EXTREMETEXTBENCH:EVALUATING LONG-FORM TEXT RENDERING AND TYPEFACE FIDELITY IN IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent image-generation models have made rapid progress in text rendering. However, most existing benchmarks evaluate rendering accuracy under conventional settings, focusing primarily on short words, phrases, or moderately long text. As model capabilities continue to advance, these settings offer increasingly limited discriminative power and provide little insight into whether models can satisfy demanding design requirements, including long-form text rendering, small-scale text rendering, and precise typeface control. We introduce **ExtremeTextBench**, a bilingual benchmark that systematically evaluates two underexplored capabilities: rendering text under extreme length and size conditions, and adhering to explicit typeface constraints. ExtremeTextBench comprises 600 English and Chinese prompts, 3,695 annotated text segments, 78 typeface families, and 24 design categories drawn from realistic text-rich applications, including posters, presentation slides, e-commerce graphics, menus, and invitations. By analyzing performance across text-length and text-size ranges (with target lengths of up to 428 words in English and 1,584 characters in Chinese, and an observed small-text range of $\leq$ 20 pixels), ExtremeTextBench separately examines how rendering performance varies along these two dimensions, providing a more discriminative assessment of current state-of-the-art models. We also introduce a typeface-control evaluation method to assess whether generated text faithfully follows explicitly specified typefaces. Experiments on 16 state-of-the-art image-generation models reveal persistent limitations in rendering ultra-long text and text at small scales, while precise typeface control remains an open challenge. ExtremeTextBench fills an important gap in existing evaluation by providing a realistic and diagnostic testbed for the next generation of text-capable image-generation models.

1 INTRODUCTION

Accurate text rendering has become an increasingly important capability for text-to-image (T2I) generation models. Recent models have made substantial progress, evolving from producing largely illegible pseudo-text to rendering short words and phrases with recognizable content (Chen et al., 2023; Tuo et al., 2024; Du et al., 2025), thereby broadening their applicability to a wider range of real-world graphic-design tasks, such as posters, presentation slides, e-commerce graphics, menus, advertisements, and social-media designs. However, these applications often require a single image to accommodate tens or hundreds of words or characters across titles, subtitles, body paragraphs, numerical labels, contact information, and brand slogans, with many text regions rendered at extremely small scales and in explicitly specified typefaces (Figure 1). Therefore, a meaningful text-rendering evaluation aligned with practical requirements should go beyond isolated, prominently displayed text. It should systematically examine whether models preserve textual correctness as content length increases and text size decreases, characterize the resulting degradation patterns, and assess whether the generated typography adheres to the requested typeface constraints.

Most existing text-rendering benchmarks, however, primarily evaluate isolated words, short phrases, or moderately long text rendered at conventional, readily visible scales (Chen et al., 2023; Tuo et al., 2024; Du et al., 2025; Geng et al., 2025). As model capabilities continue to improve, these

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.

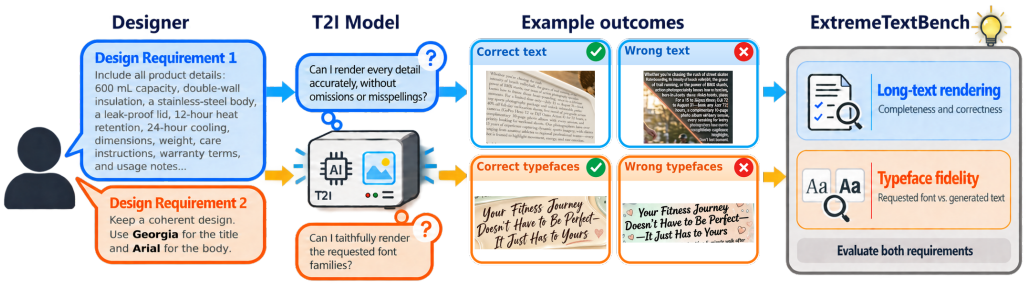


Figure 1: Motivation for ExtremeTextBench.

settings offer increasingly limited discriminative power and fail to characterize how rendering performance degrades when models must generate ultra-long content or extremely small text. Moreover, existing benchmarks commonly reduce text-rendering quality to content correctness, relying on OCR-derived metrics such as recognition accuracy and normalized edit distance (NED), while overlooking typography requirements essential to real-world graphic design, such as adherence to explicitly specified typefaces. In ExtremeTextBench, the term **extreme** therefore denotes a combination of ultra-long textual content, multiple text regions, high text density, sub-20-pixel text, visually complex backgrounds, and multi-typeface compositions with explicit typeface-family constraints. By evaluating models under these challenging conditions, ExtremeTextBench provides a more discriminative assessment of current text-rendering capabilities and tests whether progress observed in conventional settings transfers to demanding real-world T2I applications.

In summary, our main contributions are threefold:

1. We introduce ExtremeTextBench, a bilingual benchmark to probe text-rendering capability boundaries across diverse design types through stress tests spanning typeface fidelity, text length, and text size. Curated under professional designers' guidance with rigorous quality control, it comprises 600 prompts, 3,695 annotated text segments, 78 typeface families, and 24 design categories.

2. We establish a diagnostic protocol combining OCR-based text evaluation across length and size ranges with complementary MLLM-based typeface assessment. Using segment-level typeface annotations, the latter compares generated text crops against verified reference renderings to measure adherence to specified font families.

