TEXTCANVASBENCH: A COMPREHENSIVE BENCHMARK FOR COMPLEX AND LONG-TEXT IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent open- and closed-source text-to-image (T2I) models can generate text-rich images, including scientific figures and presentation slides, yet benchmark evaluation has not kept pace with this capability. Existing text-rendering benchmarks typically focus on short text, a small number of text regions, and image-level or OCR-based aggregate metrics, limiting their ability to assess long-form content, multi-region layouts, and block-level stylistic and semantic relations in practical T2I scenarios. To address these gaps, we introduce TEXTCANVASBENCH, a benchmark of 1,600 specification-guided samples spanning 8 domains and 38 subcategories of real-world text-rich visuals, from commercial posters and academic slides to digital interface panels. The required text ranges from short labels to more than 1,000 characters, and the annotations identify text blocks and their roles. We evaluate generation through four complementary dimensions, moving from the whole image to individual text blocks and regions. At the macro image level, we assess glyph quality, the completeness and accuracy of the required text, and whether the image conveys the prompt's overall meaning. At the micro block and region level, we assess position, orientation, arrangement and reading order, font, size, and color, as well as selected facts and associations tied to individual blocks. In our experiments, we evaluate nine state-of-the-art open- and closed-source T2I models (e.g., Boogu-Image and GPT-Image-2). The results show that TEXTCANVASBENCH covers a wider range of practical text-rich scenarios and reveals differences in long-text rendering capabilities that existing benchmark scores can obscure. Its detailed scores distinguish models with similar scores on other benchmarks, while its rankings agree more closely with human judgments than those of most benchmarks in our comparison.

1 INTRODUCTION

Recently, text-to-image (T2I) models (Qwen Team, 2025; OpenAI, 2026) can now generate text-rich visuals, from scientific figures and presentation slides to posters and digital interfaces. To be useful, these images must render extended passages, numbers, and units accurately, organize multiple text elements, and link them to the right visual content. Existing benchmarks assess text rendering, layout, and prompt following (Wang et al., 2025; Mao et al., 2026; Li et al., 2026b). Yet practical text-rich images may require hundreds or thousands of characters across many text blocks, whereas the English portion of LongText-Bench predominantly requires a few dozen words (Geng et al., 2025). Aggregate scores can also conceal differences in text completeness, presentation, and semantic fidelity between models with similar benchmark scores, as illustrated in Figure 1 and quantified in Table 1. This motivates an evaluation that tests content, organization, and meaning together while showing where each model fails.

As shown in Figure 1, all five models produce visually coherent travel illustrations from the short prompt. When the same theme becomes a detailed travel guide with four destination sections, paragraphs, information rows, and a structured footer, their outputs diverge: the open-source models exhibit more omissions, corrupted text, and layout problems, whereas the closed-source models preserve more of the requested content and organization. This contrast exposes an evaluation gap: short prompts can make models appear similarly capable, highlighting the need for a new benchmark that

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.

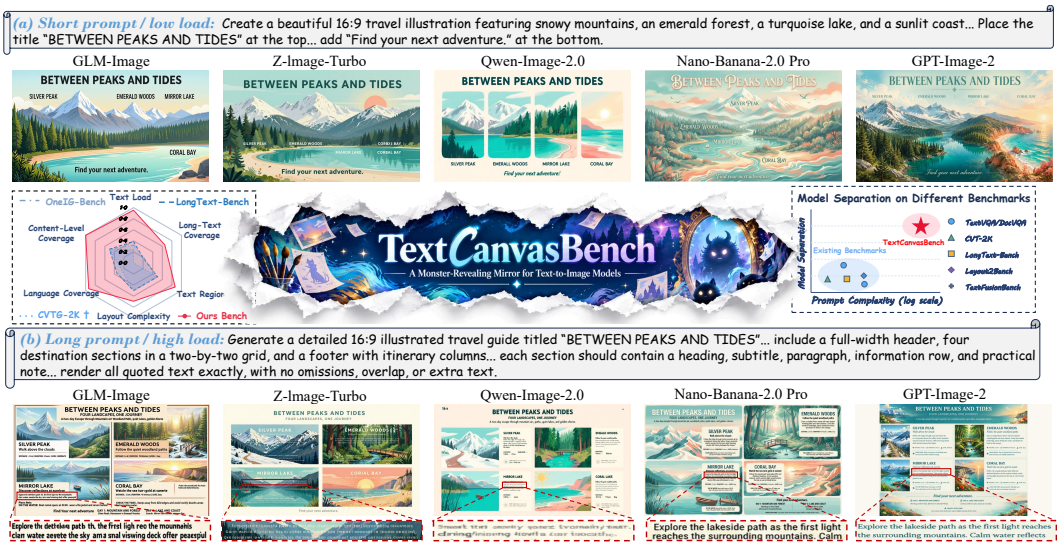

Figure 1: **Short prompts can hide substantial model differences.** (a) With a short prompt and few required text elements, open- and closed-source models produce similarly coherent travel illustrations. (b) When the same task specifies more text, multiple content sections, and a detailed layout, omissions, corruptions, and organization failures become more visible, particularly in the open-source outputs. The middle panels compare benchmark coverage across six aspects (left) and model separation on different benchmarks (right).

covers a broader mix of text lengths, long-form, text regions, layout complexity, languages, and content requirements, and more clearly differentiates the evaluated models than existing benchmarks.

We therefore introduce **TEXTCANVASBENCH**, a benchmark of 1,600 specification-guided samples spanning 8 domains and 38 subcategories of real-world text-rich visuals, from commercial posters and academic slides to digital interface panels. Each sample includes a self-contained generation prompt, the complete required text, and reviewed annotations identifying individual text blocks and their roles. The required text ranges from short labels to over 1,000 characters. Models generate images using only the prompt and required text while reference images and annotations are used to build and evaluate the benchmark.

Additionally, we evaluate generation through four complementary dimensions, moving from the whole image to individual text blocks and regions. At the macro image level, **D1** measures glyph structure quality, while **D2** measures whether the complete required text is rendered accurately. At the micro block level, **D3** evaluates prompt-specified position, block orientation, arrangement and reading order, font size, style and color. **D4** evaluates whether the rendered text preserves the required facts, intent, text–image associations, scene logic, and integration with specified text-bearing objects. D3 and D4 use fixed specifications derived before model comparison and observations collected from each generated image. This two-level evaluation distinguishes errors in the rendered text, errors in its layout and style, and errors in the meaning it conveys in context.

We evaluate nine open- and closed-source T2I models, including Boogu-Image and GPT-Image-2. TEXTCANVASBENCH covers a wider range of practical text-rich scenarios and distinguishes models with similar scores on other benchmarks. The resulting model rankings agree more closely with human judgments than those of most benchmarks in our comparison, while the four diagnostic dimensions identify errors in glyphs, text content, presentation, and meaning.

Our main contributions are:

- **TEXTCANVASBENCH.** We introduce 1,600 specification-guided samples across 8 domains and 38 subcategories. Each sample includes a generation prompt, required text, reviewed annotations of text blocks and their roles, and fixed D3/D4 specifications for presentation and semantic evaluation.

- **Macro and micro-level Evaluation.** We define four complementary dimensions: D1 for glyph quality, D2 for complete-text accuracy, D3 for layout and style, and D4 for semantic alignment from the whole image to text blocks and regions.
- **Benchmark Validation** Across nine generators, our benchmark distinguishes models with similar scores on existing benchmarks and yields rankings that agree more closely with human judgments than most benchmarks in our comparison.

