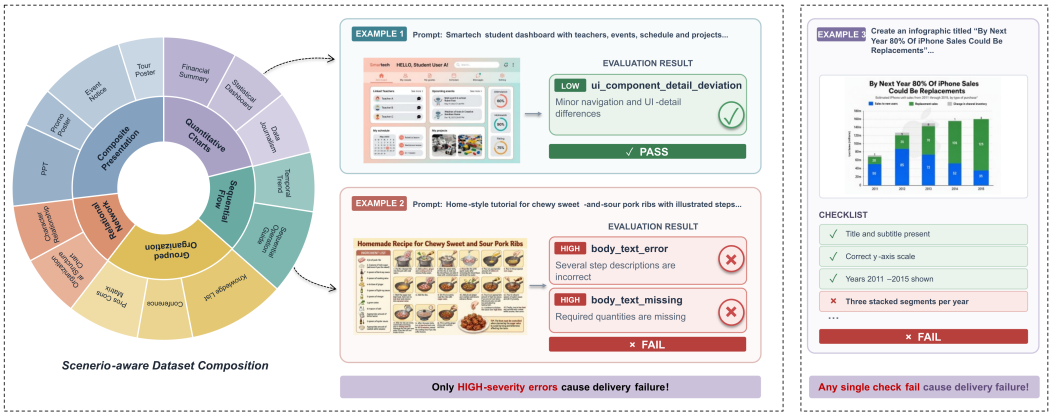
INFODIAG-BENCH: A SCENARIO-AWARE DIAGNOSTIC BENCHMARK FOR DELIVERABLE INFOGRAPHIC GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

While modern text-to-image (T2I) models produce visually compelling natural images, generating infographics suitable for practical use remains challenging. Such images require accurate rendering of dense text and numerical content, and coherent topological structures, which vary across different tasks. Consequently, infographic evaluation also requires pronounced inter-domain diversity across application scenarios. Furthermore, assessing practical deliverability necessitates distinguishing tolerable deviations from delivery-blocking errors. To address these challenges, we propose InfoDiag-Bench, a scenario-aware infographic benchmark spanning 663 realistic test cases across 14 application scenarios. Central to our evaluation framework is a set of predefined, evidence-grounded error types designed to systematically separate tolerable flaws from delivery-blocking ones. Across 10 open-source and close-source models, we find that the inability to preserve topological relationships among elements is the primary cause of delivery failure. Human evaluation further demonstrates a strong alignment between our automated judgements with human anticipation.



(a) Overview of InfoDiag-Bench *(b) Zero-Tolerance-Based Methods*

Figure 1: (a) Overview of InfoDiag-Bench including dataset composition and delivery-oriented evaluation benchmark. The dataset covers 14 application scenarios grouped into five information organization categories, encompassing 663 test cases, 79 evaluation dimensions, and 45 reusable error types. Our evaluation framework distinguishes tolerable deviations (low severity) from delivery-blocking errors (high severity), ensuring that only the latter trigger image-level delivery failure. (b) In contrast, Other methods determines image-level acceptance by zero-tolerance evaluation protocol, requiring all atomic questions to pass. The data shown here are represented by IGenBench results. Tang et al. (2026).

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

1 INTRODUCTION

Text-to-image (T2I) generation offers a promising avenue for translating complex information into accessible infographics (Labs, 2024; Esser et al., 2024; Li et al., 2024; Lu et al., 2024; Chen et al., 2024; Peng et al., 2025a; Li et al., 2025; Qiu et al., 2025; Xie et al., 2025; Wang et al., 2025a). Different from natural image synthesis, infographic generation requires explaining factual content and logical relationships that are directly applicable in real-world scenarios, rather than merely producing aesthetically pleasing visuals. A financial summary, for example, must preserve rigorous numerical values and accurate geometry-number mappings, while a procedural guide must convey the correct sequence and dependencies between steps. Errors in these core structural elements can convey a fundamentally flawed message, even when the generated image appears visually polished. Conversely, minor deviations in decorative elements or label placement may be tolerable. Evaluating infographic generation therefore requires distinguishing errors that prevent delivery from deviations that remain acceptable in practice.

This distinction presents two primary challenges. First, infographics exhibit pronounced inter-domain diversity across application scenarios, calling for adaptive, domain-specific evaluation criteria (Zhuo et al., 2026; Wei et al., 2025). For example, financial summaries emphasize numerical consistency, relationship networks rely on precise topological connections, and procedural guides demand coherent directional structures. An effective evaluation framework must comprehensively account for these distinct requirements. Second, different error types exert varying degrees of impact on real-world deliverability. An incorrect nodal connection can entirely alter the semantics of a diagram, whereas a conspicuous style-color shift might have negligible effects on information interpretation. Consequently, image-level acceptance requires explicit criteria to determine which specific deviations render an image unusable. We formalize acceptance under these rigorous practical criteria as *deliverability*.

Recent infographic benchmarks, including IGenBench (Tang et al., 2026), BizGenEval (Li et al., 2026a), and StructBench (Zhuo et al., 2026), have advanced the evaluation of visual generation through validating text/number accuracy and data encoding. IGenBench, in particular, assesses requirements such as visual encoding, demonstrating the necessity of moving beyond explicit data/text accuracy. However, translating these multi-dimensional evaluations into acceptance decisions remains an unresolved dilemma, as existing methods typically fall into two problematic extremes. On one hand, some approaches employ an averaging protocol across metrics, yielding aggregate scores that can obscure critical local failures (e.g., severe data hallucination) if the model performs well on superficial criteria. On the other hand, methods adopting rigid "all-checks-pass" (zero-tolerance) rules assign equal veto power to every evaluation dimension; this inflexible logic rejects an entire image due to minor, tolerable deviations, as illustrated in Figure 1 (b). Consequently, there lacks an explicit mapping between observable errors to justify image deliverability.

To address this gap, we introduce **InfoDiag-Bench**, an application-oriented benchmark comprising 663 test cases across 14 infographic generation scenarios. We design a scenario-aware evaluation framework where each scenario is equipped with independent evaluation criteria and configurable error types. The framework jointly assesses information fidelity and visual quality, alongside evidence-grounded error types. Each error type specifies a recognizable error type and the conditions under which it should be reported, with evidence identifying the relevant discrepancy in the generated image. Explicit acceptance rules then distinguish delivery-blocking error types from tolerable ones and map diagnostic results to image-level deliverability judgments. This design makes the rejection inspectable while supporting systematic analysis of bottleneck errors across models and application scenarios.

Our contributions are threefold:

- We introduce InfoDiag-Bench, an application-oriented benchmark comprising 663 test cases across 14 infographic generation scenarios. Organized around practical needs, these scenarios capture distinct requirements for factual content, topological structure, and visual presentation.
- We develop scenario-aware evaluation framework that jointly assesses information fidelity and visual quality. Central to the framework are evidence-grounded error types that distinguish delivery-blocking errors from tolerable deviations. By linking error diagnosis to

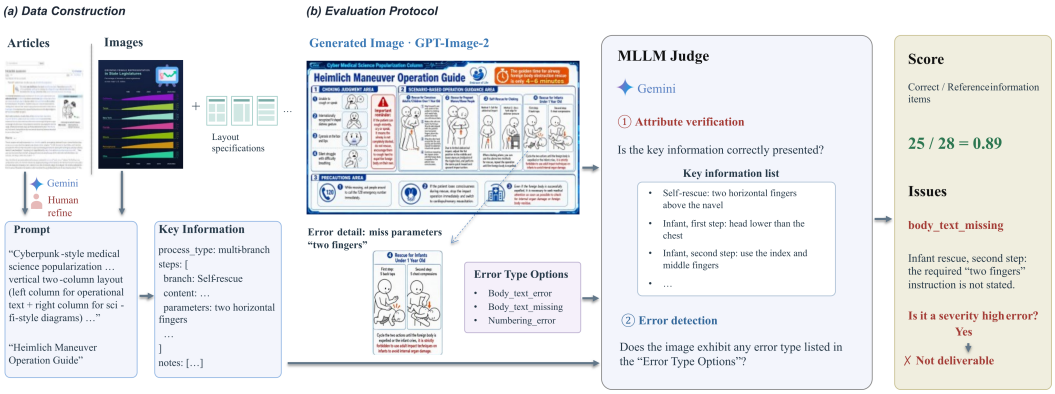


Figure 2: Overview of the InfoDiag-Bench framework. (a) Data construction: a dual-track LLM–human pipeline transforms real-world web articles and reference images into generation prompts alongside structured key information. (b) Evaluation protocol: an MLLM verifies the generated image against the key information and detects specific issues from predefined error options, yielding both scores and diagnostic results. Based on the predefined severity mapping, triggering any high-severity error renders the image undeliverable.

explicit acceptance criteria, the framework supports image-level deliverability judgments conforming to human expectations.

- We conduct comprehensive experiments on 10 state-of-the-art T2I models. By effectively isolating delivery-blocking errors from tolerable deviations, we uncover the bottlenecks in current generation models: once minor visual or stylistic deviations are appropriately tolerated, the primary cause of delivery failure is the models' barrier to preserve required topological relationships among infographic elements.

2 Related Works

2.1 Infographics Generation Models

Modern T2I generation models based on latent diffusion (Rombach et al., 2022) have shown significant ability in instruction following (Labs, 2024; 2025; Team, 2025; Wu et al., 2025). However, infographic generation remains challenging because of its strict requirements for dense text and consistent mappings between visual geometry. SenseNova-U1 (Diao et al., 2026) and BizGen (Peng et al., 2025b) advances text-rich infographic generation by architecture optimization. Zhuo et al. (2026) train a unified model that integrates a VLM with FLUX.1 Kontext. Advanced closed-source models such as Nano Banana Pro (Gemini 3 Pro Image) (Google, 2026b), Seedream 5.0 Pro (Seed, 2026b), and GPT-Image-2 (OpenAI, 2026) have also demonstrated strong text-rendering capabilities. Despite these advances, infographic generation remains susceptible to errors.

2.2 Text-to-Image Evaluation Benchmarks

General T2I benchmarks typically evaluate whether generated images satisfy prompt-derived conditions. Traditional metrics employ on vision-language models and VQA (Hessel et al., 2021; Radford et al., 2021; Han et al., 2024; Hu et al., 2023; Cho et al., 2024; Ghosh et al., 2023; Huang et al., 2023; Kajić et al., 2024; Jiang et al., 2024) In recent years, an increasing number of methods have adopted large multimodal-based automatic evaluators for fine-grained assignment (Lin et al., 2024; Liang et al., 2025; Wang et al., 2025b). Dedicated evaluation of infographic generation, however, remains less developed. StructBench (Zhuo et al., 2026) and ChartGalaxy (Li et al., 2026b) address complex charts or diagrams, emphasizing textual and numerical accuracy while providing limited assessment of real-world applicability. BizGenEval (Li et al., 2026a) and BizGen (Peng et al., 2025b) focus on text-layout-rich infographics, but their general evaluation metrics fail to identify whether an image can be delivered. Most notably, IGenBench (Tang et al., 2026) introduces Q-ACC and I-ACC to

assess correctness at the checkpoint and image levels, respectively. However, its "all-checks-pass" strategy fails to distinguish fatal semantic collapses from minor aesthetic flaws. In contrast, our benchmark defines scenario-specific evaluation dimensions and diagnostic error types to explicitly distinguish delivery-blocking errors from tolerable deviations.