3. Our extensive evaluation of 16 state-of-the-art models reveals substantial capability differences under extreme rendering demands and a gap between textual correctness and typeface fidelity. Comparisons with CVTG-2K(Du et al., 2025) and LongText-Bench(Geng et al., 2025) further show wider score separation among shared models.

2 Related Works

2.1 Benchmarks for Text-Rich Image Generation

As image generation systems increasingly target professional design applications, text-rendering benchmarks have expanded from short, isolated words to multi-region and long-form content. MARIO-Eval (Chen et al., 2023) evaluates visual text across natural images and graphic-design scenarios, while AnyText-benchmark (Tuo et al., 2024) provides English and Chinese test sets for visual text generation. CVTG-2K (Du et al., 2025) focuses on rendering multiple text regions in complex scenes, and the text-rendering subsets of OneIG-Bench (Chang et al., 2026) and WeGenBench (Liang et al., 2026) broaden bilingual evaluation across diverse scenarios. LongText-Bench (Geng et al., 2025) further examines longer content in practical categories such as posters, webpages, slides, and printed materials, using vision-language-model transcription to assess text accuracy. However, with its English long-text targets predominantly containing 30–100 words, LongText-Bench does not sufficiently challenge current leading models, limiting its ability to distinguish their text rendering capabilities. STRICT (Zhang et al., 2025) extends length-based stress testing to substantially longer multilingual text, although its fixed document-page setting provides

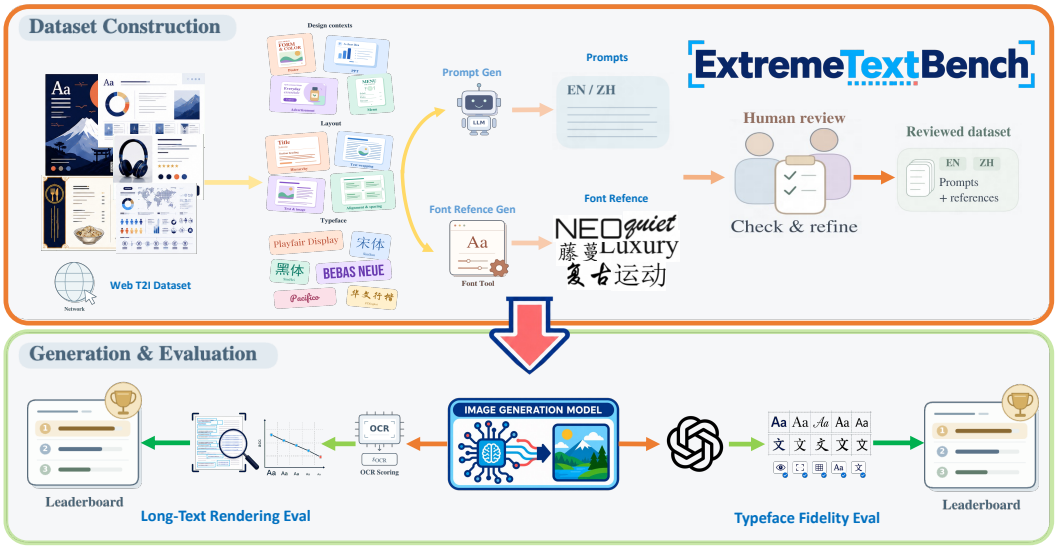


Figure 2: Overview of the construction and evaluation pipeline.

limited design and typography constraints. These benchmarks address complementary rendering challenges.

2.2 Typeface Recognition and Typography Evaluation

Typeface research spans font recognition, perceptual attributes, and evaluation of generated typography. VFR-2420/VFR-Wild (Chen et al., 2014) and AdobeVFR (Wang et al., 2015), support font recognition across synthetic and real-world text images. (O'Donovan et al., 2014) provide crowdsourced font-attribute and similarity annotations for studying human typographic perception. These resources inform font identification and comparison, while generation benchmarks examine adherence to requested typography. LeX-Bench (Zhao et al., 2025) evaluates font, color, and position attributes, with a dedicated Font-Score based on GPT-4o judgments of contrasting font styles. This measures broad stylistic distinctions, providing limited evidence of adherence to specific typeface families. Graphic-Design-Bench (Deganutti et al., 2026) includes recognition across 167 font families and evaluates font control in its Styled Text Generation task using Font Accuracy. Its emphasis on isolated text generation and masked-layout completion, however, leaves typeface fidelity in complete bilingual, text-rich designs less explored. Together, these datasets provide foundations for attribute-level and family-level evaluation.

3 ExtremeTextBench: Dataset and Evaluation Protocol

This section introduces the design principles and task definitions of ExtremeTextBench. The overall construction and evaluation process of ExtremeTextBench is illustrates in Figure 2

3.1 Design Principles and Task Definition

ExtremeTextBench targets realistic graphic-design applications, including posters, presentation slides, e-commerce graphics, and menus..., etc. It combines long-form text, multiple text regions, and explicit typeface constraints with requirements for visual subjects, backgrounds, and layouts, defining two complementary evaluation tasks.