2 TEXTCANVASBENCH CONSTRUCTION

This section presents the three-stage construction pipeline of TEXTCANVASBENCH. First, we curate the reference images and characterize the resulting task distribution (Section 2.1). Next, we convert each reference image into paired Chinese and English generation prompts that specify the required text, visual content, and organization (Section 2.2). Finally, we construct frozen D3 and D4 specifications from the finalized prompts, complete required text, and reviewed annotations. D3 formalizes presentation requirements, whereas D4 defines semantic conditions and the visual evidence needed to verify them. (Section 2.3)

2.1 DATA COLLECTION AND STATISTICS

We assembled an initial pool of approximately 10,000 candidate reference images through image search over publicly accessible websites. We retained images in which the text is legible and conveys substantive information, while preserving variation in both content and visual presentation. This curation process yielded 1,600 reference images. To represent distinct forms of textual organization in practical images, we grouped the curated set by visual medium and application context into eight domains. *Infographics* combine explanatory passages, numerical labels, legends, and charts across aligned regions. *Posters & Ads* use strong typographic hierarchies and free-form compositions to separate headlines from supporting copy. *Art & Illustration* embeds stylized text among visual elements, often with irregular placement, scale, or orientation. *Document Design* organizes paragraphs, columns, and tables under explicit reading orders. *Real-world Photos* place text on physical surfaces under variations in perspective, illumination, occlusion, and scale. *Spatial Signage* distributes short functional text across signs whose meaning depends on its spatial context. *Digital Interfaces* pack small labels, values, and controls into repeated grids or cards with precise local associations. *Academic & Slides* interleave technical text, equations, figures, captions, and annotations in dense multi-region layouts. Within each domain, content topic and use case further define 38 subcategories. Individual subcategories contain 24–84 samples, extending each visual medium across multiple concrete applications rather than a single content type. Appendix A.1 provides the complete taxonomy and sample counts. The resulting collection contains 42,853 annotated text blocks across eight domains and exhibits broad variation in required-text length, layout complexity, content specialization, and language; detailed dataset statistics are reported in Figure 2 and Appendix A.

2.2 GENERATION PROMPTS

To translate real-world text-rich images into self-contained textual inputs, we construct paired Chinese and English generation prompts from reference images (Figure 3(c)). Following a unified image-understanding instruction, Gemini 2.5 Pro extracts the required text and produces paired Chinese and English descriptions that capture the visual content and composition while specifying text placement, layout, font, size, and color. Before inclusion, human reviewers compare both descriptions and the extracted required text with the reference image, correcting inaccurate descriptions or missing text, as illustrated in Figure 3(b). The reviewed Chinese and English prompts are provided separately to generators without the reference image. Appendix B.1 presents the construction instruction and a complete example.

2.3 EVALUATION SPECIFICATION CONSTRUCTION

We first use OCR- and VLM-assisted annotation to locate text blocks in each reference image and extract their geometry and presentation attributes. Human reviewers correct the transcriptions,

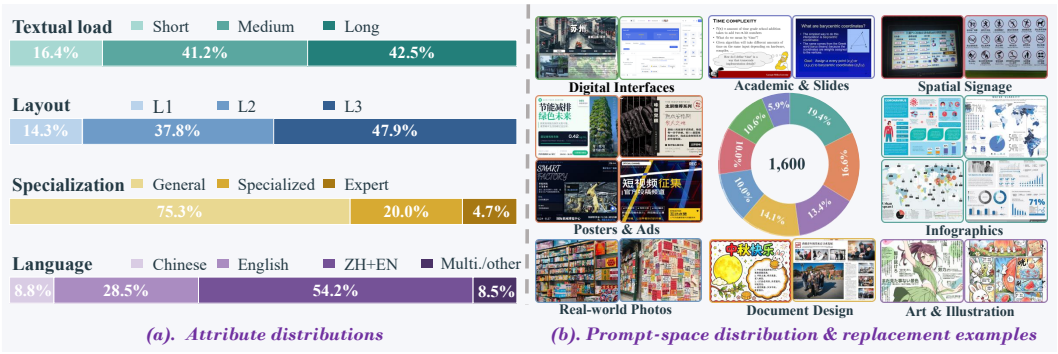

Figure 2: **Composition of TextCanvasBench.** (a) Distributions of required-text length, layout level, content specialization, and language. (b) Domain proportions and representative examples across the eight application domains.

bounding boxes, and grouping: distinct visible occurrences remain separate, related text blocks are grouped into layout regions, and line wrapping within a block does not create additional instances. We then combine these reviewed annotations with the finalized prompt and complete required text to construct two model-independent specifications. Before freezing them, human reviewers verify both specifications against the finalized prompt, complete required text, and reviewed reference annotations, correcting missing, unsupported, or incorrectly associated requirements. All models are therefore assessed against the same sample-specific specifications.

For the D3 specification, we extract every explicit presentation requirement from the finalized prompt and associate it with the relevant annotated text occurrence or occurrences. These requirements are organized into six attribute families: placement, spatial arrangement and reading order, size hierarchy, orientation, foreground color, and font style. Each criterion records its source phrase, affected text, expected value or relation, and the observation to be extracted from a generated image. Only presentation properties explicitly requested by the prompt enter the specification. This process converts the free-form prompt into structured, sample-specific presentation criteria; Appendix B.2 provides the complete construction and review procedure. **The D4 specification** defines which meanings a generated image should preserve and what visible evidence is needed to verify them. It draws on two sources. First, a structured VLM pipeline partitions the complete required text into coherent passages and extracts atomic facts while preserving their entities, attributes, values, units, conditions, negations, and associations. Semantic review removes unsupported inferences and retains at most five facts that best represent the sample's central textual content; these facts define the core textual-semantics aspect. Second, the finalized prompt supplies requirements for four complementary aspects: overall communicative intent, text–image consistency, scene logic, and text–carrier integration. For every applicable aspect, the specification records its source passage or prompt phrase, the expected semantic condition, and the visible text, entities, or relations to inspect in a generated image. Text–carrier integration is included only when the prompt explicitly specifies a physical text-bearing surface. The resulting five-part specification is reviewed and frozen before model evaluation, while the corresponding evidence is extracted later from each generated image. The construction of the D3 and D4 specifications is illustrated in the Reference Specifications module of Figure 3(c). Appendix B.3 describes the full construction and validation procedure.

3 TextCanvasBench Evaluation

Text accuracy alone does not capture failures in visual organization or semantic association. Therefore, we evaluate generation at two complementary levels using four diagnostic dimensions as shown in Figure 3(d). We use the previously frozen D3 and D4 specifications to determine what must be verified for each sample (Section 2.3). During evaluation, D1 and D2 directly analyze the generated image and its text transcription. For D3 and D4, OCR and VLM locate the relevant visible text in the generated image and extract the presentation and semantic observations required by the corresponding specifications. These observations are then compared with the specified requirements to obtain the corresponding scores. Appendix C provides the detailed evaluation rules.

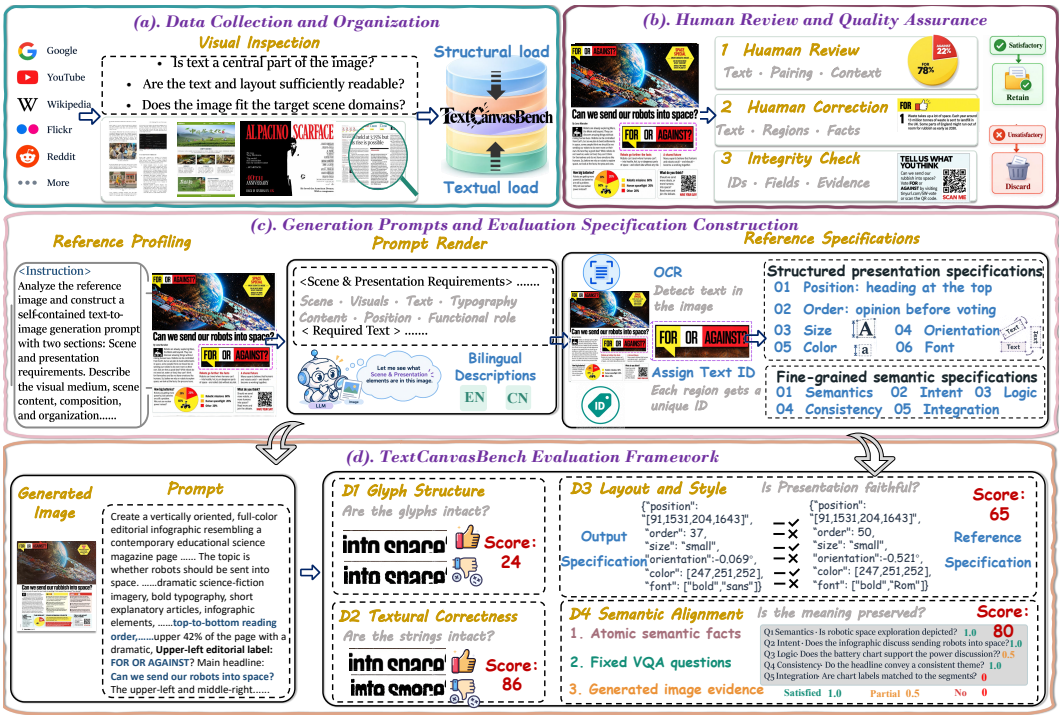

Figure 3: **Overview of data collection, benchmark construction, and evaluation.** Collected reference images are screened and annotated to produce generation inputs and evaluation specifications. Model outputs are then assessed along four diagnostic dimensions using evidence from the generated images (Section 3).