3 INFODIAG-BENCH

We propose InfoDiag-Bench, a scenario-aware benchmark to evaluate the deliverability of infographics generated by T2I models. As shown in Fig.2, we collect data from real sources, such as web articles and reference images, spanning 14 scenarios closely aligned with practical application. We employ a dual-track prompt generation pipeline to convert these sources into detailed generation prompts, accompanied by structured "key information" annotation that explicitly capture the essential semantic units and layout constraints. Then, our evaluation framework utilizes MLLMs to jointly report quality scores and evidence-grounded diagnostic issues. Finally, by mapping these detected issues against predefined severity labels and explicit acceptance criteria, we derive image-level deliverability decisions.

3.1 DATASET CONSTRUCTION

InfoDiag-Bench is grounded in real-world, high-density infographic applications, drawing on information-rich web articles and publicly available infographic images. Informed by BizGenEval (Li et al., 2026a) and IDEA-Bench (Liang et al., 2025), we categorize our data according to different information organization modalities. The dataset encompasses five high-level categories: Quantitative Charts, Sequential Flows, Grouped Organizations, Relation Networks, and Composite Presentations, spanning 14 specific application scenarios (Figure 1(a)). We construct the dataset through a five-stage, human-in-the-loop pipeline.

(1) Data Collection. We collect data from diverse origins, including the open-source dataset ChartGalaxy (Li et al., 2026b), professional infographic platforms, and targeted web crawling using scenario-specific search queries. Among the final 663 test cases, image-derived prompts account for approximately 29%, while article-derived prompts make up 71%. Each scenario covers multiple topics, as detailed in Appendix A.1.1.

(2) Prompt Construction. We first design scenario-specific prompt templates (detailed in Appendix A.1.2) and adopt separate prompt construction procedures for articles and reference images. For articles, we employ Gemini 3.1 Pro Preview (Google, 2026a) to summarize the core information. To increase diversity, we design 10 layout specifications per scenario, covering visual styles and spatial structures, and dynamically assign them to content items. The summarized information and assigned layout specifications are then integrated into prompts using the corresponding templates. For reference images, Gemini 3.1 Pro Preview follows the corresponding templates to produce detailed generative descriptions for image construction. Human annotators then verify and refine the resulting prompts to ensure information completeness.

(3) Entity Pseudonymization (Anti-memorization). For the purpose of privacy protection and to prevent model cheating, we replace specific country, organization, brand, and character names with generic terms while preserving the original information structure and logical relationships.

(4) Filtering. *Format checking:* We utilize regular expressions to eliminate prompts deviating from required structural formats or containing hallucinatory introductory texts. *Logical coherence:* We assess logical consistency, such as conflicts between numerical values and geometric layout constraints, through LLM-based verification followed by sample human spot checks.

(5) Deduplication. We remove exact duplicates via MD5 hashing and conduct semantic deduplication using `all-MiniLM-L6-v2` with an empirically set cosine similarity threshold of 0.8, ensuring high diversity across the benchmark.

3.2 Evaluation Framework

3.2.1 Information Annotation

Infographic prompts frequently interleave textual content, numerical facts, structural relationships, and visual requirements. Information annotation reorganizes these information into structured reference data for individual dimension evaluation, which is called "key information". Specifically, the "key information" inclues text, entities, numerical data, topological relationships and layout requirements.

For each scene s, we define an extraction template τ_s. Then we use Gemini 3.1 Pro Preview (Google, 2026a) to perform the "key information" extraction. To make sure "key information" coverage and consistency of the prompt, we conduct character matching check and artifical verification. An example is shown in Figure 2 "key information".

3.2.2 Scenario-Specific Evaluation Dimensions

We organize the evaluation dimensions into two groups: Information fidelity (IF) and Visual Quality (VQ). IF assesses whether the generated image accurately presents the information and satisfies the content requirements specified in the prompt, which consists 2-3 dimensions and varies across scenarios. For example, Financial Summary evaluates financial data and trend accuracy, chart-axis correctness, etc. PPT evaluates title accuracy, chart and table accuracy, etc. VQ assesses the visual presentation of the information through scenario-relevant dimensions such as layout hierarchy, readability, and theme relevance. The complete dimension registry are provided in Appendix A.2.1.

For IF dimensions, the Multimodal Large Language Model (MLLM) compares key information items with the generated image and identifies each item is correct or not. The accuracy score is computed as

$$s_d = \frac{C_d}{N_d},$$

where d is dimension, C_d is the number of correctly presented items and N_d is the total number of check items in this dimension. For VQ dimensions, we give score from 1-5. The detail criteria are provided in Appendix A.2.1.

3.2.3 Error Types and Mapping

To identify specific failures in infographic generation, we define 45 reusable error types, denoted as the set $\mathcal{E}$, and map each fine-grained error type $e \in \mathcal{E}$ to macro-level evaluation dimensions.

Initial error type pool. We generate images using GPT-Image-2 (OpenAI, 2026), Seedream 5.0 Pro (Seed, 2026b), and Qwen-Image-2512 (Wu et al., 2025), representing models with different capabilities, across 30 prompts for each scenario. Expert reviewers then analyze the generated images and annotate the error types to form the initial error type pool. Each error type e has a definition, an evidence requirement specifying its triggering conditions, and a severity label governing its impact on deliverability.

Open discovery. To enhance the comprehensiveness of error type, we perform an LLM-assisted refinement across 663 testcases. For each case, we provide an LLM with the prompt, annotated key information, scenario-specific dimension definitions, and the initial error type pool. The LLM identifies candidate error types, prioritizing existing ones while proposing new types when necessary. Furthermore, instead of restricting each error type to a single dimension, the LLM performs a multi-label mapping, explicitly linking each error type to all dimensions responsible for detecting it. For example, illegible text may penalize both textual accuracy and readability.

Human verification and freezing. Human annotators review the proposed error types and mapping assignments. Each proposed new error type and its mapping assignments are either accepted, merged into an existing error type, or rejected. Annotators also verify the severity labels, distinguishing delivery-blocking errors (high severity) from tolerable deviations (low severity). Appendix Table A.2.2 provides a detailed mapping from the error type to the evaluation dimensions for each scenario.

Table 1: Definitions and severity labels of diagnostic error types. High-severity errors block delivery, whereas low-severity errors do not.

Error type	Definition	Severity
Text hierarchy confusion	The visual priorities of elements such as the title, explanatory text, main chart, axes, and labels are disorganized, resulting in an unclear reading path.	high
UI component detail deviation	The detailed design or placement of UI components, such as card shadows, corner radii, icon styles, avatars, buttons, and input fields, deviates from the reference requirements.	low
Visual style mismatch	The poster's color scheme, icons, visual elements, or overall atmosphere does not match the reference style, or the image lacks the characteristics of a travel-guide poster.	high

Three annotators experienced in infographic generation independently annotated severity labels for error types. The exact agreement rate across all three annotators was 71.1%, which shows high agreement in the severity labels. Final severity labels were determined by majority vote. The reviewed error type definitions, mapping assignments, and severity labels were then frozen for evaluation. The final error type pool includes 45 error types including `axis_error`, `axis_missing`, `body_text_error`, `entity_relationship_mismatch`, etc. Selected error types and their severity labels are presented in Table 1, while the complete set is provided in Appendix Table A.2.2.

3.2.4 MLLM JUDGE AND DELIVERABILITY

We use MLLM to evalute images from different dimensions. The input inclues the generated image, relavant key information, dimension criteria and error types, then the evaluator jointly produces score and evidence-grounded diagnostic issues drawn from the allowed error types. The score measures the degree to which the image satisfies the dimension-specific rubric, while the issues identify concrete failures and their supporting evidence.

For a generated image i in scenario s, we evaluate it across the dimensions $\mathcal{D}_s$. For each dimension d, let $v_{i,d} = 1$ if no high-severity error type is detected, and $v_{i,d} = 0$ otherwise. Image-level deliverability is defined as

$$V_i = \prod_{d \in \mathcal{D}_s} v_{i,d}, \tag{1}$$

Thus, an image is considered deliverable only if no high-severity issue is detected in any dimension. The model-level deliverability rate is the proportion of benchmark samples satisfying this condition.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Evaluated Models. We evaluate 10 representative T2I models spanning different architectures and parameter scales, including 5 open-weight and 5 closed-source models. The five open-weight models are Qwen-Image-2512 (Wu et al., 2025), Z-Image-Turbo (Team, 2025), SenseNova-U1-8B-MoT-Infographic-V3 (SenseNova, 2026), FLUX.1-dev (Labs, 2024), and FLUX.2-dev (Labs, 2025). We also evaluate five closed-source commercial models: Seedream 5.0 Pro (Seed, 2026b), Seedream 4.5 (Seed, 2025), Gemini 3.1 Flash Image (Nano Banana 2) and Gemini 3 Pro Image (Nano Banana Pro) (Google, 2026b), and GPT-Image-2 (OpenAI, 2026). The table labels "Gemini-3.1-Flash" and "Gemini-3-Pro" refer to the two image-generation models, respectively, rather than the text-only models.

Evaluation Details. We use Gemini 3.1 Pro Preview (Google, 2026a) as the MLLM evaluator and report image-level deliverability for each scenario (Table 2) and average quality scores (Table 3). Furthermore, we use different MLLMs as evaluators on sampled datasets; the results are provided in Appendix A.3.1.

Table 2: Image-level deliverability (%; higher is better). Scenarios are ordered by their mean deliverability across models. Bold indicates the best displayed value in each column, including ties. Cell shading increases with the value.

Model	**Overall**	**Application scenario** (sorted by average)													
	$D \uparrow$	**CR**	**PP**	**SD**	**DJ**	**KL**	**TT**	**PC**	**FS**	**EN**	**PPT**	**OS**	**CS**	**TP**	**SG**
GPT-Image-2	**71.4**	**10.3**	**71.4**	**55.8**	**68.2**	**64.7**	61.1	**81.6**	**90.4**	**70.8**	**73.7**	**68.2**	**89.7**	**96.9**	**89.7**
Gemini-3-Pro	56.3	5.0	39.3	38.1	37.8	55.1	61.5	60.4	68.0	45.8	62.1	65.9	79.5	96.8	74.6
Gemini-3.1-Flash	45.5	5.0	21.4	35.7	42.9	27.5	**66.7**	42.6	49.0	36.1	63.2	56.8	61.5	45.2	78.6
Seedream-5.0	26.2	0.0	10.7	21.4	24.4	13.0	22.6	29.2	38.0	36.1	20.7	39.5	10.3	27.3	56.4
SenseNova-U1-v3	20.2	2.5	3.6	9.8	5.3	13.0	25.8	33.3	7.7	30.6	27.6	31.8	23.1	0.0	41.1
Seedream-4.5	16.1	0.0	17.9	5.0	0.0	5.9	10.7	4.1	2.0	28.6	20.7	6.8	20.5	20.7	64.9
Qwen-Image-2512	2.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.2	1.7	0.0	0.0	0.0	24.5
FLUX.2-dev	1.7	0.0	0.0	0.0	0.0	1.4	0.0	2.3	0.0	7.0	0.0	2.3	0.0	4.5	2.0
Z-Image-Turbo	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.8	0.0	2.3	0.0	4.5	0.0
FLUX.1-dev	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Average	**24.1**	**2.3**	**16.4**	**16.6**	**17.9**	**18.1**	**24.8**	**25.3**	**25.5**	**26.2**	**27.0**	**27.4**	**28.5**	**29.6**	**43.2**

CR: Character relationships; **PP**: Promotional posters; **SD**: Statistical dashboards; **DJ**: Data journalism; **KL**: Knowledge lists; **TT**: Temporal trends; **PC**: Pros–cons matrices; **FS**: Financial summaries; **EN**: Event notices; **PPT**: Presentation slides; **OS**: Organizational structures; **CS**: Conference summaries; **TP**: Tour posters; **SG**: Sequential operation guides. Each rate is pass/(pass+fail); incomplete evaluations are excluded. Overall pools completed images across scenarios, rather than averaging scenario rates. Average is the unweighted mean across the ten models.