Task I: Long-Text Rendering This task evaluates whether models can accurately and completely render specified text within a limited image canvas under multi-region design constraints. Each sample provides a complete generation prompt specifying visual content and layout requirements, together with the full target text and semantically organized text segments for evaluation. Rendering long-form content alongside headings, body text, and auxiliary information poses joint challenges

to content completeness and small-text rendering, establishing the benchmark's extreme rendering conditions. Evaluation focuses on textual correctness and examines its relationship with target-text length and observed text size. Layouts and visual elements serve as task conditions and are not independently scored for aesthetic quality.

Task II: Typeface Fidelity This task evaluates whether models render text in the explicitly requested typeface families, covering both single-typeface designs and compositions with different typefaces across segments. Each target segment is annotated with its typeface family and paired with a reference image illustrating that family and a rendering of the segment in the specified typeface. Evaluation aligns detected text regions with target segments and corresponding rendered typeface references. The task shares samples and target-text annotations with Task I, providing a complementary assessment of typeface control.

3.2 Dataset Construction and Annotation

We construct ExtremeTextBench through four stages: scenario specification, target-text generation, typeface assignment, and structured annotation. The resulting prompts combine long-form text and segment-level typeface requirements with visual content and layout constraints.

Drawing on public graphic-design datasets(Deganutti et al., 2026)(Liang et al., 2024)(Li et al., 2026), we define a scenario taxonomy with 24 design categories, 20 industries, 20 visual styles, 13 layouts, and attributes for text elements and mood. These attributes specify the application, theme, visual presentation, and information structure, forming base prompts that describe the design purpose, visual subjects, and backgrounds.

Table 1: Comparison of text-rendering benchmarks. * denotes values computed by us from officially released data where official statistics are not reported. Prior-work sizes follow the cited papers; N/R: not explicitly reported. ●: exact named-family control; ◐: partial support; ○: no systematic font control; –: no closed font inventory; N/D: no closed category taxonomy.

Benchmark	Text length		Typeface		Design
	Data	Max length	Supp.	# Fonts	# Cat.
CVTG-2K (Du et al., 2025)	2,000	30 words*	◐	4 styles	N/D
LongText-Bench (Geng et al., 2025)	160	EN: 66 words* ZH: 121 chars*	○	–	8
OneIG-Bench-Text (Chang et al., 2026)	400[a]	EN: 187 words* ZH: 540 chars*	◐	–	10
WeGenBench-Text (Liang et al., 2026)	2,000	EN: 133 chars ZH: 455 chars*	◐	–	10
STRICT (Zhang et al., 2025)	N/R	5,000 chars	○	–	1
ExtremeTextBench(Ours)	**600**	**EN: 428 words ZH: 1,584 chars**	●	**78**	**24**

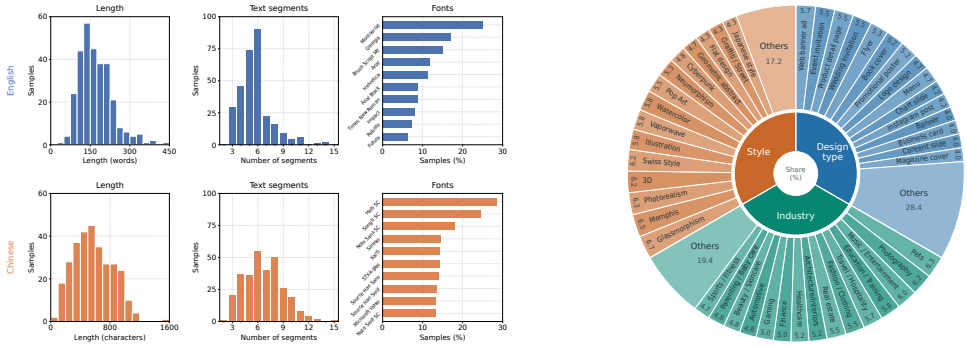
Conditioned on these base prompts, GPT-5.5 (OpenAI, 2026b) generates long-form English and Chinese target text together with layout descriptions, maintaining thematic consistency and coherent semantic organization. The text is divided into segments, such as titles, subtitles, body text, and contact information, whose roles and relationships are specified in the layout descriptions. We retain the complete target text for image-level evaluation and the segment annotations for aligning OCR-detected regions with the intended content.

We then assign a typeface family to each segment and incorporate the corresponding constraint into the prompt, supporting both single- and multi-typeface designs. Each assignment is checked against the available font resources to verify the family identifier and support for the target language and characters. We prepare reference images illustrating the assigned typeface families and render each segment in its specified typeface, providing text-specific references for typeface-fidelity evaluation.

Each sample records the complete prompt, language and design attributes, full target text and its length, indexed semantic segments, and per-segment typeface families and reference images. These linked annotations support image-level text evaluation, segment-level coverage analysis, typeface comparison, and analyses by language and design attribute.

3.3 Human Review and Quality Control

Following the construction of candidate samples, we invited 10 reviewers with professional design backgrounds to conduct quality checks. The review covered five criteria: (i) **Prompt consistency:** consistency between generation prompts and target text, and the appropriateness of the design requirements; (ii) **Text segmentation:** completeness and coherence of the target-text segmentation; (iii) **Typeface identification:** correct correspondence between font names and typeface families; (iv) **Character support:** font support for the target languages and characters; (v) **Reference com-**



(a) Text attributes

(b) Design attributes

Figure 3: (a) Text Attribute distributions. English (top, $n = 300$) and Chinese (bottom, $n = 300$) target-text lengths, segment counts, and the most frequent typeface families. (b) Design attributes. The three equal inner sectors identify annotation fields. Within each field, the outer ring shows the 15 most frequent categories and aggregates the remainder as *Others*.

pleteness: completeness of both typeface-family reference images and segment-level rendered references.