3.1 Image-Level Text Evaluation: D1 and D2

For each sample i, TextPecker (Zhu et al., 2026) receives the full generated image I_i and transcribes its visible text with glyph-error markers. The complete required text T_i from the final generation prompt is used as the content reference, retaining repeated occurrences. Let $\widehat{T}_i$ denote the parsed transcription. We use TextPecker's glyph-quality function Q and matching-based normalized edit distance GNED to obtain

$$D1_i = 100\, Q(\widehat{T}_i), \qquad D2_i = 100\big[1 - \mathrm{GNED}(\widehat{T}_i, T_i)\big]. \tag{1}$$

D1 measures glyph integrity through the proportion of marked structural errors. D2 measures textual correctness through edit costs and unmatched-text penalties, accounting for substitutions, missing text, and recognized extra or repeated text across the image. Thus well-formed but incorrect text can receive high D1 and low D2. Both scores are computed directly at macro image level; micro block is used for presentation and association analysis in D3/D4.

3.2 Layout and Style Evaluation: D3

D3 asks whether the generated text follows the layout and style specified in the prompt. Its six attribute families are region, spatial/reading order, size hierarchy, orientation, foreground color, and font style, with base weights $w = (0.25, 0.20, 0.15, 0.10, 0.15, 0.15)$ (Appendix B.2). Frozen checks target individual text instances, specified groups, or functional regions reconstructed from generated members. Deterministic rules compare observed geometry and attributes with these checks, awarding partial credit within fixed tolerances: position and order use normalized geometry; hierarchy compares compatible text scales across required subject–object pairs; orientation separates writing direction from glyph rotation; color uses HSV ranges; and style distinguishes present, absent, and unknown attributes. Unspecified appearance is unpenalized, and unobserved attributes remain unknown (Appendix C.1).

Generated-image evidence. OCR and VLM grounding align visible generated text with the frozen `text_ids`. Text-level observations provide polygons, orientation, foreground color, and style evidence; observed member instances are grouped by `region_id` to reconstruct functional regions. Each observation is linked to the check it supports. Confirmed omissions receive no credit, whereas unresolved attributes remain unknown rather than being treated as violations.

For supported checks $\mathcal{C}_{ik}$ in active component k, let $s_{ic} \in [0, 1]$ denote quality on available evidence and $e_{ic} \in [0, 1]$ the evidence fraction. The diagnostic score is

$$L_i^{(3)} = 100 \sum_{k \in \mathcal{A}_i} \widetilde{w}_{ik} \frac{\sum_{c \in \mathcal{C}_{ik}} e_{ic} s_{ic}}{|\mathcal{C}_{ik}|}, \qquad \widetilde{w}_{ik} = \frac{w_k}{\sum_{\ell \in \mathcal{A}_i} w_\ell}. \tag{2}$$

3.3 Semantic Evaluation: D4

D4 spans both evaluation levels: it tests whether the generated image conveys the intended meaning and whether individual text blocks preserve the required facts and relationships to the scene. It uses five components: core textual semantics, communicative intent, text–image consistency, semantic logic, and physical-carrier integration. Its frozen specification decomposes text-related meaning into atomic semantic facts that preserve entities, attributes, values, units, negations, conditions, and associations, together with additional prompt-supported semantic criteria (Appendix B.3). Physical-carrier integration applies only to prompt-specified physical text-bearing surfaces; font or position changes constitute semantic errors only when they alter required meanings or associations.

Generated-image evidence. The five fixed questions cite a shared output-evidence library extracted from the generated image: atomic facts require recovered text, and associations additionally require links between participating entities. Invalid or missing evidence receives one targeted repair attempt. Valid judgments score 1, 0.5, or 0 for satisfied, partially satisfied, or not satisfied. Core semantics averages applicable atomic facts equally; any unresolved fact makes this component unknown. N/A requires explicit prompt review confirming no supported atomic fact, with rationale and prompt provenance recorded; empty extraction or unresolved prompt interpretation does not suffice. Other supported components remain active.

With base weights $v = (0.45, 0.10, 0.20, 0.20, 0.05)$ and applicable components $\mathcal{B}_i$, the formal score is

$$D4_i = 100 \sum_{k \in \mathcal{B}_i} \widetilde{v}_{ik} r_{ik}, \qquad \widetilde{v}_{ik} = \frac{v_k}{\sum_{\ell \in \mathcal{B}_i} v_\ell}, \tag{3}$$

requiring a verified, frozen specification and valid judgments r_{ik} for all applicable components. Components that are not applicable to a sample are excluded from D4 aggregation. D4 is computed over the remaining applicable components(Appendix C.2).

3.4 Overall Score and Reporting

All dimensions range from 0 to 100. Samples with four valid, applicable scores receive the equally weighted summary

$$S_i = 0.25D1_i + 0.25D2_i + 0.25D3_i + 0.25D4_i. \tag{4}$$

We report the four dimension scores alongside S_i so that the summary does not hide different failure modes. Missing or N/A dimensions are neither zero-filled nor reweighted to produce S_i.

4 Experiments

Our experiments test whether TextCanvasBench separates current generators, how performance changes as the required text and independently organized blocks increase, and which failures text accuracy alone misses. We first compare model scores and rankings across benchmarks, then stratify results by character and text-block counts, inspect presentation and semantic errors, and isolate the effect of benchmark inputs under a shared evaluator. Nine generators (six open-weight, three proprietary) receive the same 1,600 prompts without reference images or evaluation specifications. We retain the first valid output within a fixed retry budget and distinguish generation failures from unresolved evaluation evidence. Each analysis states its scoring protocol and uses records available for every model being compared.

Table 1: **Cross-benchmark comparison on a 0–1 scale.** We compare published scores from CVTG (Tai et al., 2025), LongText-Bench (Geng et al., 2025), and OneIG-Bench (Chang et al., 2026) with the English- and Chinese-prompt TextCanvasBench scores. Scores originally reported on a 0–100 scale are divided by 100. Variance is the population variance over the seven models shared by all columns; larger values indicate greater cross-model dispersion. Spearman ρ compares each benchmark ranking with the human ranking over its available models.

Generator	CVTG	LongText		OneIG	Ours	
	Word	EN	ZH	Ovr. EN	S_{EN}	S_{ZH}
Open-weight models						
FLUX.1 [dev] (Black Forest Labs, 2024)	0.4965	0.607	0.005	0.434	0.4460	0.3049
SD3.5 Large (Stability AI, 2024)	0.6548	0.4460	0.0079	0.462	0.3243	0.1372
Qwen-Image (Wu et al., 2025)	0.8288	0.943	0.946	0.539	0.7355	0.7619
Z-Image-Turbo (Cai et al., 2025)	0.8585	0.917	0.926	0.528	0.7653	0.7789
GLM-Image (Z.ai Team, 2026)	0.9116	0.952	0.979	0.528	0.7582	0.7594
Boogu-Image-0.1-Turbo (Chen et al., 2026)	0.819	0.944	0.977	—	0.7891	0.8066
Proprietary models						
GPT-Image-2 (OpenAI, 2026)	0.9003	0.964	0.983	0.516	**0.8884**	**0.8905**
Nano-Banana-Pro (Google DeepMind, 2025)	0.779	0.976	0.949	0.583	0.8706	0.8759
Seedream-5.0-Pro (ByteDance Seed, 2026)	—	0.941	0.946	0.553	0.8500	0.8616
Variance ↑	0.0195	0.0388	0.1846	0.0021	0.0396	0.0760
Human ρ ↑	0.619	0.619	0.527	0.539	**0.950**	**0.917**

Table 2: **TextCanvasBench diagnostic profiles.** Dimension scores for the English- and Chinese-prompt tracks. All scores are reported on a 0–100 scale.