Table 3: Benchmark result scores (0–5; higher is better). Models are shown as columns. IF: information fidelity; VQ: visual quality. Overall is the equal-weight mean of IF and VQ.

Quality metric	**GPT-Image-2**	**Gemini-3-Pro**	**Gemini-3.1-Flash**	**Seedream-5.0**	**SenseNova-U1-v3**	**Seedream-4.5**	**Qwen-Image-2512**	**FLUX.2-dev**	**Z-Image-Turbo**	**FLUX.1-dev**	**Average**
Overall	**4.590**	4.461	4.207	3.987	3.632	3.536	2.404	2.574	2.102	1.255	**3.275**
IF	**4.533**	4.384	4.148	3.896	3.561	3.414	2.030	2.237	1.639	0.542	**3.038**
VQ	**4.648**	4.538	4.266	4.077	3.704	3.658	2.779	2.911	2.565	1.968	**3.511**

Scores are transcribed from the HTML report at its displayed precision. Each metric uses valid-record weighting after conversion to a common 5-point scale. Average is the unweighted mean across models; bold marks the best model in each metric.

4.2 Main Results

As shown in Tables 2 and 3, we evaluate 10 representative T2I models across 14 scenarios on InfoDiag-Bench, reporting both deliverability and quality scores. GPT-Image-2 achieves the highest score on both overall metrics, followed by Gemini-3-Pro and Gemini-3.1-Flash. Open-weight models generally lag behind their closed-source counterparts. SenseNova-U1-v3, the strongest open-weight model in our evaluation, achieves a deliverability rate of 20.2%, compared with 71.4% for GPT-Image-2, highlighting a substantial gap between the best-performing models in the two groups.

The expressions of topological relationships among elements remain challenging As illustrated in Table 2, all models achieve low deliverability in the character relationship (CR) scenario. This finding indicates that current models struggle to accurately depict complex relationships between entities. Test cases in this scenario often include multiple lines connecting characters to communicate their relationships intuitively. We further analyze the most frequent error types, as shown in Figure 3. The most frequent error types are `relationship_edge_error`, `relationship_edge_missing`, and `entity_relationship_mismatch`, indicating that models often generate edges connected to the wrong characters or omit required edges. Although GPT-Image-2 attains the highest CR deliverability (10.3%), relationship-edge errors account for 67.5% of all annotated errors in its CR outputs. Thus, its failed CR cases are dominated by structural relationship errors. The examples further reveal directionality errors and incorrect entity binding. These findings highlight a significant challenge for T2I models.

High-density and precise rendering remain challenging We use prompt length as a proxy for information density. Within each scenario, we rank prompts by length and divide them into three equal-sized groups: sparse, medium, and dense. As illustrated in Table 4, deliverability is generally higher for sparse prompts than for dense prompts. This finding indicates that T2I models still

Table 4: Average delivery rates (%) of different models across prompt information density levels. Retention measures the ratio of the dense-group to sparse-group deliverability rate (%).

Difficulty	GPT	Gemini-3-Pro	Gemini-3.1-Flash	Seed5.0	SenseNova-U1-V3	Seedream4.5	Qwen-25	Flux2	Z-Image	Flux1
Sparse	77.83	65.61	54.75	37.44	26.76	24.88	4.81	4.39	1.45	0.00
Medium	72.85	59.46	46.12	28.30	21.13	15.94	3.43	0.50	0.50	0.00
Dense	63.47	43.89	35.07	12.32	12.62	7.37	0.00	0.00	0.00	0.00
Retention (%) ↑	81.55	66.90	64.05	32.91	47.16	29.62	0.00	0.00	0.00	–

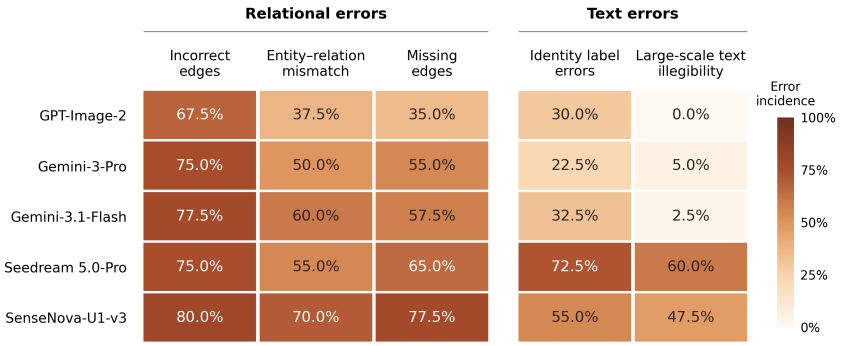


Figure 3: High-frequency error types in the character relationship scenario.

struggle with high-density, precise rendering. We further calculate retention, defined as the ratio of dense-group to sparse-group deliverability. Its ranking is broadly consistent with that of absolute deliverability in Table 2, except that SenseNova-U1-V3 preserves deliverability better than Seedream 5.0 as information density increases. We believe that robustness to increasing information density is also an important evaluation criterion.

Qualitative Comparisons Figure 4 compares GPT-Image-2, Gemini-3.1-Flash, and Seedream 5.0 across the Character Relationship and Pros & Cons Matrix scenarios. In the Character Relationship example, all three outputs contain delivery-blocking relationship errors (incorrect or missing edges); Seedream 5.0 additionally exhibits body-text errors and large-scale text garbling. In the Pros & Cons Matrix example, the GPT-Image-2 and Gemini-3.1-Flash outputs are deliverable because no blocking errors are detected, whereas the Seedream 5.0 output is non-deliverable because required list content is missing. These cases illustrate how the proposed diagnostic error types connect local evidence to image-level deliverability decisions.

4.3 Human Evaluation

We further assess the alignment between our model evaluation and human judgments. We conduct a human evaluation on a subset of the dataset, assessing one dimension in each of four scenarios. These comprise two Information Score dimensions, `financial_summary_insights_text_accuracy` and `ppt_chart_table_accuracy`, and two Visual Quality dimensions, `knowledge_list_layout_hierarchy` and `tour_poster_textlegibility`. We additionally assess deliverability in the PPT and Financial Summary scenarios. We randomly sample 25 cases per scenario and evaluate the outputs of all 10 T2I models, yielding 1,000 images across the four scenarios. Three professional annotators with extensive image-annotation experience independently assess the images under blinded conditions. For Information Score dimensions, annotators answer yes/no for each reference-information item, after which we automatically compute an accuracy score. For Visual Quality dimensions, annotators assign each image a score from 1 to 5. For deliverability, annotators provide a binary judgment of whether each image is deliverable. The human evaluation results are presented in Table 5. The ICC(A,1) is 0.723, indicating strong agreement among annotators on scores. The mean Pearson's correlation across the four dimensions (1,000 test points) is 0.828, and the balanced accuracy rates for deliverability in both scenarios exceed 0.75, further supporting alignment with human judgments. The results show strong alignment between automatic and human evaluations,

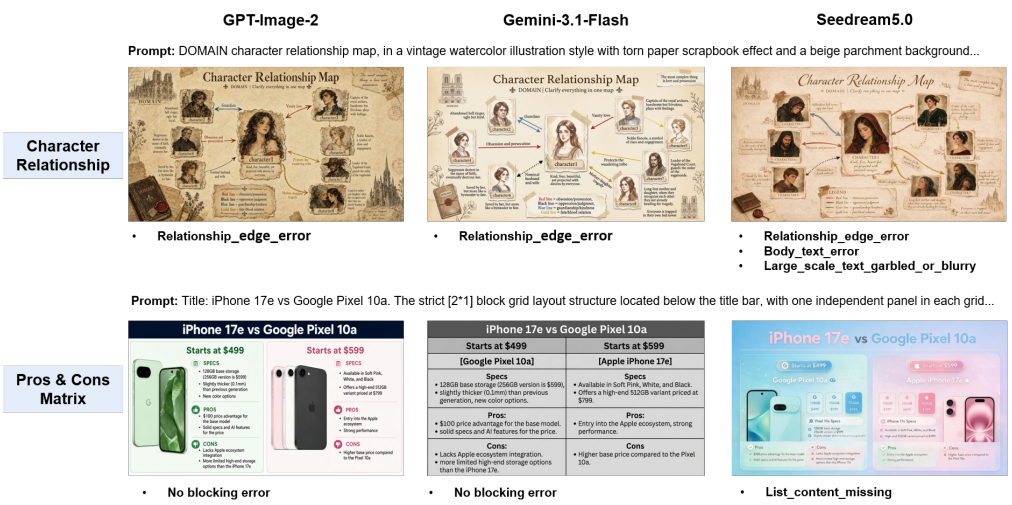


Figure 4: Qualitative comparisons across representative infographic-generation scenarios.

Table 5: Consistency between model scores and human evaluations in four scenarios. ICC(A,1) measures inter-annotator agreement; Pearson's r is the mean correlation between model and human scores; BA denotes the avarage balanced accuracy of model deliverability judgments to human judgment.

ICC(A,1)	Pearson's r	Financial Summary deliverability BA	PPT deliverability BA
0.723	0.828	0.851	0.765

together with high inter-annotator agreement, supporting the reliability of the evaluation protocol on the assessed subset.

4.4 Comparison Experiments

To validate the effectiveness of our deliverability rules, we conduct comparison experiments with IGenBench (Tang et al., 2026). We sample 57 prompts from its dataset and generate images using GPT-Image-2 (OpenAI, 2026), Seedream 5.0 Pro (Seed, 2026b), and Qwen-Image-2512 (Wu et al., 2025). We then evaluate the images using both our protocol and the IGenBench method. Doubao Seed 2.0 Pro (Seed, 2026a) serves as the MLLM evaluator in both experiments. Two annotators independently judge the deliverability of each image, achieving a Cohen's κ of 0.929, which indicates high consistency. We identify the test samples with inter-annotator disagreement and calculate Cohen's κ between each method and the human evaluation on 99 images. Our method achieves 0.755, whereas IGenBench achieves 0.689, indicating that our method better diagnoses deliverability.

5 Conclusion

We introduce InfoDiag-Bench, a scenario-aware benchmark for evaluating the deliverability of generated infographics across 14 practical application scenarios. Beyond measuring information fidelity and visual quality, the framework provides evidence-grounded error diagnoses. We define a reusable error-type pool and severity mapping that distinguish delivery-blocking failures from tolerable deviations, and applies explicit acceptance rules to derive image-level deliverability judgments. Evaluations of 10 representative T2I models reveal two findings: state-of-the-art models struggle to preserve complex entity relationships, and deliverability generally decreases as prompt information density increases. These findings underscore the need to improve structural correctness and precise information rendering alongside visual quality. Our future work will expand scenario coverage and investigate how diagnostic feedback can guide image correction.

6 AI USE STATEMENT

In this work, we used generative AI tools to assist with paper research, manuscript refinement, prompt construction, annotation generation, MLLM-based judgement, and code development. We have not used generative AI tools for the evaluation framework design, prompt-template and annotation-template formulation, pipeline construction, and result analysis. All AI-produced outputs have undergone rigorous manual review by the authors and annotators to verify correctness. The authors independently examined each referenced paper to confirm authenticity and relevance. Both authors and three dedicated annotators jointly jointly validated prompt logic and annotation fidelity for the benchmark dataset. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

ByteDance. Doubao seed 2.0 lite. https://seed.bytedance.com/en/seed2, 2026. Model checkpoint: doubao-seed-2-0-lite-260428, Volcano Engine API documentation, Accessed: 2026-09-26.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024.