Each sample underwent verification and revision by at least three reviewers, and only samples that passed this review process were retained. Through screening and revision, we retained 600 samples from approximately 2,000 candidates to form the final evaluation dataset.

3.4 Dataset Statistics

The evaluation set contains 600 prompts (300 in English and 300 in Chinese) with 3,695 annotated text segments and 78 distinct typeface families after cross-language deduplication. It covers 24 design categories, 20 industries, and 20 visual styles. Figure 3 presents their text and design attributes.

Text attributes. Figure 3a summarizes three text attributes: target-text length, the number of text segments, and typeface-family distribution. Text length is measured in words for English and characters for Chinese, ranging from 46 to 428 words and 93 to 1,584 characters, respectively. Segment count measures the number of annotated semantic units per prompt, averaging 5.66 for English and 6.68 for Chinese, with ranges of 3–14 and 2–15 segments, respectively. For typefaces, the English and Chinese sets cover 71 and 15 families, respectively, with eight shared across languages. The figure displays the ten most frequent families in each language.

Design attributes. Figure 3b shows the distributions of design types, industries, and visual styles. Web banner ads, event invitations, product detail pages, and wedding invitations are the most frequent design types. The most represented industries are pets, photography, and music and entertainment, while glassmorphism, Memphis, and photorealism are the most common visual styles.

3.5 Eval Protocol

3.5.1 Eval pipeline

ExtremeTextBench employs a shared pipeline for image preprocessing, text detection and recognition, and alignment with target annotations to support both long-text rendering and typeface-fidelity evaluation. Each model receives the complete generation prompt for a sample and produces an image with fixed aspect ratio or resolution. An OCR system then extracts text, bounding boxes, polygon coordinates, and recognition confidence scores.

The recognized text is compared with the full target text for image-level content evaluation, while detected text regions are aligned with target semantic segments for local analysis. The geometry of the detected regions is used to estimate rendered text size and analyze performance across size

ranges. Regions satisfying the text-matching and size criteria are cropped and paired with the corresponding typeface-rendered references. These pairs are provided to a multimodal judge to assess whether the generated text adheres to the requested typeface family. The two evaluation tracks share generated images, OCR outputs, and target annotations, while reporting text-content and typeface-fidelity results separately.

3.5.2 LONG-TEXT RENDERING METRICS

For each generated image, we concatenate the recognized text from all OCR boxes and compare it with the complete target text. English box texts are joined with spaces and tokenized by whitespace, whereas Chinese box texts are concatenated directly and evaluated character by character. Let G_i and O_i denote the multisets of target and recognized text units for image i, respectively, and let $M_i = |G_i \cap O_i|$ be the multiset intersection size. Each distinct unit contributes the smaller of its occurrence counts in the target and recognized text. We define the image-level metrics as

$$\text{TextRecall}_i = \frac{M_i}{|G_i|}, \qquad \text{TextPrecision}_i = \frac{M_i}{|O_i|}, \qquad \text{TextF1}_i = \frac{2M_i}{|G_i| + |O_i|}. \tag{1}$$

Text Recall measures the proportion of requested units recovered by OCR, while Text Precision measures the proportion of recognized units supported by the target text, penalizing excess occurrences and extraneous content. Text F1 is their harmonic mean (Manning et al., 2008), accounting for both missing and additional text. For dataset-level reporting within each language, we sum the per-image matched, target, and recognized unit counts and apply the same formulas to these totals, yielding micro-averaged scores. Results are reported in Table 2 as **R**,**P**,**F1**.

3.5.3 TEXT-SIZE-AWARE EVALUATION

We estimate rendered text size using the shortest nonzero edge of each OCR polygon in the evaluation image, with the shorter bounding-box side as a fallback. Each region is aligned with a local target-text span using bounded longest common subsequence (LCS) (Cormen et al., 2022) matching against the annotated segments, as detailed in AppendixA.3.

Within each size bin, we sum the matched, target, and recognized unit counts across regions and compute micro-averaged Text Recall, Precision, and F1 using the multiset-based definitions in Section3.5.2. Results are reported separately for each model and language. This analysis measures local text quality at observed sizes, while undetected text is assessed through the image-level metrics.

3.5.4 TYPEFACE FIDELITY EVALUATION

Typeface fidelity evaluation reuses the OCR regions and bounded LCS matching procedure introduced in Section 3.5.3. By aligning each region with an annotated target-text segment, we identify its requested typeface family and retrieve the corresponding pre-rendered reference image of that segment. The region is then cropped from the generated image and paired with this reference for visual comparison. We retain regions with an LCS-based matching score of at least 0.5 and a text size of at least 50 pixels, to provide a reliable reference association and sufficient glyph detail.

