution."
    }
}
``
**Example Response for "Both good":**
``json
{
    "comparison_id": "example_2",
    "dimensional_comparison": {
        "aesthetic_and_design_quality": {"winner": "Both good",
"reasoning": "Both figures demonstrate professional color usage
and contemporary design suitable for publication. Figure A uses a
warmer palette while Figure B uses cooler tones, but both are equally
sophisticated."},
        "visual_expressiveness": {"winner": "Both good", "reasoning":
"Both figures effectively use icons and visual metaphors to communicate
concepts. Figure A emphasizes flowcharts while Figure B emphasizes
component diagrams, but both are equally expressive."}
    },
    "final_decision": {
        "winner": "Both good",
        "confidence": "High",
        "reasoning": "Both figures meet professional publication
standards with excellent design quality. The differences are primarily
stylistic preferences rather than quality differences."
    }
}
``

**Example Response for "Both bad":**
``json
{
    "comparison_id": "example_3",
    "dimensional_comparison": {
        "aesthetic_and_design_quality": {"winner": "Both bad",
"reasoning": "Both figures use clashing colors and poor typography
that would not be acceptable in professional publications. Neither
demonstrates adequate visual design skills."},
        "professional_polish": {"winner": "Both bad", "reasoning":
"Both figures have alignment issues, inconsistent styling, and poor
attention to detail. Neither shows professional-level execution."}
    },
    "final_decision": {
        "winner": "Both bad",
        "confidence": "High",
        "reasoning": "Both figures have significant design deficiencies
that would require major revisions before publication. Neither meets
minimum professional standards for scientific illustrations."
    }
}
``
```

Content Type Context Definitions

```
**Paper (Research Paper Methodology):**
   - **Content Name:** research paper methodology
   - **Evaluation Focus:** The figure should accurately represent the
research methodology, clearly showing the workflow, data flow, or system
architecture described in the paper.
   - **Audience:** academic researchers and peers in the field

**Survey (Survey/Review Article Content):**
   - **Content Name:** survey/review article content
   - **Evaluation Focus:** The figure should provide a clear overview or
taxonomy that helps readers understand relationships between different
approaches, methods, or concepts covered in the survey.
   - **Audience:** researchers seeking to understand the landscape of a
field

**Textbook (Educational Textbook Material):**
   - **Content Name:** educational textbook material
   - **Evaluation Focus:** The figure should support learning
objectives, making complex concepts accessible and easy to understand
for students.
   - **Audience:** students and educators

**Blog (Blog Article Content):**
   - **Content Name:** blog article content
   - **Evaluation Focus:** The figure should engage readers and make
the content more accessible, with emphasis on visual appeal and easy
comprehension.
   - **Audience:** general readers and practitioners
```