Jaemin Cho, Yushi Hu, Jason Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *International conference on learning representations*, volume 2024, pp. 15625–15645, 2024.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, et al. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Google. Gemini 3.1 pro preview model, 2026a. URL https://deepmind.google/models/gemini-image/pro.

Google. Gemini 3 pro image and gemini 3.1 flash image, 2026b. URL https://deepmind.google/models/gemini-image/pro.

Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chunle Guo, et al. Evalmuse-40k: A reliable and fine-grained benchmark with comprehensive human annotations for text-to-image generation model evaluation. *arXiv preprint arXiv:2412.18150*, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021. doi: 10.18653/v1/2021.emnlp-main.595.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *2023 ieee/cvf international conference on computer vision (iccv)*, pp. 20349–20360. IEEE, 2023.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. Genai arena: An open evaluation platform for generative models. *Advances in Neural Information Processing Systems*, 37:79889–79908, 2024.

Ivana Kajić, Olivia Wiles, Isabela Albuquerque, Matthias Bauer, Su Wang, Jordi Pont-Tuset, and Aida Nematzadeh. Evaluating numerical reasoning in text-to-image models. *Advances in Neural Information Processing Systems*, 37:42211–42224, 2024.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2`, 2025.

Hao Li, Yang Zou, Ying Wang, Orchid Majumder, Yusheng Xie, R. Manmatha, Ashwin Swaminathan, Zhuowen Tu, Stefano Ermon, and Stefano Soatto. On the scalability of diffusion-based text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9400–9409, June 2024.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Arsh Koneru, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Reflect-dit: Inference-time scaling for text-to-image diffusion transformers via in-context reflection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15657–15668, October 2025.

Yan Li, Zezi Zeng, Ziwei Zhou, Xin Gao, Muzhao Tian, Yifan Yang, Mingxi Cheng, Qi Dai, Yuqing Yang, Lili Qiu, et al. Bizgeneval: A systematic benchmark for commercial visual content generation. *arXiv preprint arXiv:2603.25732*, 2026a.

Zhen Li, Duan Li, Yukai Guo, Xinyuan Guo, Bowen Li, Lanxi Xiao, Shenyu Qiao, Jiashu Chen, Zijian Wu, Hui Zhang, et al. Chartgalaxy: A dataset for infographic chart understanding and generation. In *International Conference on Learning Representations*, volume 2026, pp. 52484–52546, 2026b.

Chen Liang, Lianghua Huang, Jingwu Fang, Huanzhang Dou, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Junge Zhang, Xin Zhao, and Yu Liu. Idea-bench: How far are generative models from professional designing? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18541–18551. IEEE, 2025.

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

Zeyu Lu, Zidong Wang, Di Huang, Chengyue Wu, Xihui Liu, Wanli Ouyang, and Lei Bai. Fit: Flexible vision transformer for diffusion model. *arXiv preprint arXiv:2402.12376*, 2024.

OpenAI. Gpt-image-2, 2026. URL `https://developers.openai.com/api/docs/models/gpt-image-2`.

OpenAI. Gpt-5.5. `https://openai.com/index/introducing-gpt-5-5`, apr 2026. Accessed: 2026-09-26.

Gao Peng, Le Zhuo, Dongyang Liu, DU DU, Xu Luo, Longtian Qiu, Yuhang Zhang, Rongjie Huang, Shijie Geng, Renrui Zhang, et al. Lumina-t2x: Scalable flow-based large diffusion transformer for flexible resolution generation. In *International Conference on Learning Representations*, volume 2025, pp. 37254–37292, 2025a.

Yuyang Peng, Shishi Xiao, Keming Wu, Qisheng Liao, Bohan Chen, Kevin Lin, Danqing Huang, Ji Li, and Yuhui Yuan. Bizgen: Advancing article-level visual text rendering for infographics generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23615–23624. IEEE, 2025b.

Weimin Qiu, Jieke Wang, and Meng Tang. Self-cross diffusion guidance for text-to-image synthesis of similar subjects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23528–23538, June 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

ByteDance Seed. Seedream 4.5, 2025. URL https://seed.bytedance.com/en/seedream4_5.

ByteDance Seed. Seed 2.0 pro, 2026a. URL https://seed.bytedance.com/en/seed2.

ByteDance Seed. Seedream 5.0 pro, 2026b. URL https://seed.bytedance.com/en/seedream5_0_pro.

SenseNova. Neo-unify: Building native multimodal unified models end to end, 2026. URL https://huggingface.co/blog/sensenova/neo-unify.

Yinghao Tang, Xueding Liu, Boyuan Zhang, Tingfeng Lan, Yupeng Xie, Jiale Lao, Yiyao Wang, Haoxuan Li, Tingting Gao, Bo Pan, et al. Igenbench: Benchmarking the reliability of text-to-infographic generation. *arXiv preprint arXiv:2601.04498*, 2026.

Z-Image Team. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Andrew Z. Wang, Songwei Ge, Tero Karras, Ming-Yu Liu, and Yogesh Balaji. A comprehensive study of decoder-only llms for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28575–28585, June 2025a.

Jiarui Wang, Huiyu Duan, Yu Zhao, Juntong Wang, Guangtao Zhai, and Xiongkuo Min. Lmm4lmm: Benchmarking and evaluating large-multimodal image generation with lmms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17312–17323, October 2025b.

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

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *The Thirteenth International Conference on Learning Representations*, 2025.

Le Zhuo, Songhao Han, Yuandong Pu, Boxiang Qiu, Sayak Paul, Yue Liao, Yihao Liu, Jie Shao, Xi Chen, Si Liu, et al. Factuality matters: When image generation and editing meet structured visuals. In *International Conference on Learning Representations*, volume 2026, pp. 28249–28275, 2026.

A Appendix

A.1 Data Collection

Note that the data construction process described here is the initial step, so the quantity is greater than the final test set size.

A.1.1 Data Collection Details

We collect source materials from the open-source ChartGalaxy dataset (Li et al., 2026b), professional infographic platforms, and targeted web crawling. The data topics and their sources are listed below.

Data Journalism Topics (165 records)
Election security, voter participation, electoral integrity, gun-control opinion; air quality, renewable energy, reforestation, biodiversity, coral reefs, and bird recovery; inequality, gender representation, poverty, maternal health, telehealth, education, employment, and public debt.

Statistical Dashboard Topics (10 records)
Trade and container throughput; intellectual-property compliance; startup employment; vehicle maintenance costs; green-building policy adoption; transport fuel savings; AI-assisted development; GDPR/CCPA awareness; bike-share and delivery-time comparisons.

Financial Summary Topics (24 records)
Company and sector performance, including cigarette sales, taxi revenue, and electric-vehicle market share; agricultural trade; public spending and government expenditure; housing-price indices, GDP per capita, investment, revenue loss, and other financial indicators.

Temporal Trend Topics (70 records)
Multi-year trends and timelines for turnout, air quality, GDP and income, renewable energy, coral reefs and wildlife, and livestock losses.

Conference Topics (57 records)
AI competition and leadership; management hierarchies, workflows, and digital employees; manufacturing and global trade; interest rates, exchange rates, central-bank policy, public talks, podcasts, and international policy forums.

Knowledge List Topics (70 records)
Public health, including antimicrobial resistance, diabetes, obesity, and open health data; macroeconomics and trade; food and agriculture; science and technology policy; standards and governance; energy and climate; space and astronomy.

PPT Topics (63 records)
Corporate briefings, policy interpretation, academic presentations, teaching materials, industry data analysis, process and workflow briefings, and project-achievement reporting; examples include project scheduling, risk quantification, procurement strategy, clinical participation paths, and cost/probability analysis.

Character Relationship Topics (80 records)
Superhero comic and cinematic universes, live-action television and streaming series, and space-opera IP. Relationship maps cover character identity, faction, family, romance, alliance, mentorship, hostility, and power relations.

Organizational Structure Chart Topics (74 records)
Corporate hierarchy (66), government structure (6), and board governance (2), including reporting chains from executives to departments and roles, institutional units, and governance relationships.

Pros–Cons Matrix Topics (51 records)
Consumer-electronics comparisons, including tablets, phones, and devices; automotive features and trims; SaaS and cloud pricing; competitive feature scorecards and reviews; business SWOT and strategic options.

Sequential Operation Guide Topics (60 records)
Emergency and public-health procedures, including CPR and the Heimlich maneuver; cooking and food preparation; software and developer workflows; data, analytics, and cloud operations; device setup and account security; creative, publishing, collaboration, service, and offline procedures.

Event Notice Topics (70 records)
Academic conferences (24), government and community events (19), campus student activities (12), industry seminars (11), and online webinars (4), including workshops, lectures, training, and registration announcements.

Tour Poster Topics (32 records)
Tourism promotion (12), city short itineraries (4), cultural-heritage routes (4), nature-park routes (4), regional road trips (3), island and coastal escapes (3), and event or seasonal travel (2). Content covers route order, destination highlights, transport arrangements, practical visitor notes, cultural background, natural features, conservation guidance, and seasonal timing.

Promo Poster Topics (46 records)
Product promotion (12), film promotion (11), brand campaigns (11), television-series promotion (10), and cultural-IP promotion (2).

Sources The following is a list of data sources.

Source	Relevant Content	Public Website
ChartGalaxy: A Dataset for Infographic Chart Understanding and Generation (Li et al., ICLR 2026)	Data Journalism and infographic-chart data	`https://huggingface.co/datasets/ChartGalaxy/ChartGalaxy`
World Health Organization (WHO)	Public-health knowledge	`https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance`
Centers for Disease Control and Prevention (CDC)	Public-health open data	`https://data.cdc.gov/`
National Aeronautics and Space Administration (NASA)	Space and astronomy	`https://science.nasa.gov/`
Food and Agriculture Organization of the United Nations (FAO)	Food and agriculture	`https://www.fao.org/`
Board of Governors of the Federal Reserve System (Fed)	Monetary-policy meeting records	`https://www.federalreserve.gov/monetarypolicy/fomcminutes20240131.htm`
World Trade Organization (WTO)	International trade conferences	`https://www.wto.org/english/thewto_e/minist_e/mc13_e/mc13_e.htm`
Organisation for Economic Co-operation and Development (OECD)	Artificial intelligence and policy forums	`https://www.oecd.org/en/topics/artificial-intelligence.html`

A.1.2 PROMPT TEMPLATES

Infographic prompt generation requires precise logical constraints and structured layout descriptions. Figures 5–11 present the prompt-generation templates used for the seven scenario families represented in the source template collection.

Generation Instruction for LLMs (Character Relationship)

I. System role

You are an expert character-relationship infographic prompt editor. Rewrite an existing English character-relationship prompt by changing only the overall visual-style description and the connection-color rules.

II. Preservation constraints

1. Preserve every relationship fact, character role, hierarchy tier, relative position, layout rule, and diagram constraint.
2. Do not add characters, relationship edges, modules, or factual claims.
3. Keep the output as a relationship infographic rather than a narrative scene illustration.
4. Do not include URLs, source links, commentary, JSON, or Markdown fences.
5. Use English only and return only the rewritten generation prompt.
6. Use determined instructions; avoid optional alternatives.