Given the generated crop and its reference, GPT-5.5 (OpenAI, 2026b) assesses how faithfully the generated text reproduces the requested typeface. The comparison focuses on glyph structure, proportions, stroke terminals, serifs, and distinctive design details, rather than text content, color, size, layout, or background. Reasonable variations within the same typeface family and minor rendering distortions are tolerated. The judge returns an integer similarity score from 0 to 3, representing negligible, slight, moderate, and high similarity, respectively. A score of 3 indicates strong agreement in the defining glyph features, consistent with the same typeface family; sharing only a broad category such as serif or sans serif is insufficient for a high score.

$$\text{FontFidelity} = \frac{1}{3|\mathcal{E}|} \sum_{c \in \mathcal{E}} s_c, \qquad s_c \in \{0, 1, 2, 3\}. \tag{2}$$

For each model and language, we report this normalized mean on a 0–100 scale as **Font** in Table 2. Let $\mathcal{E}$ denote the sampled crops retained and s_c the score of crop c. We compute typeface fidelity as Equation 2

Table 2: Main results for long-text rendering and typeface fidelity. Recall (R), Precision (P), and F1 are percentages. For F1 and Font, $\pm$ denotes the half-width of the approximate 95% confidence interval in percentage points. Higher is better for all scores. Rows are ordered by English Text F1. Bold and underlined scores indicate the best and second-best values within each language and metric.

	English				Chinese			
Model	R ↑	P ↑	F1 ↑ (95% CI)	Font ↑ (95% CI)	R ↑	P ↑	F1 ↑ (95% CI)	Font ↑ (95% CI)
Qwen-Image 3.0 Pro	<u>91.52</u>	**92.87**	**92.19** $\pm$ 1.28	57.67 $\pm$ 4.19	<u>93.56</u>	**94.24**	**93.90** $\pm$ 0.82	62.12 $\pm$ 3.78
GPT-Image 2	**92.51**	88.91	<u>90.67</u> $\pm$ 1.22	<u>66.36</u> $\pm$ 4.23	**94.61**	91.77	<u>93.17</u> $\pm$ 0.82	<u>63.70</u> $\pm$ 3.20
GPT-Image 1.5	91.22	89.90	90.55 $\pm$ 1.38	65.82 $\pm$ 4.16	91.22	89.85	90.53 $\pm$ 1.06	63.04 $\pm$ 3.22
Nano Banana 2.0	86.15	<u>90.32</u>	88.18 $\pm$ 1.88	**77.31** $\pm$ 5.13	87.06	87.31	87.18 $\pm$ 1.47	**65.89** $\pm$ 4.09
Seedream 4.5	80.65	80.16	80.40 $\pm$ 1.84	46.02 $\pm$ 5.70	88.03	83.54	85.72 $\pm$ 1.14	53.87 $\pm$ 4.24
Wan2.7-Image	74.46	73.91	74.19 $\pm$ 2.18	54.73 $\pm$ 5.20	88.98	85.07	86.98 $\pm$ 1.01	60.11 $\pm$ 4.81
ERNIE-Image	70.56	75.36	72.88 $\pm$ 1.83	47.11 $\pm$ 4.88	83.92	89.64	86.68 $\pm$ 1.16	55.67 $\pm$ 4.29
Qwen-Image 2.0 Pro	68.33	77.41	72.59 $\pm$ 3.02	53.31 $\pm$ 5.81	90.22	<u>93.14</u>	91.66 $\pm$ 1.11	56.24 $\pm$ 4.90
Seedream 5.0 Pro	67.95	72.52	70.16 $\pm$ 2.95	59.03 $\pm$ 4.71	71.40	76.44	73.84 $\pm$ 2.15	57.14 $\pm$ 4.33
Qwen-Image 2.0	59.68	73.12	65.72 $\pm$ 3.41	52.55 $\pm$ 6.55	79.39	88.06	83.50 $\pm$ 1.88	56.57 $\pm$ 4.52
FLUX.2 Dev	55.54	66.60	60.57 $\pm$ 2.09	58.19 $\pm$ 6.58	25.23	60.75	35.65 $\pm$ 1.62	51.52 $\pm$ 4.31
Z-Image-Turbo	57.43	60.17	58.77 $\pm$ 2.34	44.61 $\pm$ 4.99	56.06	83.22	66.99 $\pm$ 1.96	44.27 $\pm$ 3.78
HiDream-O1-Image	48.80	63.79	55.30 $\pm$ 2.29	49.29 $\pm$ 4.66	69.99	76.76	73.22 $\pm$ 1.51	46.10 $\pm$ 3.75
GLM-Image	38.59	58.18	46.40 $\pm$ 1.63	40.95 $\pm$ 5.14	38.98	70.03	50.08 $\pm$ 1.97	44.83 $\pm$ 4.74
FLUX.2 Klein 9B	29.45	35.96	32.38 $\pm$ 1.59	49.00 $\pm$ 5.99	24.04	37.58	29.32 $\pm$ 1.01	41.30 $\pm$ 4.55
FLUX.2 Klein 4B	16.33	20.67	18.24 $\pm$ 0.86	46.30 $\pm$ 6.19	19.24	28.74	23.05 $\pm$ 0.94	40.60 $\pm$ 6.80