Generator	English prompt				Chinese prompt			
	D1	D2	D3	D4	D1	D2	D3	D4
Open-weight models								
FLUX.1 [dev] (Black Forest Labs, 2024)	73.37	12.91	58.00	34.13	57.00	7.52	53.38	4.06
SD3.5 Large (Stability AI, 2024)	53.63	6.36	41.10	28.61	11.53	0.85	41.73	0.77
Qwen-Image (Wu et al., 2025)	85.49	41.21	85.24	82.27	86.75	43.56	87.31	87.12
Z-Image-Turbo (Cai et al., 2025)	93.28	49.80	88.86	74.17	93.07	51.57	86.94	79.99
GLM-Image (Z.ai Team, 2026)	92.88	53.10	83.35	73.97	93.22	54.17	78.94	77.43
Boogu-Image-0.1-Turbo (Chen et al., 2026)	89.28	55.02	87.56	83.80	89.14	54.80	90.05	88.66
Proprietary models								
GPT-Image-2 (OpenAI, 2026)	**96.76**	**69.63**	**90.75**	**98.21**	**96.42**	**68.88**	**92.19**	**98.69**
Nano-Banana-Pro (Google DeepMind, 2025)	95.33	66.15	88.83	97.93	95.61	66.44	89.87	98.42
Seedream-5.0-Pro (ByteDance Seed, 2026)	94.18	63.17	89.84	92.80	94.41	63.67	91.38	95.19

4.1 Performance as Text Amount and Layout Complexity Increase

Table 1 shows that TextCanvasBench separates models that existing benchmarks place close together, revealing performance differences that are less apparent in their reported scores. GLM-Image and GPT-Image-2, for example, differ by only 0.004 on LongText Chinese, but by 0.1311 on TextCanvasBench Chinese. The English track shows a similar pattern at the model-group level. The closed–open mean gap is 0.233 on TextCanvasBench, compared with 0.078 on CVTG, 0.159 on LongText English, and 0.052 on OneIG. TextCanvasBench therefore provides greater separation both for this closely scored model pair and between the closed and open model groups. Its rankings also agree more closely with human judgments, reaching Spearman ρ values of 0.950 in English and 0.917 in Chinese, versus at most 0.619 for the included external benchmarks.

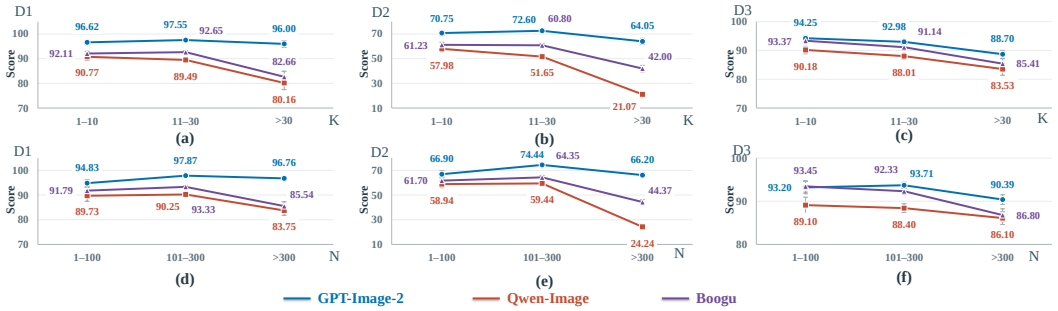

Figure 4: **Performance as textual requirements increase.** (a–c) D1–D3 by non-empty text-block count K; (d–f) D1–D3 by required character count N. Points are group means over common inputs across models and dimensions; unavailable records are excluded jointly without zero imputation. D3 uses the resolved-requirement quality score in Equation 2.

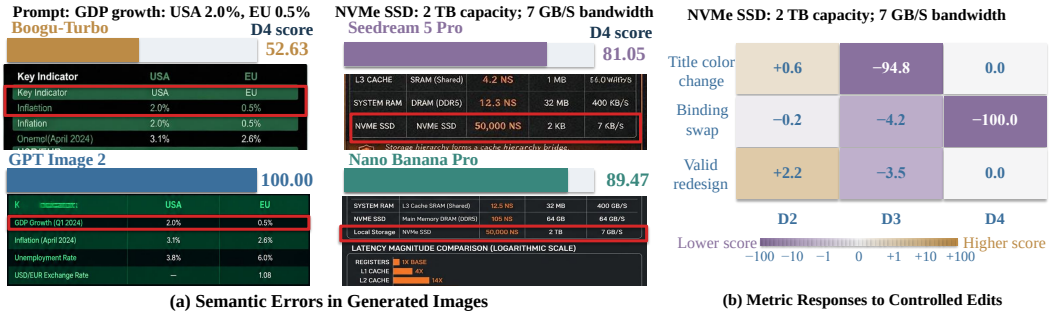

Figure 5: **Semantic errors and responses to controlled edits.** (a) Label–value and unit errors in generated images; red boxes identify the discussed details. (b) Mean score changes relative to each unedited image for title-color changes, text–object binding swaps, and valid redesigns.

Table 2 identifies the source of this separation by breaking down performance across individual capabilities. On the Chinese track, GLM-Image and GPT-Image-2 differ by only 3.20 points in D1, but by 14.71 in D2, 13.25 in D3, and 21.26 in D4. The difference between the models is therefore not uniform across dimensions: their D1 scores are relatively close, whereas the remaining dimensions show substantially larger gaps, with the largest difference in D4. Their similar LongText score consequently conceals substantial differences in text completeness, presentation fidelity, and semantic preservation. TextCanvasBench makes these differences visible in its overall scores and attributes them to specific capabilities through its diagnostic profiles, showing both how far the models differ and where that separation arises.

4.2 What Text Accuracy Misses

Figure 5a illustrates two limitations of text-only evaluation. In the GDP example, Boogu renders the requested 2.0% and 0.5% values but associates them with inflation rather than GDP growth; GPT-Image-2 preserves the intended association. Because the required strings remain visible, text accuracy alone cannot identify this error. In the storage example, Seedream-5.0-Pro changes 2 TB and 7 GB/s to 2 KB and 7 KB/s, whereas Nano-Banana-Pro preserves both units. A text metric can detect these substitutions, but D4 additionally captures their effect on the stated storage capacity and bandwidth. Figure 5b evaluates whether D3 and D4 respond selectively to their intended failure types. Across 24 base images, changing the title color reduces D3 by 94.8 points while leaving D4 unchanged. Swapping text–object associations reduces D4 by 100.0 points, while D2 changes by only -0.2. By contrast, a meaning-preserving redesign leaves D4 unchanged and causes only small changes in D2 and D3 ($+2.2$ and -3.5). These results show that D3 detects presentation errors, D4 detects semantic-association errors, and both remain stable under valid redesigns.

4.3 Ablation Study

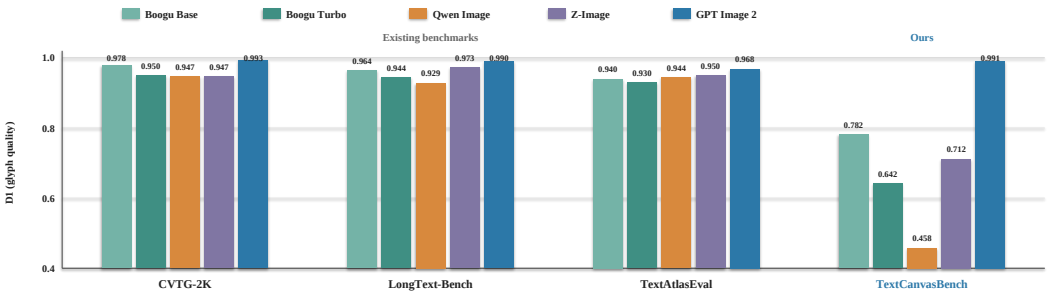

Figure 6: **Shared-evaluator benchmark ablation.** Five generators are scored with the same TextPecker D1 protocol. Existing benchmark subsets yield similar glyph-quality scores across models, whereas TextCanvasBench reveals a much wider spread.