III. Required output structure

Overall Style.: Specify one coherent visual language, palette, typography system, portrait treatment, background treatment, and information density.
Title and Labels.: Preserve all exact titles, subtitles, character names, role labels, and relationship labels.
Character Placement.: Preserve the center character, surrounding tiers, relative coordinates, grouping, and reading order.
Relationship Encoding.: Enumerate every required edge, direction, endpoint, label, line type, and label placement.
Connection Colors.: Assign one unambiguous color rule to each relationship class without changing its semantics.
Topology Constraints.: Require non-crossing connectors where possible, correct arrow direction, attachment to node boundaries, and no omitted or duplicated edges.

IV. Output contract

Produce a single self-contained English prompt. The visual-style and connection-color wording may change; all semantic and topological information must remain unchanged.

Figure 5: Prompt-generation template for the Character Relationship scenario.

Generation Instruction for LLMs (Event Notice)

I. System role

You are an expert event-notice infographic prompt designer. Rewrite the supplied event notice as a complete, high-density English infographic-generation prompt.

II. Factual constraints

1. Use only facts in the source prompt and item metadata. Preserve every clearly specified title, organizer, date, time, venue, speaker, fee, registration detail, agenda item, audience note, accessibility note, contact, and footer.
2. Do not invent organizations, logos, URLs, QR destinations, or missing event details.
3. Never expose source-material links. A QR element may only be a generic registration or event-access placeholder.
4. Output English only and return only the redesigned prompt, without JSON, Markdown, explanations, or meta-level assurances.
5. Use concise, determined wording. Do not offer alternatives or specify numeric color codes.

III. Required prompt structure

Opening.: "Create a complete [event type] event notice poster titled '[exact title]'."
Visual Style.: Fix the announcement style, mood, natural-language palette, typography hierarchy, icon style, spacing, density, and readability.
Identity and Title Band.: Specify title, subtitle or theme, organizer line, supported logo placeholders, registration QR placeholder, and corner treatment.
Main Layout.: Use a top-to-bottom poster with named modules and a central card grid. Assign fixed icons: calendar to date, clock to time, map pin to venue, ticket to registration, people to audience, microphone to speakers, checklist to agenda, envelope to contact, accessibility to accommodations, and globe to online access.
Content Modules.: When supported by the source, describe Time and Venue, Registration, Organizer, Event Description, Agenda or Program, Audience and Accessibility, Contact, and Footer. Preserve exact visible wording.

IV. Source insertion

Append the unmodified source item after the template and produce one final prompt that is sufficiently explicit for direct text-to-image generation.

Figure 6: Prompt-generation template for the Event Notice scenario.

Generation Instruction for LLMs (Financial Summary)

I. System role

You are a professional vision-language prompt writer. Parse the financial-summary image, recover its visual organization and all visible text and numerical values, and convert them into a faithful image-generation prompt.

II. Core rules

1. Output no greeting, explanation, or summary outside the prompt.
2. Preserve all legible text, numerical values, units, dates, percentages, currencies, legends, labels, and footnotes without omission.
3. Do not fabricate uncertain content or infer unsupported values.
4. For every chart, state the chart type, topic, visual encoding, axes, ticks, legend, and structured data.
5. Preserve the mapping between each number and its corresponding bar, line, sector, table cell, icon, or callout.

III. Required output template

Style Settings.: Describe the fixed palette, typography, and financial-report visual language.
Layout Structure.: Describe the page-level hierarchy and the position of every information region.
Region n ***- Position and Name.***: State the visual form and exact visible text for each title, KPI card, insight block, table, legend, or footer.
Chart Region n***.***: "Use [chart type] to present [topic]. Data is differentiated by [visual mapping]. The x-axis represents [meaning]. The y-axis represents [meaning and ticks]." Then list every series and value in structured form.
Consistency Constraints.: Require readable labels, correct totals, correct signs and units, and exact number-to-geometry correspondence.

IV. Output contract

Return one complete English generation prompt that exhausts all regions from top to bottom and left to right.

Figure 7: Prompt-generation template for the Financial Summary scenario.

Generation Instruction for LLMs (Organizational Chart)

I. System role

You are a senior prompt engineer for organizational charts and topology diagrams. Fully extract hierarchy, containers, connectors, and global page components from the raw prompt.

II. Output format

Return exactly two English paragraphs separated by one blank line.
Paragraph 1 - Global view. Describe the background, title and header treatment, divider and footer elements, global zones, topology type, total hierarchy depth, node count or range, flow direction, color system, node families, connector conventions, containers, and whether the layout is symmetric, asymmetric, matrix-like, radial, vertical, horizontal, or mixed.
Paragraph 2 - L1-to-leaf topology. Use explicit L1, L2, L3, and subsequent labels. Preserve every visible node name, title, department, position, and group. Specify node geometry, fill, border, portrait or icon placement, text alignment, connector attachment points, direct-reporting lines, collaboration links, container boundaries, and grouping logic.

III. Hard constraints

1. Preserve the complete superior-subordinate topology and all source text.
2. Preserve asymmetric structures; do not force visual symmetry.
3. Use only elements present or implied by the source. Add no unrelated decoration.
4. Connectors must be non-crossing where possible and anchored to geometric midpoints of node borders.
5. Distinguish direct reporting from horizontal collaboration through explicit line styles and directions.
6. Prohibit truncated text, overlapping nodes, ghost text, crossed connectors, and blank nodes.
7. Do not mention URLs, source pages, aspect ratios, paper sizes, pixel dimensions, or resolutions.
8. Output only the two-paragraph prompt; no JSON, Markdown, commentary, or explanation.

Figure 8: Prompt-generation template for the Organizational Chart scenario.

Generation Instruction for LLMs (Poster)

I. System role

You are an image-to-text-to-image prompt writer. Inspect the reference poster and return strict JSON containing an English title, difficulty label, and faithful English and Chinese generation prompts.

II. JSON schema

```
{ "title": "", "difficulty": "easy|medium|hard", "prompt_en": "", "prompt_zh": "" }
```

III. Extraction requirements

1. Preserve all clearly visible Chinese and English text in both prompt fields whenever possible; quote source-language text rather than silently altering it.
2. Describe the poster theme, title and slogan, principal visual subject, typography hierarchy, palette, background, icons, illustrations, and design style.
3. Decompose the composition into fixed zones: top identity area, main visual, selling-point or information cards, and bottom footer, price, call-to-action, or QR area.
4. Preserve product names, specifications, prices, dates, quantitative claims, labels, badges, and footer text exactly when legible.
5. Describe the count, placement, alignment, and visual form of cards, badges, logos, icons, and data blocks.
6. Do not invent uncertain text and do not mention platform watermarks or source links.

IV. Prompt organization

Overall Style.: Theme, palette, lighting, illustration or photography treatment, and typography.
Top Identity Area.: Brand, title, subtitle, slogan, and logo placement.
Main Visual.: Subject, orientation, scale, environment, and focal hierarchy.
Information Modules.: Exact module count, headings, body copy, values, icons, and relative positions.
Footer.: Product or event details, price, CTA, legal text, and generic QR placeholder if present.

V. Output contract

Return valid JSON only, with no Markdown or additional fields.

Figure 9: Prompt-generation template for the Poster scenario.

Generation Instruction for LLMs (PPT/Slide)

I. System role

You are a professional vision-language model. Parse the slide image, extract all visual features, text, and numbers, and return strict JSON with an English title, difficulty, and English image-generation prompt.

II. JSON schema

```
{ "title": "", "difficulty": "easy|medium|hard", "prompt_en": "" }
```

III. Core rules

1. Extract every visible text string and number without omission. Translate visible Chinese into English in the generation prompt.
2. Do not fabricate uncertain content and do not output explanations or Markdown.
3. Preserve the complete page hierarchy: title, subtitle, body blocks, logo, footer, background, decorative illustration, portrait, icons, borders, and page number.
4. For a table, provide its row and column counts, exact headers, and exact content of every cell.
5. For each chart, provide type, title, axes, ticks, legend, data values, visual encoding, grid, and appearance.

IV. Required prompt template

Style Settings.: Color palette and visual style.
Layout Structure.: Macro layout and reading order.
Region n ***- Position and Name.***: Visual form and exact text.
Chart Region n***.***: "Use [chart type] to present [topic]. Data is differentiated by [encoding]. The x-axis represents [meaning]. The y-axis represents [meaning and ticks]." List all data series and values.
Table Region n***.***: State dimensions, headers, and all cell values in reading order.

V. Output contract

Exhaust all slide regions and return valid JSON only.

Figure 10: Prompt-generation template for the PPT/Slide scenario.

Generation Instruction for LLMs (Statistical Dashboard)

I. Title and context

Application Name/Brand.: Exact application name and logo description.
Main Header.: Page title, search bar, account information, and utility controls.
Navigation.: Exact sidebar or top-navigation items in display order.

II. Layout structure

Describe the full dashboard grid, including navigation width, main content area, row and column organization, module spans, margins, alignment, and reading order.

III. Module template

Repeat the following block for every visible module:

- ***Module*** n ***- Name and Function.***
- **Type**: KPI card, progress indicator, chart, table, list, calendar, map, filter, or control group.
- **Key Information**: Exact headings, labels, values, units, changes, dates, and status text.
- **Visuals**: Icons, micro-trends, avatars, buttons, badges, tooltips, and decorative details.

IV. Chart-specific fields

For each chart, specify chart type, topic, series names, legend, axis meanings, ticks, units, color mapping, all visible values, trend characteristics, grid, annotations, and correspondence between data and geometry.

V. Style and layout

Specify one fixed dashboard style, background, primary and accent colors, card treatment, corner radius, shadows, borders, spacing, typography hierarchy, icon style, and interaction-state appearance. Preserve all sensitive or factual content exactly; add no unsupported modules.

VI. Output contract

Return one complete generation prompt ordered from global context to modules and from top-left to bottom-right.

Figure 11: Prompt-generation template for the Statistical Dashboard scenario.

A.2 Evaluation Details

A.2.1 Evaluation Dimension Registry

Each scenario has a set of Information Fidelity (IF) dimensions and Visual Quality (VQ) dimensions. IF dimensions verify content requirements, including text, numerical facts, entity attributes, relationships, and data–visual mappings. VQ dimensions assess presentation requirements, including layout hierarchy, readability, and thematic consistency.

Character Relationship

Dimension	Concise evaluation criterion
`character_relationship_illusion`	Checks whether the image contains content that is unrelated to or fabricated beyond the specified character-relationship diagram.
`character_relationship_l1_accuracy`	Checks the accuracy of character nodes, names, identity labels, and key text.
`character_relationship_l2_accuracy`	Checks the accuracy of inter-character relationships, connection directions, and relationship semantics.
`character_relationship_readability`	Checks whether nodes, connectors, and text are clearly distinguishable without overlap or occlusion.
`character_relationship_theme_relevance`	Checks whether the overall image matches the specified relationship-diagram theme and narrative setting.

Conference

Dimension	Concise evaluation criterion
`information_ThemeRelevance`	Checks whether the conference visuals and theme match the event context.
`information_keys_accuracy`	Checks the accuracy of key information such as dates, locations, people, and agenda items.
`information_statement_accuracy`	Checks whether specified statements, descriptions, and factual text are presented accurately.
`information_textHierarchy`	Checks whether the hierarchy among titles, agenda items, and explanatory text is clear.
`information_textlegibility`	Checks whether conference-related text is clear, complete, and readable.

Data Journalism

Dimension	Concise evaluation criterion
`data_journalism_l1_data_accuracy`	Checks the accuracy of data points, values, units, and data correspondences.
`data_journalism_l1_text_accuracy`	Checks the accuracy of titles, axis labels, categories, and explanatory text.
`data_journalism_l2_color_legend`	Checks whether colors and legends correctly correspond to categories or series.
`data_journalism_l2_geometry_trend`	Checks whether chart types, geometric encodings, and trends are represented correctly.
`data_journalism_readability`	Checks the readability of chart text, data labels, and the overall layout.
`data_journalism_text_hierarchy`	Checks whether the textual hierarchy among the title, main chart, annotations, and legend is clear.
`data_journalism_theme_relevance`	Checks whether the chart theme and content are relevant to the specified news narrative.