4 Evaluation Results

4.1 Experimental Setup

We evaluate 16 leading T2I models: ERNIE-Image (Liu et al., 2026), FLUX.2 Dev (Black Forest Labs, 2026), FLUX.2 Klein 4B (Black Forest Labs, 2026), FLUX.2 Klein 9B (Black Forest Labs, 2026), Gemini 3.1 (NanoBanana, 2026), GLM-Image (Z.ai, 2026), GPT-Image 1.5 (OpenAI, 2026a), GPT-Image 2 (OpenAI, 2026a), HiDream-O1-Image (Cai et al., 2026), Qwen-Image 2.0 (Zhao et al., 2026), Qwen-Image 2.0 Pro (Zhao et al., 2026), Qwen-Image 3.0 Pro (Alibaba Qwen Team, 2026), Seedream 4.5 (Seedream et al., 2025), Seedream 5.0 Pro (ByteDance Seed, 2026), Wan2.7-Image (Alibaba, 2026), and Z-Image-Turbo (Cai et al., 2025). All generated images are evaluated at a common input resolution of 1024×1024 pixels. Since the evaluated versions of Seedream 4.5 and HiDream-O1-Image do not support generation at this resolution, we generate their images at the supported resolution of 2048×2048 pixels and downsample them to 1024×1024 pixels before evaluation. For OCR-based text rendering metric and text size aware evaluation, we use PaddleOCR 3.7.0 (Zhang et al., 2026) with the PP-OCRv6 (Zhang et al., 2026) medium detector (`PP-OCRv6_medium_det`) and recognizer (`PP-OCRv6_medium_rec`), running on PaddlePaddle GPU 3.3.0 (Ma et al., 2019). For Text F1 and Font, we perform 20,000 prompt-level bootstrap replicates (Efron & Tibshirani, 1994) for each model and language, resampling prompts with replacement and recomputing micro-averaged F1 and crop-weighted mean typeface fidelity, respectively. We report approximate 95% confidence intervals as the point estimate $\pm 1.96\,\mathrm{SE}_{\mathrm{boot}}$, where $\mathrm{SE}_{\mathrm{boot}}$ is the sample standard deviation of the bootstrap estimates.

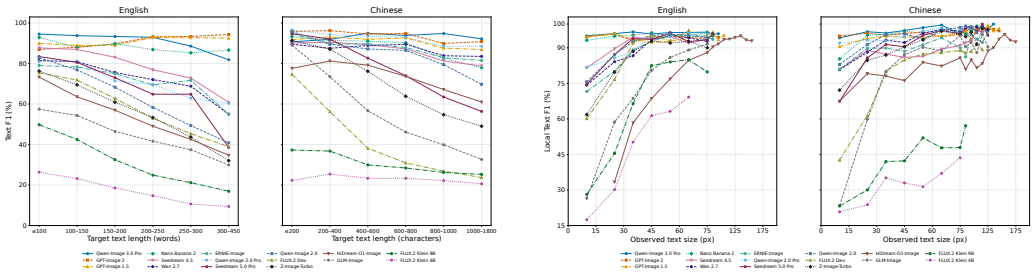
4.2 Main Results

Text-rich design remains challenging for many evaluated models (Table 2). 11 of the 16 models score below 80 in English Text F1, while the highest point estimates reach 92.19 in English and 93.90 in Chinese.

Limited coverage of the requested content is a recurring weakness. In each language, 12 of the 16 models have higher Precision than Recall. For example, Z-Image-Turbo achieves a Chinese Precision of 83.22 but a Recall of 56.06. This gap shows that relatively accurate recognized text can coexist with incomplete recovery of the requested content, highlighting the importance of assessing completeness alongside correctness.

Table 3: Scores and dispersion on common-model sets. External scores are CVTG-2K Word Accuracy from the GLM-Image report (Z.ai, 2026) and LongText-Bench scores from Table 3 of the Boogu-Image report (Chen et al., 2026); Ours is ExtremeTextBench Text F1.

	CVTG-2K English ($n = 4$)		LongText-Bench English ($n = 7$)		LongText-Bench Chinese ($n = 7$)	
Model	External	Ours	External	Ours	External	Ours
GPT-Image 2	–	–	96.00	90.67	96.10	93.17
Nano Banana 2.0	77.88	88.18	98.10	88.18	94.90	87.18
Seedream 4.5	89.90	80.40	98.90	80.40	98.70	85.72
ERNIE-Image	–	–	98.00	72.88	96.60	86.68
Z-Image-Turbo	85.85	58.77	91.70	58.77	92.60	66.99
HiDream-O1-Image	–	–	97.90	55.30	97.80	73.22
GLM-Image	91.16	46.40	95.20	46.40	97.90	50.08
Score range	13.28	41.78	7.20	44.27	6.10	43.09
Standard deviation	5.19	16.67	2.31	15.91	1.94	14.00
Mean absolute pairwise gap	7.32	24.50	2.80	20.97	2.48	17.44



(a) (b)

Figure 4: Text rendering performance for 16 models: (a) Text F1 by target length; (b) local Text F1 by observed text size. Each subfigure shows English (left) and Chinese (right).

Strong text rendering does not ensure adherence to the requested typefaces. Qwen-Image 3.0 Pro has the highest Text F1 point estimates in both languages, but its English Font score is 57.67, compared with 77.31 for Nano Banana 2.0. Even among eligible regions with sufficient glyph detail, the median English Font score across models is only 52.93. These results motivate evaluating textual correctness and typeface fidelity separately.