To separate the effect of benchmark inputs from that of the scoring function, we apply the same full-image TextPecker protocol used by D1/D2 to outputs from five generators on CVTG-2K, LongText-Bench, TextAtlasEval, and TextCanvasBench. Figure 6 compares D1 scores across the four benchmarks. The cross-model ranges are 0.046 on CVTG-2K, 0.061 on LongText-Bench, and 0.038 on TextAtlasEval, compared with 0.533 on TextCanvasBench. Hence, the same evaluator produces tightly clustered scores on the three existing subsets but exposes substantially larger differences on TextCanvasBench. This result attributes the increased separation to the benchmark inputs rather than a change in the D1 scoring rule.

5 Related Work

Text-rich generation benchmarks. TextAtlasEval studies dense text generation (Wang et al., 2025), CVTG-2K evaluates multiple text instances in complex scenes (Du et al., 2025), and LongText-Bench tests extended bilingual text rendering (Geng et al., 2025). TextGround4M emphasizes prompt–layout alignment (Mao et al., 2026). OneIG-Bench covers multiple generation dimensions (Chang et al., 2026), while BizGenEval examines text, layout, attribute binding, and reasoning in commercial visuals (Li et al., 2026b). TextCanvasBench combines reference-grounded tasks across eight domains with stable text-block identities and separate presentation and semantic specifications. Character and block counts enable complementary analyses of textual and structural load. **Diagnostic evaluation.** TextPecker quantifies structural glyph anomalies (Zhu et al., 2026). For semantic evaluation, LongT2IBench uses entity–attribute–relation annotations (Yang et al., 2026), while IGenBench verifies infographic requirements through atomic questions (Tang et al., 2026). Building on these directions, our protocol separates glyph integrity, content correctness, layout/style fidelity, and semantic alignment. Frozen specifications define shared targets across generators; explicit evidence states distinguish confirmed violations from unresolved observations. Controlled edits further probe whether presentation and association changes affect the intended diagnostic dimensions.

6 Conclusion

We introduced TextCanvasBench, a benchmark of 1,600 specification-guided samples across eight domains and 38 subcategories of practical text-rich visuals. Its four dimensions assess glyph quality, complete-text accuracy, prompt-specified layout and style, and semantic alignment from the whole image to text blocks and regions. Across nine open- and closed-source generators, the benchmark distinguishes models with similar scores on existing benchmarks, and its rankings agree more closely with human judgments than those of most benchmarks in our comparison.

AI USE STATEMENT

We used large language models (LLMs) for language polishing and code debugging. Generative models used as components of benchmark construction and evaluation are described in the corresponding methodology sections. All AI-assisted outputs were reviewed and verified by the authors. The authors take full responsibility for the final content of this manuscript.

REFERENCES

Black Forest Labs. FLUX.1 [dev]: Model Card. Official model card, 2024. URL https://huggingface.co/black-forest-labs/FLUX.1-dev. Accessed September 16, 2026.

ByteDance Seed. Seedream 5.0 Pro. Official model page, 2026. URL https://seed.bytedance.com/en/seedream5_0_pro.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. *Advances in Neural Information Processing Systems*, 38, 2026.

Guoxuan Chen, Chufeng Xiao, Haoran Yang, Siyue Xie, Binxiao Huang, Ming Zhang, Cheuk Him Chau, Xinyu Fu, Yingzhao Lian, Tom S. Y. Li, Jintao Lin, Bowen Dong, Zian Qian, Yuhao Liu, Yuxuan Hu, Weikang Shi, Bin Zou, Bowen Zheng, Haoxuan Che, Chang Chen, Yuyang He, Heyang Sun, Tianyu Huang, Chong Hou Choi, Cheng Gong, Han Shi, Haoli Bai, Xihui Liu, Hongsheng Li, Qifeng Chen, Chao Huang, Rui Liu, and Chenyang Lei. Boogu-Image-0.1: Boosting open-source unified multimodal understanding and generation, 2026. URL https://arxiv.org/abs/2607.13125.

Nikai Du, Zhennan Chen, Zhizhou Chen, Shan Gao, Xi Chen, Zhengkai Jiang, Jian Yang, and Ying Tai. Textcrafter: Accurately rendering multiple texts in complex visual scenes. *arXiv e-prints*, pp. arXiv–2503, 2025.

Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, Linus, Di Wang, and Jie Jiang. X-Omni: Reinforcement learning makes discrete autoregressive image generative models great again, 2025. URL https://arxiv.org/abs/2507.22058.

Google DeepMind. Introducing Nano Banana Pro. Official announcement; Gemini 3 Pro Image, 2025. URL https://blog.google/innovation-and-ai/products/nano-banana-pro/. Accessed September 16, 2026.

Bingnan Li, Chen-Yu Wang, Haiyang Xu, Xiang Zhang, Ethan Armand, Divyansh Srivastava, Shan Xiaojun, Zeyuan Chen, Jianwen Xie, and Zhuowen Tu. Overlaybench: A benchmark for layout-to-image generation with dense overlaps. *Advances in Neural Information Processing Systems*, 38, 2026a.

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*, 2026b.

Dongxing Mao, Yilin Wang, Linjie Li, Zhengyuan Yang, and Alex Jinpeng Wang. Textground4m: A prompt-aligned dataset for layout-aware text rendering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7918–7926, 2026. doi: 10.1609/aaai.v40i10.37736.

OpenAI. GPT-Image-2 Model. Official model documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-image-2. Accessed September 16, 2026.

Qwen Team. Qwen-Image-2512. Official model card, 2025. URL https://huggingface.co/Qwen/Qwen-Image-2512. Accessed September 16, 2026.

Stability AI. Stable Diffusion 3.5 Large: Model Card. Official model card, 2024. URL https://huggingface.co/stabilityai/stable-diffusion-3.5-large. Accessed September 16, 2026.

Ying Tai, Nikai Du, Rui Xie, Zhennan Chen, Qian Wang, Zhengkai Jiang, Kai Zhang, and Jian Yang. Investigating text insulation and attention mechanisms for complex visual text generation. *arXiv preprint arXiv:2503.23461*, 2025.

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.

Alex Jinpeng Wang, Dongxing Mao, Jiawei Zhang, Weiming Han, Zhuobai Dong, Linjie Li, Yiqi Lin, Zhengyuan Yang, Libo Qin, Fuwei Zhang, Lijuan Wang, and Min Li. Textatlas5m: A large-scale dataset for dense text image generation, 2025. URL https://arxiv.org/abs/2502.07870.

Zengbin Wang, Xuecai Hu, Yong Wang, Feng Xiong, Man Zhang, and Xiangxiang Chu. Everything in its place: Benchmarking spatial intelligence of text-to-image models. *arXiv preprint arXiv:2601.20354*, 2026.

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.

X-Omni Team. X-omni: Official inference code and longtext-bench. Official benchmark repository, 2025. URL https://github.com/X-Omni-Team/X-Omni. Category labels verified from textbench/text_prompts.jsonl, accessed September 14, 2026.

Zhichao Yang, Tianjiao Gu, Jianjie Wang, Feiyu Lin, Xiangfei Sheng, Pengfei Chen, and Leida Li. Longt2ibench: A benchmark for evaluating long text-to-image generation with graph-structured annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11820–11828, 2026. URL https://arxiv.org/abs/2512.09271.

Z.ai Team. GLM-Image: Auto-regressive for Dense-knowledge and High-fidelity Image Generation. Official model repository, 2026. URL https://github.com/zai-org/GLM-Image. Accessed September 16, 2026.

Hanshen Zhu, Yuliang Liu, Xuecheng Wu, An-Lan Wang, Hao Feng, Dingkang Yang, Chao Feng, Can Huang, Jingqun Tang, and Xiang Bai. Textpecker: Rewarding structural anomaly quantification for enhancing visual text rendering. *arXiv preprint arXiv:2602.20903*, 2026. URL https://arxiv.org/abs/2602.20903.