Event Notice

Dimension	Concise evaluation criterion
event_notice_keys_accuracy	Checks the accuracy of key information such as the event name, time, location, and target audience.
event_notice_layout_hierarchy	Checks whether the layout hierarchy among the notice title, body, and action information is clear.
event_notice_match_accuracy	Checks whether all notice content matches the specified event requirements.
event_notice_textlegibility	Checks whether notice text is complete, clear, and readable.
event_notice_theme_relevance	Checks whether the visual style matches the event theme, audience, and context.

Financial Summary

Dimension	Concise evaluation criterion
financial_summary_chart_axes_logic	Checks whether chart types, axes, scales, and data logic are correct.
financial_summary_context_title_accuracy	Checks the accuracy of titles, time ranges, and contextual descriptions.
financial_summary_financial_data_trend_accuracy	Checks the accuracy of financial values, units, comparisons, and trends.
financial_summary_insights_text_accuracy	Checks the accuracy of insights, conclusions, and explanatory text.
financial_summary_macro_layout_visual_elements	Checks whether the overall modular layout and core visual elements are reasonable and complete.
financial_summary_table_typography_readability	Checks the quality and readability of tables, typography, and typesetting.
financial_summary_zero_hallucination	Checks that no unrequested financial information or graphics are fabricated.

Knowledge List

Dimension	Concise evaluation criterion
knowledge_list_design_aesthetic	Checks the overall visual design, aesthetics, and consistency of the list.
knowledge_list_keys_accuracy	Checks the accuracy of list items, names, numbers, and key fields.
knowledge_list_layout_hierarchy	Checks whether grouping, order, hierarchy, and layout are clear.
knowledge_list_match_accuracy	Checks whether each item matches the specified knowledge or information requirement.
knowledge_list_no_hallucination	Checks that the list contains no irrelevant, misleading, or fabricated content.
knowledge_list_textlegibility	Checks whether list text is clear, complete, and easy to read.

Organizational Structure Chart

Dimension	Concise evaluation criterion
organizational_structure_chart_keys_accuracy	Checks the accuracy of organization names, positions, personnel, and key labels.
organizational_structure_chart_layer_accuracy	Checks the accuracy of hierarchy, reporting relationships, and organizational structure.
organizational_structure_chart_no_hallucination	Checks that no unrequested positions, departments, or relationships are introduced.
organizational_structure_chart_textlegibility	Checks whether text, nodes, and connectors in the organizational chart are clear and readable.

PPT

Dimension	Concise evaluation criterion
ppt_chart_table_accuracy	Checks the accuracy of charts, tables, and data representations on the slide.
ppt_no_hallucination	Checks that no unrequested content, facts, or visual elements are introduced.
ppt_supporting_accuracy	Checks the accuracy of supporting descriptions, body text, and supplementary information.
ppt_textlegibility	Checks whether slide text is clear, complete, and readable.
ppt_theme_relevance	Checks whether the overall theme, style, and visual content satisfy the requirements.
ppt_title_accuracy	Checks whether the slide title accurately reflects the central topic.

Promo Poster

Dimension	Concise evaluation criterion
promo_poster_keys_accuracy	Checks the accuracy of key information such as the poster name, selling points, time, and location.
promo_poster_layout_hierarchy	Checks whether the layout hierarchy among the title, main subject, selling points, and action information is clear.
promo_poster_no_hallucination	Checks that no unrequested products, people, text, or facts are introduced.
promo_poster_textlegibility	Checks whether poster text is clear, complete, and readable.
promo_poster_theme_relevance	Checks whether the visual style and content match the promotional theme and target audience.

Pros–Cons Matrix

Dimension	Concise evaluation criterion
pros_cons_matrix_hierarchy	Checks whether the positive and negative columns, groupings, and information hierarchy correspond clearly.
pros_cons_matrix_illusion	Checks that no irrelevant or fabricated comparison content is included.
pros_cons_matrix_readability	Checks whether the matrix text, columns, and visual organization are clear and readable.
pros_cons_matrix_text_accuracy	Checks the accuracy of pros-and-cons items, headings, and explanatory text.
pros_cons_matrix_theme_relevance	Checks whether the comparison dimensions and visual theme match the specified topic.

Sequential Operation Guide

Dimension	Concise evaluation criterion
sequential_operation_guide_extension_rationality	Checks whether supplementary steps, explanations, and extensions are reasonable and relevant.
sequential_operation_guide_hierarchy	Checks whether step grouping, hierarchy, and key information are clear.
sequential_operation_guide_l1_step_info	Checks the accuracy of each step's name, action, materials, and text.
sequential_operation_guide_l2_sequence	Checks the correctness of step order, connections, and process logic.
sequential_operation_guide_readability	Checks whether steps, arrows, and text are clear and readable.
sequential_operation_guide_theme_relevance	Checks whether the process content matches the specified task theme.

Statistical Dashboard

Dimension	Concise evaluation criterion
statistical_dashboard_chart_axes_trend_logic	Checks whether chart types, axes, scales, and trend logic are correct.
statistical_dashboard_context_text_accuracy	Checks the accuracy of titles, descriptions, filters, and contextual text.
statistical_dashboard_layout_structure_accuracy	Checks the accuracy of dashboard modules, regions, and hierarchical structure.
statistical_dashboard_numerical_accuracy	Checks the accuracy of values, units, aggregates, and data correspondences.
statistical_dashboard_typography_readability	Checks whether text, numbers, labels, and tables are clear and readable.
statistical_dashboard_ui_component_details	Checks the completeness and correctness of interface components such as filters, cards, and legends.
statistical_dashboard_visual_style_color_logic	Checks whether the visual style, color semantics, and series differentiation are appropriate.
statistical_dashboard_zero_hallucination	Checks that no unrequested data, modules, or factual content are introduced.

Temporal Trend

Dimension	Concise evaluation criterion
temporal_trend_data_accuracy	Checks the accuracy of time-series data points, values, and units.
temporal_trend_geometry	Checks the correctness of geometric representations such as trend lines, bars, and nodes.
temporal_trend_illusion	Checks that no information unrelated to or fabricated beyond the specified temporal trend is included.
temporal_trend_readability	Checks whether the time axis, annotations, trend graphics, and text are clear and readable.
temporal_trend_text_accuracy	Checks the accuracy of dates, titles, labels, and explanatory text.
temporal_trend_theme_relevance	Checks whether the trend topic and visual narrative satisfy the requirements.

Tour Poster

Dimension	Concise evaluation criterion
tour_poster_keys_accuracy	Checks the accuracy of destinations, routes, attractions, dates, and advisory information.
tour_poster_layout_hierarchy	Checks whether the layout hierarchy among the title, route, attractions, and travel tips is clear.
tour_poster_textlegibility	Checks whether route descriptions, explanations, and advisory text are clear and readable.
tour_poster_theme_relevance	Checks whether the travel theme, style, and destination presentation are consistent.

A.2.2 Error-Type Mapping and Definitions

Error-Type Definition

Table 6: Complete diagnostic error-type catalog used in InfoDiag-Bench. Definitions are taken from the final English tag schema; high-severity tags block delivery, whereas low-severity tags do not.

Tag ID	Definition	Severity
axis_error	The coordinate axis is presented, but its content type, values, units, scales, direction, or mapping to the graphic fails to meet reference requirements. This tag shall not be triggered if only the axis line, axis label is missing or the placement is incorrect, while content and mapping along that dimension remain clear.	high
axis_missing	Missing coordinate axis: The entire required horizontal or vertical axis content specified by reference is absent, or key categories, scales, units, etc. are missing to the extent that content along that dimension and its mapping to the graphic cannot be identified. Axis lines or axis-type titles such as "category axis" are not mandatory. As long as categories or scales for the specified dimension are recognizable and mapping is clear, absence of axis lines, descriptive terms or fixed placement shall not be judged as missing. If the axis is presented yet contains errors in type, value, direction or mapping, use axis_error instead.	high
body_text_error	Body text character error: Readable information in the image is inconsistent with reference information. When data or text labels are correct, readable and clearly associated with target objects, mere positional offset shall not trigger this tag, nor shall absence from the expected position alone be treated as missing. This positional tolerance does not apply if object attribution, scale or data mapping becomes wrong, or content becomes unreadable due to cropping or occlusion.	high
body_text_missing	Missing body-text content: The required corresponding content per reference is entirely absent or key parts are missing in the image. When data or text labels are correct, readable and clearly associated with target objects, mere positional offset shall not trigger this tag, nor shall absence from the expected position alone be treated as missing. This positional tolerance does not apply if object attribution, scale or data mapping becomes wrong, or content becomes unreadable due to cropping or occlusion.	high
card_structure_error	Card structure error: Layout, hierarchy, sequence, topology or boundary structure in the image does not comply with reference requirements.	high
card_text_error	Card text character error: Readable information in the image is inconsistent with reference information. When data or text labels are correct, readable and clearly associated with target objects, mere positional offset shall not trigger this tag, nor shall absence from the expected position alone be treated as missing. This positional tolerance does not apply if object attribution, scale or data mapping becomes wrong, or content becomes unreadable due to cropping or occlusion.	high
chart_missing	Missing chart: The required chart content specified by reference is entirely absent or key components are missing in the image.	high
chart_type_error	Chart type error shall be reported only when the chart explicitly required by reference is obviously different from the rendered chart and the core data representation method is changed, e.g. a bar chart is shown while a line chart is required. This tag is not triggered for different naming as long as graphical meaning and representation are similar, such as line chart versus line-with-markers chart. Differences limited to markers, line styles, color scheme, layout or minor graphical details also do not trigger this tag.	high

Continued on next page

Table 6 (continued)

Tag ID	Definition	Severity
chart_value_error	Chart value error: Readable information in the image is inconsistent with reference information. When data or text labels are correct, readable and clearly associated with target objects, mere positional offset shall not trigger this tag, nor shall absence from the expected position alone be treated as missing. This positional tolerance does not apply if object attribution, scale or data mapping becomes wrong, or content becomes unreadable due to cropping or occlusion.	high
connector_style_mismatch	Connector style mismatch: Attributes of connecting lines including color, thickness, line pattern are inconsistent with reference requirements.	low
content_overflow_boundary	Content overflow boundary: Mandatory text, data labels, charts or nodes extend beyond canvas edges or explicit container boundaries, resulting in cropped / missing content, or intrusion into other modules with ambiguous ownership. Normal line breaks, permitted multi-column layout, decorative bleed and edge-alignments without content impact shall not trigger this tag.	high
data_graphic_mismatch	Data-graphic mismatch is triggered only when a complete standard basic statistical chart (e.g. full pie chart, standard line chart or bar chart) is rendered, and readable numbers contradict their corresponding geometric encoding within the image itself. Examples: vertical positions for labelled values 25 and 35 contradict axis value direction, or a smaller value maps to a larger pie sector. Judgement depends solely on the image, not original prompt, key-information or other reference materials. Half-circular charts, creative infographics, non-standard or incomplete statistical charts are out-of-scope.	high
element_overlap_or_occlusion	Element overlap or occlusion: Two or more elements overlap and directly obstruct reading of essential text, causing incoherent semantics, misrepresented information or unintelligible overall logic. Mere visual overlap, crossed connectors, text-graph overlay or decorative occlusion does not trigger this tag as long as text remains readable and semantics, information and logic stay clear.	high
entity_relationship_mismatch	Entity-relationship mismatch: Object-label, color, connector or data-mapping relationships shown in the image are inconsistent with reference.	high
flow_order_error	Flow-order error: Layout, hierarchy, sequence, topology or boundary structure in the image fails to satisfy reference requirements.	high
wrong_hallucination	Irrelevant hallucinated information: The image introduces substantial information not required by reference, unrelated to topic or module core content, which distracts understanding or misleads interpretation. Reasonable extensions that stay topic-relevant, aid comprehension and do not alter factual reference content shall not trigger this tag.	high
identity_label_error	Identity-label error: Readable information in the image is inconsistent with reference information. When data or text labels are correct, readable and clearly associated with target objects, mere positional offset shall not trigger this tag, nor shall absence from the expected position alone be treated as missing. This positional tolerance does not apply if object attribution, scale or data mapping becomes wrong, or content becomes unreadable due to cropping or occlusion.	high
identity_label_missing	Missing identity-label: Required identity-label content per reference is entirely absent or key parts are missing in the image. When data or text labels are correct, readable and clearly associated with target objects, mere positional offset shall not trigger this tag, nor shall absence from the expected position alone be treated as missing. This positional tolerance does not apply if object attribution, scale or data mapping becomes wrong, or content becomes unreadable due to cropping or occlusion.	high
image_or_icon_error	Image or icon error: Visual elements, graphic shape or type in the image do not match reference requirements.	high