4.3 Main Findings

Score Spread Across Benchmarks. ExtremeTextBench reveals larger score differences between models that are closely clustered on existing benchmarks (Table 3). For example, HiDream-O1-Image and ERNIE-Image differ by only 0.10 points on LongText-Bench EN but by 17.58 points on ExtremeTextBench EN. Across the common-model comparisons with CVTG-2K and LongText-Bench, score ranges, standard deviations, and mean pairwise gaps are all larger. These results provide descriptive evidence of score separation under different metrics and evaluation protocols.

Rendering Quality Across Text Lengths and Sizes. Most models show lower full-image Text F1 on longer target-text subsets (Figure 4a), with median decreases of 26.74 and 12.58 percentage points from the shortest to the longest bins in EN and ZH, respectively. Stronger models nevertheless retain high scores under demanding conditions: GPT-Image 2 achieves 94.35 Text F1 on English prompts exceeding 300 words, while Qwen-Image 3.0 Pro achieves 95.04 local Text F1 on English text of at ≤ 20 pixels (Figure 4b).

Typeface Fidelity Across Font Families. Most models reproduce Helvetica and Heiti SC more faithfully than Montserrat and STXingkai (Figure 5). Among models with adequate samples for

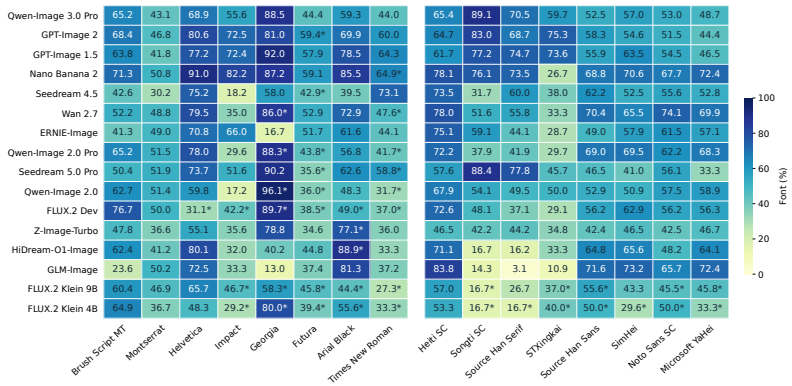


Figure 5: Typeface fidelity by font family for English (left) and Chinese (right). Rows denote models; columns show font families. Cell values and colors indicate normalized mean font similarity. An asterisk (∗) marks a model–font pair with fewer than 20 credible evaluation samples.

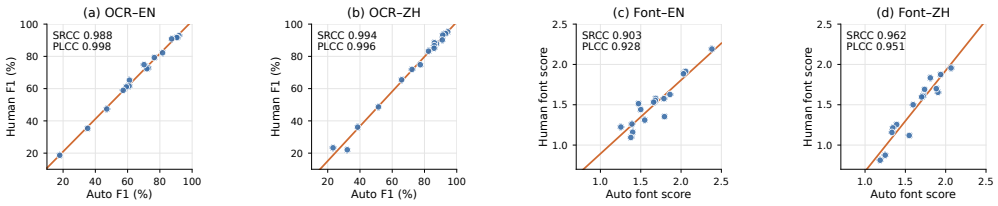


Figure 6: Model-level agreement between automatic and human evaluation for English and Chinese OCR-based text rendering and typeface fidelity. Each point represents one evaluated model.

both families, Helvetica leads Montserrat by a median of 25.78 points. Simpler glyphs and greater pretraining exposure may explain these differences. Stronger Font models can offer broader, more balanced fidelity: across the eight Chinese families, GPT-Image 2 scores 44.4–83.0 versus GLM-Image's 3.1–83.8, achieving a much higher minimum despite similar peaks.

4.4 Evaluator validation

We conduct human validation with 10 annotators. Each item requires at least three valid annotations from distinct annotators, and an answer is accepted as the human reference only when a majority of the valid annotations agree. Items without majority agreement undergo a second round of review by multiple annotators, none of whom participated in the first round for that item. Annotation interfaces as AppendixA.5

Figure 6 shows strong agreement between automatic and human scores at the model level. For text rendering, Spearman/Pearson correlations (SRCC/PLCC) are 0.988/0.998 in English and 0.994/0.996 in Chinese; for typeface fidelity, they are 0.903/0.928 and 0.962/0.951.

5 Conclusion

We introduce **ExtremeTextBench**, a bilingual benchmark filling a gap in jointly evaluating long-text rendering, small-text rendering, and adherence to specified typefaces in realistic design scenarios. Experiments on 16 models reveal wider score separation among shared models than reported on CVTG-2K and LongText-Bench. Under extreme rendering demands, the benchmark enables a more precise assessment of model capability boundaries across typeface fidelity, text length, and text size. We hope ExtremeTextBench will guide future research toward reliable text-rich image generation that satisfies both content and typography requirements in practical design workflows.

AI USE STATEMENT

In this work, we used generative AI tools for paper polishing and translation. We have not used generative AI tools for developing the research idea, designing the experiments, literature review, and drafting the main body of the paper. Additionally, we used ChatGPT for the above tasks. We have reviewed all AI-assisted work. Specifically, the AI-assisted text was manually reviewed by the authors to verify its correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the AI.

REFERENCES

Alibaba. wan2.7-image. https://help.aliyun.com/en/model-studio/wan2-7-image, 2026. Multimodal image generation and editing system.

Alibaba Qwen Team. Qwen-image-3.0-pro. Qianwen AI Platform, 2026. URL https://platform.qianwenai.com/docs/developer-guides/getting-started/image-models.md. Accessed: 2026-09-26.