A Dataset Statistics and Benchmark Comparison

Because evaluation proceeds from the whole image down to individual text blocks, we report dataset statistics at the corresponding macro and micro levels. Macro-level statistics characterize the complete required-text inventory of each sample. Micro-level statistics characterize its decomposition into localized text blocks with functional roles. (Table 3)

Table 3: **Macro- and micro-level dataset statistics.** All counts use the 1,600-input inventory. The two right-hand groups are separate cross-tabulations of the same samples; K bins are not layout levels. P90 denotes the 90th percentile, computed by linear interpolation.

Per-image quantity	Mean	Median	P90	Range
Characters N	636.52	241	1,643.20	4–8,423
Text blocks K	26.76	17	64	1–248

		Layout level			Text blocks K		
Characters N	Samples	L1	L2	L3	1–10	11–30	> 30
1–100	262	77	108	77	212	49	1
101–300	658	106	296	256	186	424	48
> 300	680	46	200	434	54	229	397
Total	1,600	229	604	767	452	702	446

A.1 Domain and Subcategory Coverage

The eight domains expose different combinations of textual content and visual organization (Table 4). Academic documents and slides contain the most required text on average, whereas infographics have the highest median block count and L3 share. Digital interfaces contain a median of 27 blocks despite a mean of only 324.7 characters, reflecting their dense use of short labels and values. Together, these profiles show that the same amount of required text can be organized differently across practical visual media.

Table 4: **Domain-level text and layout profiles.** Long-text and L3 percentages use the corresponding domain count as denominator.

Domain	Samples	Mean chars.	Median blocks	$N > 300$ (%)	L3 (%)
Infographics	310	1340.4	57	95.48	89.35
Posters & Ads	265	258.9	15	23.77	45.28
Art & Illustration	215	298.0	9	22.79	46.51
Document Design	225	730.1	14	37.33	19.56
Real-world Photos	160	204.7	11	16.88	32.50
Spatial Signage	160	217.6	9.5	19.38	35.62
Digital Interfaces	170	324.7	27	41.76	51.76
Academic & Slides	95	1928.1	21	62.11	30.53

Table 5 expands the eight domains into 38 concrete content topics and use cases. Domain-specific subcategory counts range from two for Academic & Slides to nine for Infographics, while individual subcategories contain 24–84 samples. This hierarchy preserves broad visual media coverage while preventing the benchmark from collapsing into a single content type within each domain.

A.2 Comparison with Related Benchmarks

Table 6 positions TextCanvasBench among benchmarks for visual text generation and related compositional tasks. Our design combines practical task coverage with text-block identities, functional roles, explicit presentation requirements, and atomic semantic questions. These annotations support evaluation from the whole image down to individual blocks, separating glyph structure quality, textual content correctness, layout and style, and text–image semantic alignment that aggregate benchmark scores may obscure.

The comparison separates text-instance annotations from the object layouts used in adjacent tasks. Joint strata retain each benchmark's native text-unit definition, while explicit checks include prompt-

Table 5: **Taxonomy and sample counts across 38 subcategories.** Parentheses report subcategory counts; the final column reports domain totals. Each of the 1,600 samples belongs to exactly one subcategory.

Domain	Subcategories (samples)	Total
Infographics	Society & humanities (29); Geography (35); Data charts (29); Life & health (29); Everyday guides (29); Science & engineering (52); Economy & business (28); Natural sciences (47); Arts & culture (32).	310
Posters & Ads	Public-service posters (60); Commercial ads (41); Film & TV posters (49); Recruitment posters (43); Event posters (72).	265
Art & Illustration	Traditional painting (41); Illustrations (84); Comics & anime (46); Modern collage (44).	215
Document Design	Newspaper pages (26); Resumes (42); Thank-you letters (40); Holiday greetings (46); Invitations (71).	225
Real-world Photos	Meetings (24); Event scenes (37); Objects & products (50); Everyday settings (25); Natural settings (24).	160
Spatial Signage	Traffic signage (44); Public facilities (33); Commercial spaces (50); Industrial signage (33).	160
Digital Interfaces	Apps (44); Game interfaces (38); Web pages (25); Software interfaces (63).	170
Academic & Slides	Presentation slides (50); Paper pages (45).	95
Total	**38 subcategories across eight domains**	**1,600**

Table 6: **Benchmark annotations and evaluation design.** ✓: supported; —: unverified; N/A: outside the stated task. Unknown entries do not indicate absence. Positions refer to annotations, not evaluator-predicted boxes; reference images need not be generation inputs.

Benchmark	Annotations					Evaluation design		
	Text instances	Position	Style	Roles	Ref. images	Joint strata	Explicit checks	Semantic method
TextAtlasEval (Wang et al., 2025)	—	—	—	—	✓	—	—	CLIP
CVTG-2K (Du et al., 2025)	✓	Text	✓	—	—	—	—	CLIP / VQA
LongText-Bench (X-Omni Team, 2025)	—	—	—	—	—	—	—	—
OneIG-Bench (Chang et al., 2026)	—	—	—	—	—	—	—	VLM
TextGround-Bench (Mao et al., 2026)	✓	Bbox	—	—	✓	✓	—	CLIP
BizGenEval (Li et al., 2026b)	—	Text	✓	—	✓	—	✓	Checklist VQA
IGenBench (Tang et al., 2026)	—	—	—	—	✓	—	✓	Atomic VQA
LongT2IBench (Yang et al., 2026)	N/A	N/A	N/A	N/A	—	N/A	✓	Graph-based VLM
SpatialGenEval (Wang et al., 2026)	N/A	N/A	N/A	N/A	—	N/A	—	Spatial MCQ
OverLayBench (Li et al., 2026a)	N/A	N/A	N/A	N/A	✓	N/A	N/A	—
TextCanvasBench (ours)	✓	Text	✓	✓	✓	✓	✓	Atomic VQA

derived questions and graph criteria; neither implies an identical evaluation protocol. OneIG's semantic entry refers to its broader evaluation framework.

B D3 AND D4 SPECIFICATION PREPARATION

This section expands the specification-preparation stages introduced in Sections 3.2 and 3.3 through a single workplace-health poster. The final generation prompt and required text define the requirements to be evaluated, while reviewed reference annotations identify the intended text occurrences during preparation. The resulting D3 and D4 specifications are fixed before any model output is evaluated.

B.1 Generation Prompt Construction

Figure 7 illustrates the conversion of a reference image into a self-contained generation prompt. Gemini 2.5 Pro drafts Chinese and English scene descriptions and extracts the required text. The selected scene description specifies the visual medium, composition, text organization, and explicit presentation requirements; a separate required-text section preserves wording, capitalization, numbers, punctuation, and repeated occurrences. A rendering instruction requires complete and accurate reproduction of this text. The generator receives the assembled prompt without the reference image.

The description makes selected visual requirements explicit: "In the top left" specifies placement, "below which" specifies spatial order, and "white text" specifies foreground color. These phrases provide source evidence for the presentation specifications in Appendix B.2. Reference-image details omitted from the final prompt do not become additional evaluation requirements.

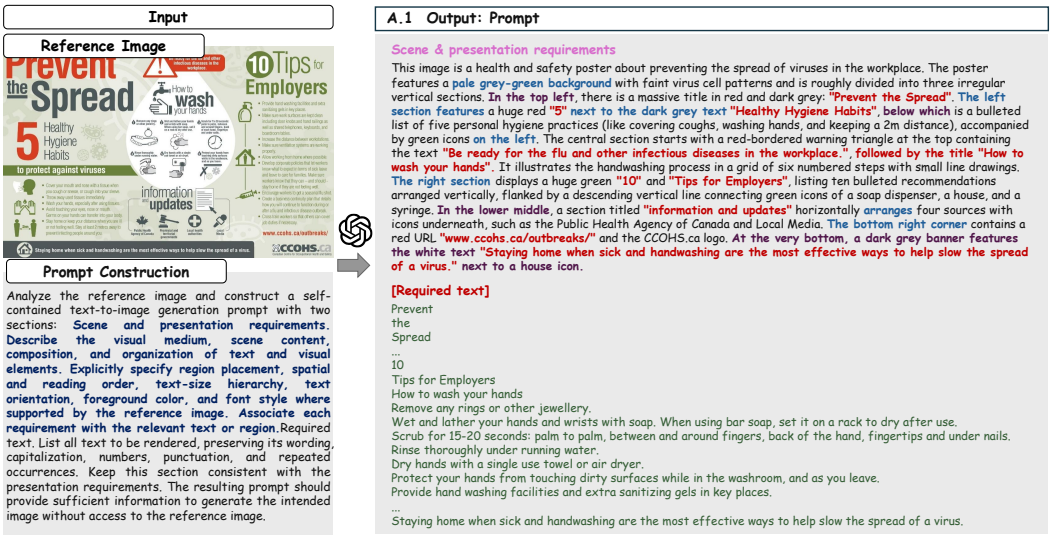

Figure 7: **Generation prompt construction.** The reference image and construction instruction yield a scene description with explicit presentation requirements and a separate required-text section. The displayed text is excerpted; the complete generation input retains all required strings.