Continued on next page

Table 6 (continued)

Tag ID	Definition	Severity
icon_missing	Missing icon / image: Required image or icon content specified by reference is entirely absent or key components are missing in the image.	low
large_scale_graphic_error	Large-scale graphic error: Within the observable viewport, a complete core infographic module contains holistic errors, or at least 50 % of independent graphic units share the same category of errors in shape, type, structure or visual encoding, distorting core meaning or relationships. Examples: wrong chart type for an entire chart, reversed main org-chart structure, or systematic distortion of nearly half of data graphics. Isolated minor icon defects, purely aesthetic preferences, text-only errors and harmless decorations do not trigger this tag; do not auto-migrate all rendering-error cases into this tag.	high
large_scale_text_error	Large-scale text error: Text is legible, yet an entire core text module is replaced by incorrect content, or at least 50 % of independent text units contain material errors versus reference that impair primary message delivery. Independent units refer to titles, full sentences, labels or data entries; multiple miswritten characters inside one word count as a single unit. A single local typo, permitted synonymous rephrasing, pure absence or illegible garbled text shall not trigger this tag. When data or text labels are correct, readable and clearly associated with target objects, mere positional offset shall not trigger this tag, nor shall absence from the expected position alone be treated as missing. This positional tolerance does not apply if object attribution, scale or data mapping becomes wrong, or content becomes unreadable due to cropping or occlusion.	high
large_scale_text_garbled_or_blurry	Large-scale garbled or blurry text: After reasonable magnification at native image resolution, an entire core text module remains unreliably readable, or at least 50 % of independent text units are unrecognizable due to fake characters, merged glyphs, broken shapes or blurriness, making essential information inaccessible. Minor single-character flaws, tiny thumbnails while original image stays readable, legible content errors and normal font-style variations do not trigger this tag. Unreadability caused by explicit overlay objects or canvas cropping shall be assigned to occlusion or overflow to avoid duplicate reporting of the same defect.	high
legend_error	Legend error: Observable defects exist within the legend in the image. When data or text labels are correct, readable and clearly associated with target objects, mere positional offset shall not trigger this tag, nor shall absence from the expected position alone be treated as missing. This positional tolerance does not apply if object attribution, scale or data mapping becomes wrong, or content becomes unreadable due to cropping or occlusion.	high
legend_missing	Missing legend: Required legend content specified by reference is entirely absent or key parts are missing in the image. When data or text labels are correct, readable and clearly associated with target objects, mere positional offset shall not trigger this tag, nor shall absence from the expected position alone be treated as missing. This positional tolerance does not apply if object attribution, scale or data mapping becomes wrong, or content becomes unreadable due to cropping or occlusion.	high
list_content_missing	Missing list content: Required list content per reference is entirely absent or key entries are missing in the image. When data or text labels are correct, readable and clearly associated with target objects, mere positional offset shall not trigger this tag, nor shall absence from the expected position alone be treated as missing. This positional tolerance does not apply if object attribution, scale or data mapping becomes wrong, or content becomes unreadable due to cropping or occlusion.	high

Continued on next page

Table 6 (continued)

Tag ID	Definition	Severity
list_order_error	List-order error: Layout, hierarchy, sequence, topology or boundary structure of the list in the image does not comply with reference requirements.	high
list_text_error	List-item text error: Readable information within list items in the image is inconsistent with reference information. When data or text labels are correct, readable and clearly associated with target objects, mere positional offset shall not trigger this tag, nor shall absence from the expected position alone be treated as missing. This positional tolerance does not apply if object attribution, scale or data mapping becomes wrong, or content becomes unreadable due to cropping or occlusion.	high
numbering_error	Numbering error: Layout, hierarchy, sequence, topology or boundary structure reflected by numbering in the image fails reference requirements.	high
numeric_unit_error	Numeric-unit error: Readable unit information in the image is inconsistent with reference information. When data or text labels are correct, readable and clearly associated with target objects, mere positional offset shall not trigger this tag, nor shall absence from the expected position alone be treated as missing. This positional tolerance does not apply if object attribution, scale or data mapping becomes wrong, or content becomes unreadable due to cropping or occlusion.	high
numeric_value_error	Numeric-value error: The core numeric value of a data item in the image disagrees with reference. Correct core value plus correct unit is sufficient. Adding, removing or replacing regular whitespace between number and unit, as well as formatting variations that do not alter numeric meaning, shall not trigger this tag. Use numeric_unit_error for wrong units. Mere positional shift does not trigger this tag.	high
numeric_value_missing	Missing numeric value: Required numeric-value content specified by reference is entirely absent or key parts are missing in the image. When data or text labels are correct, readable and clearly associated with target objects, mere positional offset shall not trigger this tag, nor shall absence from the expected position alone be treated as missing. This positional tolerance does not apply if object attribution, scale or data mapping becomes wrong, or content becomes unreadable due to cropping or occlusion.	high
relationship_edge_error	Relationship-edge error: Object-label, color, connector or data-mapping relationships represented by edges in the image are inconsistent with reference.	high
relationship_edge_missing	Missing relationship edge: A required relationship connector specified by reference is entirely absent or its key connecting segment is missing in the image.	high
structure_relation_error	Structure-relationship error: Structural relationships such as hierarchical ownership, relation direction and reading path shown in the infographic conflict with reference requirements.	high
table_missing	Missing table: The required table specified by reference is entirely absent or key components are missing in the image.	high
table_structure_error	Table-structure error: Layout, hierarchy, sequence, topology or boundary structure of the table in the image does not meet reference requirements.	high
table_text_error	Table-cell text error: Readable content inside table cells in the image is inconsistent with reference information. When data or text labels are correct, readable and clearly associated with target objects, mere positional offset shall not trigger this tag, nor shall absence from the expected position alone be treated as missing. This positional tolerance does not apply if object attribution, scale or data mapping becomes wrong, or content becomes unreadable due to cropping or occlusion.	high

Continued on next page

Table 6 (continued)

Tag ID	Definition	Severity
table_typography_error	Table typography error: Formatting details including header style, row-column alignment, special-column background / text style and table borders deviate from reference specifications.	high
text_hierarchy_confusion	Text-hierarchy confusion: Visual priority among chart elements such as title, description, chart plot area, axes and labels is disordered, yielding an unclear reading flow.	high
text_low_contrast	Low-contrast text: Poster text cannot be read clearly due to insufficient text-background contrast, overly small font size, line-art occlusion or text truncation.	high
title_text_error	Title core-content error: The title misrepresents core topic, entities or key messages. Verbatim match is not required; rephrasing, abbreviation or omitted descriptive words that preserve core meaning shall not trigger this tag, e.g. ”Finances Chart” shortened to ”Finances”.	high
ui_component_detail_deviation	UI-component detail deviation: UI components including card shadow, corner radius, icon style, avatar, button and input box deviate from reference design or placement; or components are missing or rendered crudely.	low
unreasonable_extension_content	Unreasonable extended content: The image adds supplementary content not explicitly required by reference, which is irrelevant to operational goals, misleads operations, or breaks original steps and logical relationships. Plain text garbling, visual decorations, layout defects and direct factual alterations of reference content are out-of-scope for this tag.	high
visual_style_mismatch	Visual-style mismatch: Poster color palette, icons, visual elements and overall mood diverge from the required poster style, or the output lacks properties of a travel-guide poster.	high

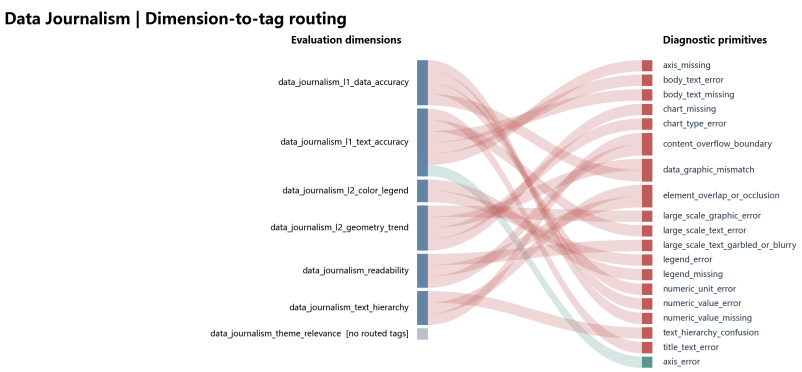


Figure 12: Data Journalism

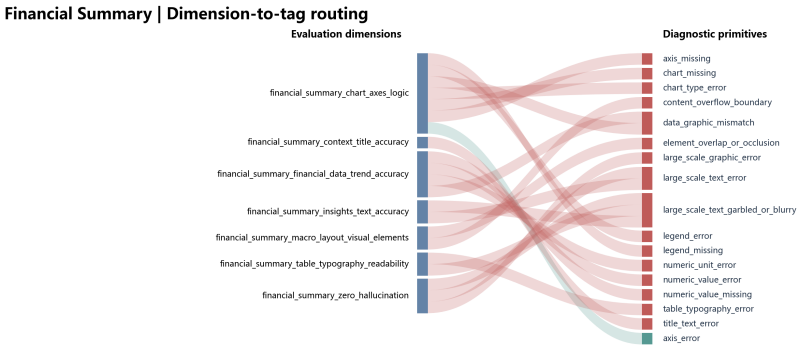


Figure 13: Financial Summary

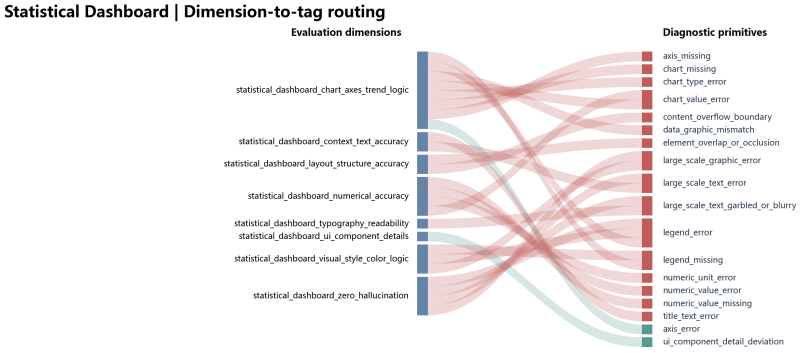


Figure 14: Statistical Dashboard

Error-Type Mapping with Dimensions

A.3 Additional Results

A.3.1 Comparison of MLLM Evaluators

We employ ”GPT-5.5”(OpenAI, 2026) and ”doubao-seed-2-0-lite-260428”(ByteDance, 2026) as evaluators for our benchmark. The corresponding results are presented in Table 7.