Black Forest Labs. Flux.2. https://huggingface.co/black-forest-labs/FLUX.2-dev, 2026. Multimodal image generation and editing system.

ByteDance Seed. Bytedance seed. https://seed.bytedance.com/en/blog/beyond-generation-it-understands-design-introducing-seedream-5-0-pro, 2026. Multimodal image generation and editing system.

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.

Qi Cai, Jingwen Chen, Chengmin Gao, Zijian Gong, Yehao Li, Yingwei Pan, Yi Peng, Zhaofan Qiu, Kai Yu, Yiheng Zhang, et al. Hidream-o1-image: A natively unified image generative foundation model with pixel-level unified transformer. *arXiv preprint arXiv:2605.11061*, 2026.

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.

Guang Chen, Jianchao Yang, Hailin Jin, Jonathan Brandt, Eli Shechtman, Aseem Agarwala, and Tony X Han. Large-scale visual font recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3598–3605, 2014.

Guoxuan Chen, Chufeng Xiao, Haoran Yang, Siyue Xie, Binxiao Huang, Ming Zhang, Cheuk Him Chau, Xinyu Fu, Yingzhao Lian, Tom SY Li, et al. Boogu-image-0.1: Boosting open-source unified multimodal understanding and generation. *arXiv preprint arXiv:2607.13125*, 2026.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems*, 36:9353–9387, 2023.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, 4 edition, 2022. ISBN 9780262046305. URL https://mitpress.mit.edu/9780262046305/introduction-to-algorithms/.

Adrienne Deganutti, Elad Hirsch, Haonan Zhu, Jaejung Seol, and Purvanshi Mehta. Graphic-design-bench: A comprehensive benchmark for evaluating ai on graphic design tasks. *arXiv preprint arXiv:2604.04192*, 2026.

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.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, et al. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025.

Yan Li, Zezi Zeng, Ziwei Zhou, Xin Gao, Muzhao Tian, Yifan Yang, Mingxi Cheng, Qi Dai, Yuqing Yang, Lili Qiu, Zhendong Wang, Zhengyuan Yang, Xue Yang, Lijuan Wang, Ji Li, and Chong Luo. Bizgeneval: A systematic benchmark for commercial visual content generation, 2026. URL https://arxiv.org/abs/2603.25732.

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?, 2024. URL https://arxiv.org/abs/2412.11767.

Qian Liang, Xiaomin Li, Ying Zhang, Jia Xu, Lihao Ni, Hongrui Li, Jingjing Li, Jing Lyu, and Chen Li. Wegenbench: A multidimensional diagnostic benchmark towards text-to-image model optimization. *arXiv preprint arXiv:2606.20100*, 2026.

Jiaxiang Liu, Zhida Feng, Pengyu Zou, Zhenyu Qian, Tianrui Zhu, Jun Xia, Yuehu Dong, Yanzheng Lin, Honglin Xiong, Anqi Chen, et al. Ernie-image technical report. *arXiv preprint arXiv:2605.25347*, 2026.

Yanjun Ma, Dianhai Yu, Tian Wu, and Haifeng Wang. Paddlepaddle: An open-source deep learning platform from industrial practice. *Frontiers of Data and Domputing*, 1(1):105–115, 2019.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. URL https://nlp.stanford.edu/IR-book/.

NanoBanana. Nanobanana. https://nanobananaimg.com/, 2026. Multimodal image generation and editing system.

Peter O'Donovan, Jānis Lībeks, Aseem Agarwala, and Aaron Hertzmann. Exploratory font selection using crowdsourced attributes. *ACM transactions on graphics (TOG)*, 33(4):1–9, 2014.

OpenAI. Openai. https://openai.com/, 2026a. Multimodal image generation and editing system.

OpenAI. GPT-5.5 system card. https://openai.com/index/gpt-5-5-system-card/, 2026b.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. In *International Conference on Learning Representations*, volume 2024, pp. 56783–56799, 2024.

Zhangyang Wang, Jianchao Yang, Hailin Jin, Eli Shechtman, Aseem Agarwala, Jonathan Brandt, and Thomas S Huang. Deepfont: Identify your font from an image. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 451–459, 2015.

Z.ai. Glm-image. https://z.ai/blog/glm-image, 2026. Multimodal image generation and editing system.

Tianyu Zhang, Xinyu Wang, Lu Li, Zhenghan Tai, Jijun Chi, Jingrui Tian, Hailin He, and Suyuchen Wang. Strict: Stress-test of rendering image containing text. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21148–21161, 2025.

Yubo Zhang, Xueqing Wang, Manhui Lin, Yue Zhang, Penglongyi Deng, Ting Sun, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Changda Zhou, et al. Pp-ocrv6: From 1.5 m to 34.5 m parameters, surpassing billion-scale vlms on ocr tasks. *arXiv preprint arXiv:2606.13108*, 2026.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, et al. Qwen-image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026.

Shitian Zhao, Qilong Wu, Xinyue Li, Bo Zhang, Ming Li, Qi Qin, Dongyang Liu, Kaipeng Zhang, Hongsheng Li, Yu Qiao, et al. Lex-art: Rethinking text generation via scalable high-quality data synthesis. *arXiv preprint arXiv:2503.21749*, 2025.