B.2 Structured Presentation Specifications (D3)

D3 specification preparation proceeds in four stages. First, we decompose each explicit presentation statement in the final prompt into atomic requirements. Each requirement retains its exact source phrase, the affected text, one of six attribute families—placement, spatial/reading order, size hierarchy, orientation, foreground color, and font style—and the requested value or relation. Compound statements are split so that every scored requirement tests a single attribute.

Second, a semantic review classifies each extracted requirement as supported, ambiguous, or unsupported by its quoted prompt phrase. A reverse audit then checks that every explicit presentation statement in the prompt is represented. This prevents unsupported substitutions between presentation attributes and prevents global scene descriptions from becoming unstated text-level constraints.

Third, supported requirements are associated with the corresponding annotated text occurrences. Stable identifiers are retained in the machine-readable record to distinguish repeated strings. A requirement applying to all text is expanded over the relevant occurrences, while a relational requirement retains all participating occurrences. These associations specify what must be inspected in a generated image; they do not transfer reference-image pixel coordinates or colors into the evaluation target.

Finally, each reviewed requirement is compiled into a scoring entry containing the target text, attribute family, expected value or relation, applicable rule, and generated-image observations required by that rule. Automated validation checks source phrases, identifiers, duplicate entries, prompt–

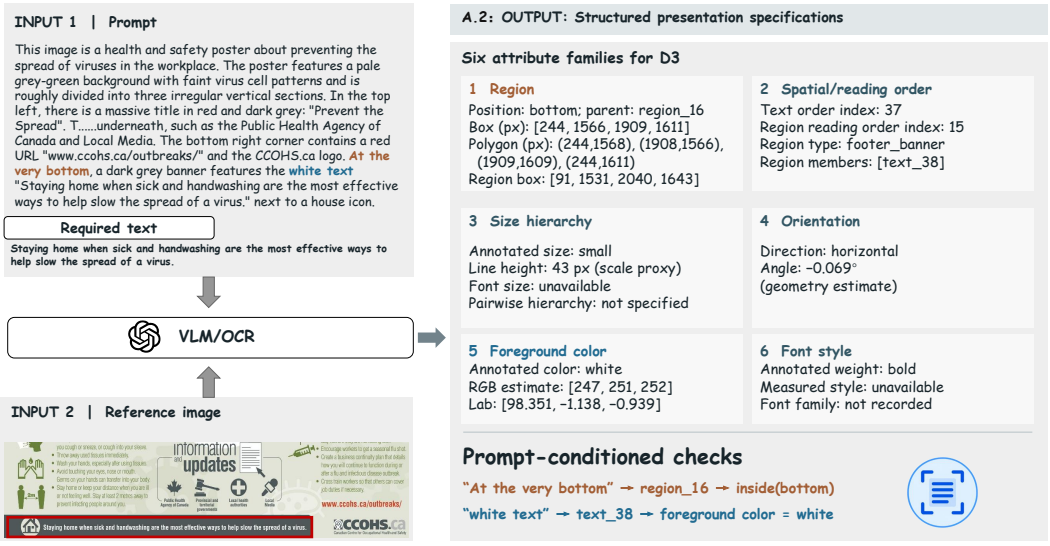

Figure 8: **D3 specification preparation.** Explicit prompt statements are reviewed, associated with the intended text, and compiled into requirements from the six D3 attribute families.

requirement correspondence, and rule availability. The specification is frozen only after this audit and human review.

B.3 Fine-Grained Semantic Specifications (D4)

D4 is prepared from the final prompt and required text without inspecting model outputs. We first partition the required text into coherent passages and extract atomic semantic facts. Each fact preserves its subject, attribute, value, unit, condition, negation, and association when present, together with its source passage. A second semantic review removes unsupported inferences and locally repairs facts whose wording changes the source meaning. At most five facts that best preserve the sample's core textual content are selected for the first D4 question.

The prompt supplies the remaining semantic requirements. We identify the requested overall theme and communicative intent, associations between text and visual content, scene-level logical constraints, and text–carrier integration when a physical carrier is explicitly specified. Each requirement retains its source phrase and states which visible text, entity, or relation must be inspected in the generated image. Carrier integration is marked inapplicable when the prompt does not specify a text-bearing surface.

The selected facts and prompt-derived requirements populate the five fixed questions in Table 7. Specification preparation defines the meaning to be preserved and the observations needed for assessment; it does not judge whether a generated image satisfies them. The complete D4 specification undergoes a final semantic review and is then frozen across all evaluated models.

Table 7: The five fixed D4 questions.

Component	Fixed question
Core textual semantics	Does the rendered text preserve the selected core meanings, conditions and associations?
Intent	Does the text in context convey the requested communicative purpose, theme and attitude?
Text–image	Is the meaning of the rendered text consistent with the objects, graphics or scene it refers to?
Logic	Are meanings within and between text passages and their associated graphics logically consistent with the requested conditions?
Carrier	Does the text physically integrate with its specified carrier, including material, deformation, occlusion and lighting where observable?

Prompt

Required text | text_38

Staying home when sick and handwashing are the most effective ways to help slow the spread of a virus.

Scene description (excerpts)

"a health and safety poster about preventing the spread of viruses in the workplace"

"ten bulleted recommendations"
"five personal hygiene practices"

Source illustration | same sample



A.3: OUTPUT: Fine-grained semantic specifications

1 Core textual semantics | F004 → text_38

Staying home when sick and handwashing are the most effective ways to help slow the spread of a virus.

→ **Required evidence:** actual generated text preserving both actions, the "when sick" condition, and their effect on virus spread.

2 Communicative intent | C001

Source: "Prevent … Spread" (text_1/3); "Tips for Employers" (text_6); "Healthy Hygiene Habits" (text_8); "How to wash your hands" (text_9).

→ **Criterion: communicate workplace guidance for preventing virus spread.**

→ **Required evidence:** readable headings and associated content in the generated image that convey this purpose.

3 Text-image consistency | C002

→ **Required evidence:** "5" and "10" must correspond to their titled lists.
→ **Required evidence:** visible numerals, headings, list items and their associations.

4 Semantic logic | C003

→ **Required evidence:** Each stated count must agree with its list.
→ **Required evidence:** distinct item counts and the numeral–list relation.

5 Physical-carrier integration | N/A

No physical text-bearing surface is explicitly required.

Figure 9: **D4 specification preparation.** Core facts and the remaining prompt-derived semantic requirements retain their input sources and define the observations needed for the five fixed questions.

F001 | text_13 Readable text required Evidence scope: local	Meaning: Scrubbing hands should last 15-20 seconds covering palm to palm, between and around fingers, back of the hand, fingertips and under nails.	Source: "Scrub for 15-20 seconds: palm to palm, between and around fingers, back of the hand, fingertips and under nails."
F002 | text_31 Readable text required Evidence scope: local	Meaning: When ill, stay at least 2 metres away to prevent infecting people around you.	Source: "Stay at least 2 metres away to prevent infecting people around you."
F003 | text_30 Readable text required Evidence scope: local	Meaning: Germs on your hands can transfer into your body.	Source: "Germs on your hands can transfer into your body."
F004 | text_38 Readable text required Evidence scope: local	Meaning: Staying home when sick and handwashing are the most effective ways to help slow the spread of a virus.	Source: "Staying home when sick and handwashing are the most effective ways to help slow the spread of a virus."
F005 | text_24 Readable text required Evidence scope: local	Meaning: Create a business continuity plan detailing how you will continue to function during or after a flu and infectious disease outbreak.	Source: "Create a business continuity plan that details how you will continue to function during or after a flu and infectious disease outbreak."