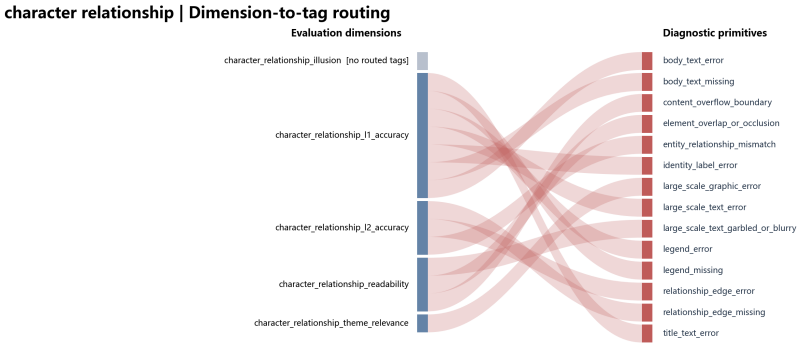


Figure 15: Character Relationship

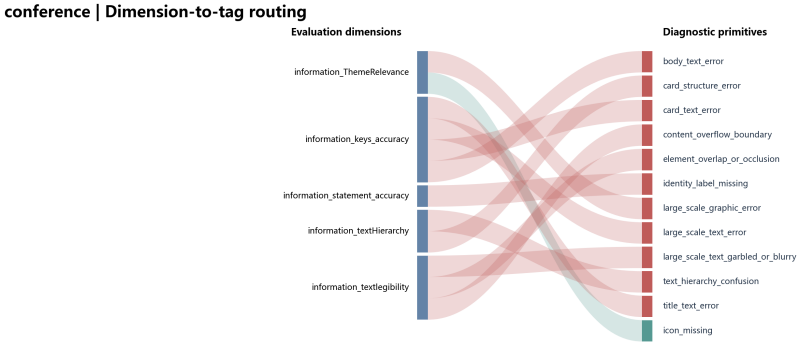


Figure 16: Conference

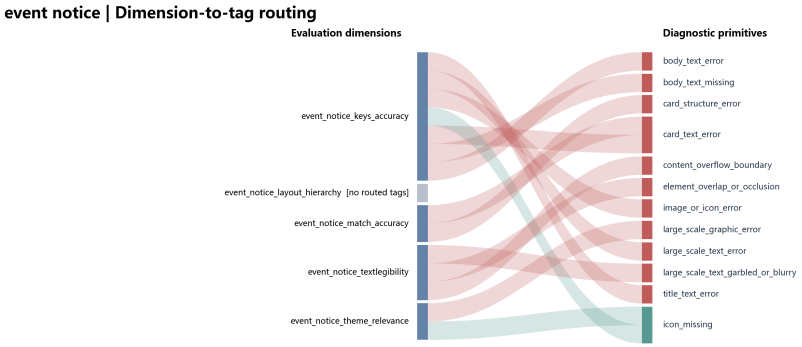


Figure 17: Event Notice

Table 7: Comparison of Evaluator Consistency across Different Result Sets.

Result Set / Evaluator	**GPT**	**Seed-5.0**	**SenseNova-U1-V3**
GPT-5.5 Sampled Eval.	70.33% (64/91)	29.76% (25/84)	17.33% (13/75)
Doubao Sampled Eval.	84.00% (84/100)	41.00% (41/100)	23.00% (23/100)
0920final Historical Results	77.78% (77/99)	35.11% (33/94)	23.00% (23/100)

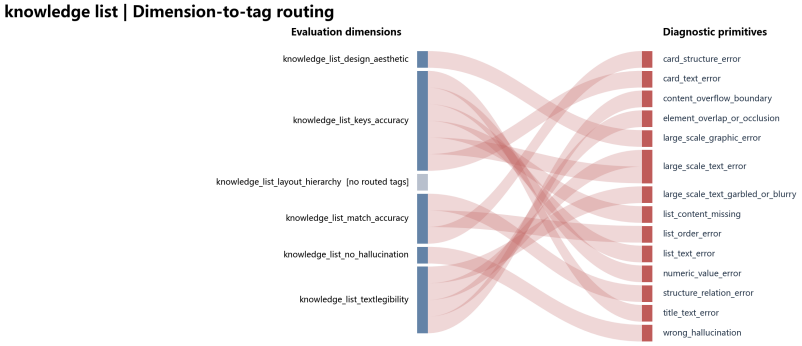


Figure 18: Knowledge List

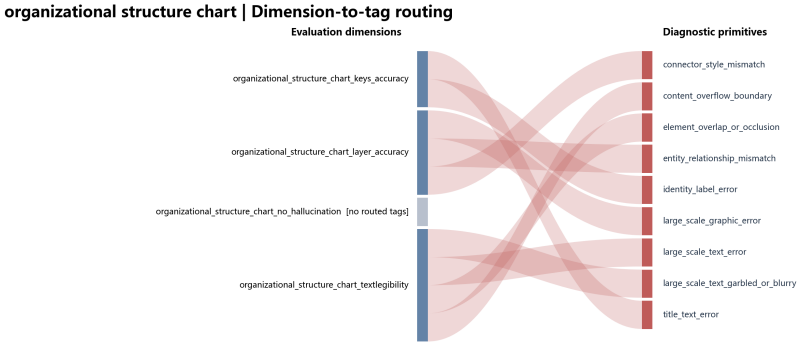


Figure 19: Organizational Structure Chart

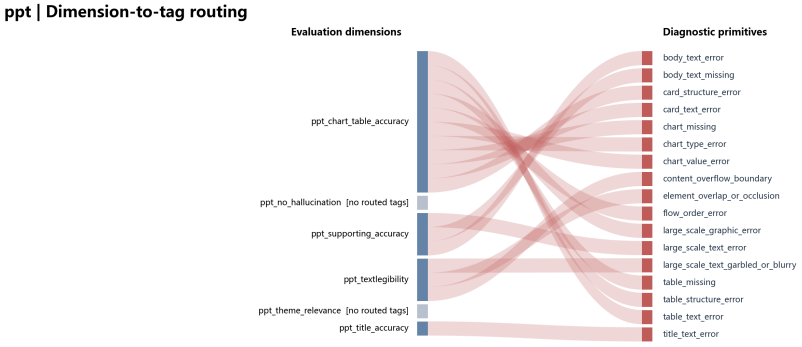


Figure 20: PPT

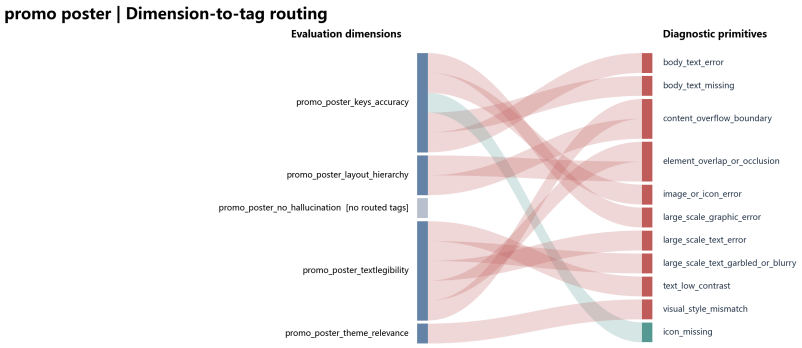


Figure 21: Promo Poster

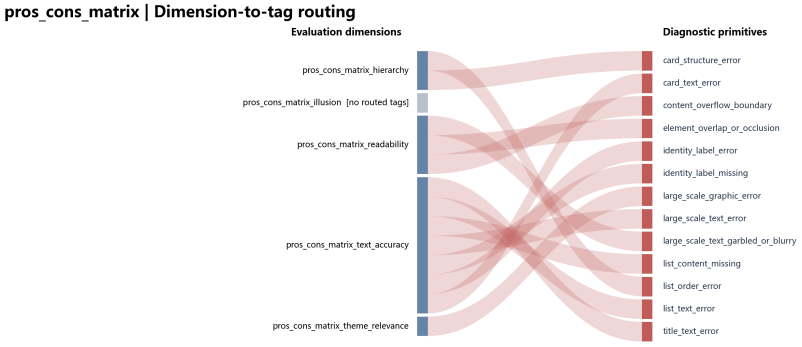


Figure 22: Pros Cons Matrix

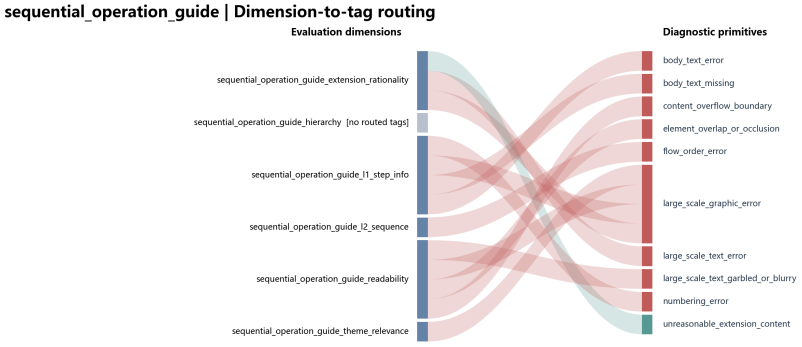


Figure 23: Sequential Operation Guide

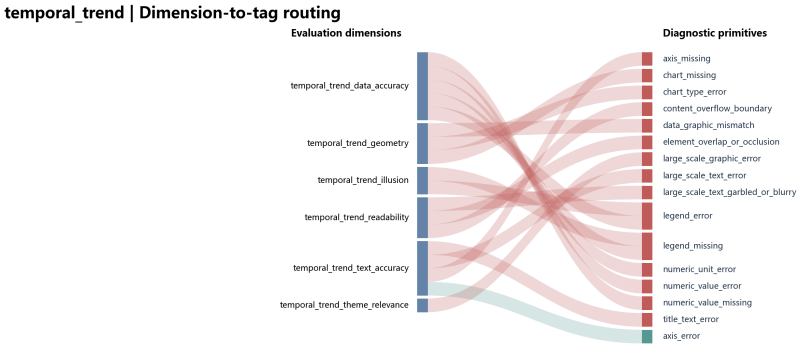


Figure 24: Temporal Trend

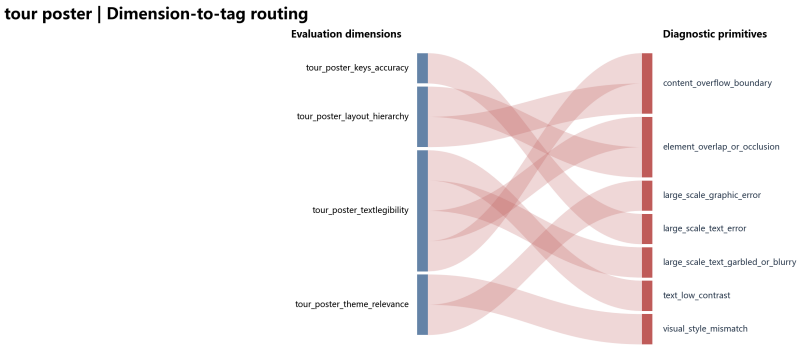


Figure 25: Tour Poster