Figure 10: **Selected core semantic facts.** Meanings and source passages define five facts for the core-textual-semantics question.

B.4 Validation and Specification Freezing

Human review checks required text, text localization, and grouping against the reference image. For D3, it verifies that every requirement follows from its quoted prompt phrase and refers to the intended text. For D4, it verifies that each selected fact preserves its source meaning and conditions, and that the remaining criteria and their applicability follow from the final prompt. Automated checks validate source spans, identifiers, duplicate entries, requirement completeness, the five-fact limit, and scoring-rule availability. The reviewed specifications are versioned with the final input hash and fixed across evaluated generators; an input revision triggers renewed validation. Generated-image observations and scores are produced separately under the protocol in Appendix C.

C D3 and D4 Evaluation Details

The frozen D3 and D4 specifications define the properties and meanings to be tested, while each generated image supplies the observations used in those tests. Evaluation compares generated-image observations with prompt-derived requirements and never compares generated pixels directly with reference pixels.

C.1 D3: Layout and Style Evaluation

Observation extraction. PP-OCRv5 first proposes text locations in the generated image. A VLM then matches these detections to the target text for the sample and measures only the attributes requested by the frozen D3 specification. The observer is not shown the expected attribute values. When the initial localization or attribute decision is ambiguous, a targeted VLM pass re-examines the relevant text and surrounding context.

Each prompt-derived requirement receives an observation for its subject and, when applicable, its comparison target. The corresponding rule converts the observed placement, relation, scale, orientation, foreground color, or font style into a requirement score $q_{ic} \in [0, 1]$. A confirmed missing target receives zero; all evaluated requirements are resolved before aggregation.

Per-requirement scores. All coordinates are normalized to $[0, 1]^2$, and $[x]_0^1 = \min(1, \max(0, x))$. Named canvas regions are fixed boxes formed by thirds and halves of the image: for example, top is $[0, 0, 1, \frac{1}{3}]$, bottom-left is $[0, \frac{2}{3}, \frac{1}{3}, 1]$, center is $[\frac{1}{3}, \frac{1}{3}, \frac{2}{3}, \frac{2}{3}]$, and right-half is $[\frac{1}{2}, 0, 1, 1]$. Table 8 gives the score assigned to each supported requirement.

D3 aggregation. Let $\mathcal{C}_{ik}$ denote the prompt-derived requirements in attribute family k for sample i, and let $q_{ic} \in [0, 1]$ be the score assigned by the corresponding rule. We first average requirements within each applicable family, then combine family scores:

$$D3_i = 100 \sum_{k \in \mathcal{A}_i} \widetilde{w}_{ik} \frac{\sum_{c \in \mathcal{C}_{ik}} q_{ic}}{|\mathcal{C}_{ik}|}, \qquad \widetilde{w}_{ik} = \frac{w_k}{\sum_{j \in \mathcal{A}_i} w_j}, \tag{5}$$

where $\mathcal{A}_i$ contains the applicable families and $w = (0.25, 0.20, 0.15, 0.10, 0.15, 0.15)$ follows the order in Table 8. Averaging within a family prevents prompts with many requirements of one type from overwhelming the remaining attributes. Weights are renormalized when a family is not requested. The specification, observation procedure, tolerances, and aggregation rule are fixed across all evaluated models.

C.2 D4: Semantic Evaluation

Evaluation inputs and outputs. For each sample i, the D4 evaluator receives the generated image I_i and the frozen D4 specification constructed before model comparison; the reference image is not used for scoring. The specification supplies the applicable requirements under five semantic aspects: core textual semantics, communicative intent, text–image consistency, semantic logic, and physical-carrier integration. The evaluator produces an evidence record and a judgment for every applicable requirement, five aspect-level results, and the final scalar score $D4_i$.

Table 8: **D3 per-requirement scoring.** Each rule maps the observed attribute or relation to a score in $[0, 1]$.

Attribute	Score calculation
Placement	For target center (x, y) and allowed box $B = [l, t, r, b]$, let $\delta_B = \max\{l - x, 0, x - r, t - y, 0, y - b\}$. The score is $\max(0, 1 - \delta_B/0.05)$: a center inside B receives 1, and credit falls linearly to 0 within 5% of image size outside the boundary.
Order	For left/right/above/below or reading-order relations, let Δ be the signed separation of the two target centers in the required direction. The score is $[\Delta/0.02]_0^1$; a tie or reversed order receives 0. For a same-row constraint, the score is $\max(0, 1 - (\max y - \min y)/0.02)$ over all participating centers.
Size hierarchy	Text scale is measured from comparable glyph geometry. For "larger than," the score is $[(\min h_s / \max h_o - 1)/0.15]_0^1$; "smaller than" reverses the ratio. For "same level," the score is $[1 - |\max h / \min h - 1|/0.15]_0^1$. Thus a 1.15 ratio gives full credit for a larger/smaller relation, and every required subject–object pair must satisfy it.
Orientation	A categorical horizontal/vertical/angled observation is matched directly. For a requested numeric rotation, angular error up to 5° receives full credit, then decreases linearly to 0 at 20°. Glyph rotation uses circular error modulo 360°; writing direction uses a baseline axis modulo 180°.
Foreground color	Foreground-stroke RGB is converted to HSV and compared with the requested named color family, rather than with an exact reference RGB. Examples include red (0–20° or 340–360°), green (80–160°), blue (200–260°), cyan (160–200°), and purple (260–315°). Hue outside a family falls linearly to 0 within 15°; light/dark modifiers additionally constrain value. Black, white, and gray use saturation/value thresholds.
Font style	Serif/sans and explicit attributes such as bold, italic, handwritten, calligraphic, and outline are observed independently. A confirmed match receives 1 and a confirmed contradiction receives 0. Ambiguous initial observations are re-examined before scoring. Attribute observations require confidence of at least 0.8.

Step 1: evidence extraction. For every applicable requirement, a VLM inspects I_i and first records what is visibly present. For core textual facts, it transcribes the relevant generated text and checks the required subject, attribute, value, unit, condition, negation, and association. For communicative intent, text–image consistency, and semantic logic, it inspects the participating text, entities, graphics, and relations. When physical-carrier integration is applicable, it also examines whether the lettering is attached to and visually integrated with the specified surface, including deformation, occlusion, material, and lighting cues. Every judgment must cite the corresponding image observations.

Step 2: evidence validation. The validation stage checks that every applicable requirement has been addressed, that each cited observation supports the requirement to which it is linked, and that the aspect-level answer agrees with its requirement-level judgments. If an observation is incomplete or ambiguous, a targeted second VLM pass re-examines only the unresolved requirement and its local context. After validation, each applicable requirement is assigned $q \in \{1, 0.5, 0\}$ for satisfied, partially satisfied, or not satisfied.

Step 3: aspect scoring and aggregation. The core textual-semantics aspect scores its selected facts individually. If it contains M_i facts, its score is $r_{i,\text{core}} = M_i^{-1} \sum_{m=1}^{M_i} q_{im}$; thus a partially satisfied fact contributes $0.5/M_i$ rather than capping the entire aspect at 0.5. Each of the other four aspect scores is obtained from its validated requirements. With $v = (0.45, 0.10, 0.20, 0.20, 0.05)$ for core semantics, intent, text–image consistency, logic, and carrier integration, respectively,

$$D4_i = 100 \sum_{k \in \mathcal{B}_i} \widetilde{v}_{ik} r_{ik}, \qquad \widetilde{v}_{ik} = \frac{v_k}{\sum_{j \in \mathcal{B}_i} v_j}, \tag{6}$$

where $\mathcal{B}_i$ contains the applicable aspects. An inapplicable aspect is excluded, and the remaining weights are renormalized. The reported output retains the supporting observations and requirement-level judgments alongside the five aspect results and $D4_i$. The same specification, observation procedure, validation rules, and aggregation are applied to every evaluated model.