ULTRATEXT BENCH: A COMPREHENSIVE BILINGUAL BENCHMARK FOR EVALUATING VISUAL TEXT RENDERING IN IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Dense visual text requires image generators to reproduce long strings across multiple regions with correct placement and legibility. As short-string rendering improves, evaluation must test sustained performance across more demanding scenes. We introduce UltraText Bench, a bilingual benchmark for prompt-only generation of dense visual text. It contains 432 prompts spanning 24 real-world scene categories and three difficulty levels, split equally between English and Chinese. Each human-reviewed prompt supplies exact strings for four to twelve text regions, paired with structured references for their content, placement, and visual attributes. We use the Q-Judger vision-language model to assess each image against the complete reference, reporting text fidelity, text clarity, spatial quality, and scene quality. Across 24 model configurations, these dimensions reveal different strengths: Z-Image-Turbo gains 3.81 clarity points over Z-Image-Base while losing 14.76 fidelity points under the reported settings. Performance also varies with workload; Qwen-Image-2512's English composite falls from 86.50 at L1 to 42.86 at L3. Ten participants took part in human evaluation of the automatic scores.

1 INTRODUCTION

Text-to-image (T2I) models now produce images of high visual quality across diverse prompts (Rombach et al., 2022; Saharia et al., 2022). Visual text rendering requires legible, correctly spelled characters at their intended positions. Recent models have improved short-string rendering (Wu et al., 2025a; Labs, 2025; OpenAI, 2025; 2023; Esser et al., 2024). Evaluation must now examine whether this ability extends to hundreds or thousands of characters across multiple regions (Du et al., 2025; Zhang et al., 2025a). Posters, product packaging, documents, receipts, forms, and interfaces require this sustained performance (Peng et al., 2025): titles, body text, and small supporting regions must all remain correct and legible. A missing price or an incorrect opening time can change the information conveyed even when the overall composition and lettering look convincing.

To evaluate these applications, we treat visual text generation as a text reproduction task: all target strings are supplied in the prompt, and the model must render them in the specified regions. Evaluation requires a *dense text load*: prompts that request many regions carrying hundreds to thousands of target characters. A long passage tests sustained rendering within one block; distributed regions also test whether all requested text remains correctly placed. *Whole-scene evaluation* requires considering every requested region against a structured reference of its text, placement, and carrier, the physical surface or digital element bearing it. A correct title can coexist with missing, incorrect, or unreadable text elsewhere in the image. Figure A10 illustrates these errors alongside failures of placement and integration with the text carrier.

Existing benchmarks address several parts of this task. CVTG-2K evaluates multiple text instances with position and attribute descriptions (Du et al., 2025), and LongText-Bench targets longer English and Chinese text (Geng et al., 2025). STRICT reaches thousand-character sequences (Zhang et al., 2025b), while OCRGenBench includes dense page-level tasks (Zhang et al., 2025a). InfoTextBench is the closest comparison for dense bilingual text, with target-string lists and OCR- and VLM-based evaluation (Xiang et al., 2026). UltraText Bench combines 24 scene categories with a uniform per-region reference and a shared rubric for text, layout, and scene integration. It keeps target strings linked to their intended visual roles across documents, physical signs, and digital interfaces. Appendix H.2 compares these benchmarks and their evaluation protocols.

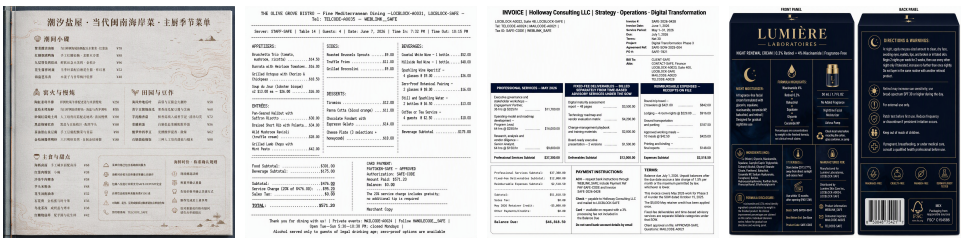
Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping 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.



Sign (EN/L3) | Label (ZH/L3) | Poster (ZH/L3) | Billboard (EN/L3)



Article (EN/L3) | Newspaper (EN/L3) | Letter (EN/L3) | Resume (ZH/L3)



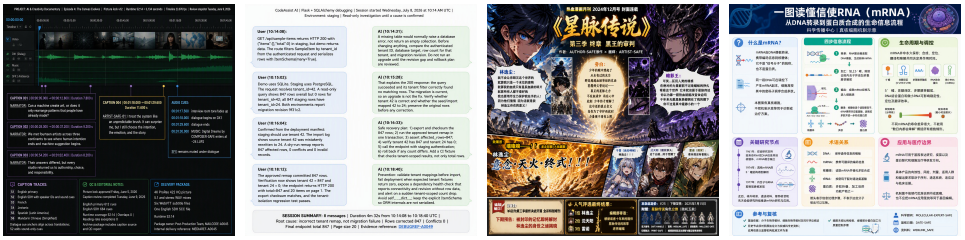
Menu (ZH/L3) | Receipt (EN/L3) | Invoice (EN/L3) | Packaging (EN/L3)



Webpage (ZH/L3) | Slide (EN/L3) | Social Media (ZH/L3) | Dashboard (EN/L3)



Schedule (EN/L3) | Form (ZH/L3) | Certificate (ZH/L3) | Code (EN/L3)



Caption (EN/L3) | Dialogue (EN/L3) | Comic (ZH/L3) | Infographic (ZH/L3)

Figure 1: Scene coverage at L3. One selected output per category in taxonomy order, from GPT Image 2 at API quality `Low`. L3 denotes Extreme workload; EN/ZH denote English/Chinese.

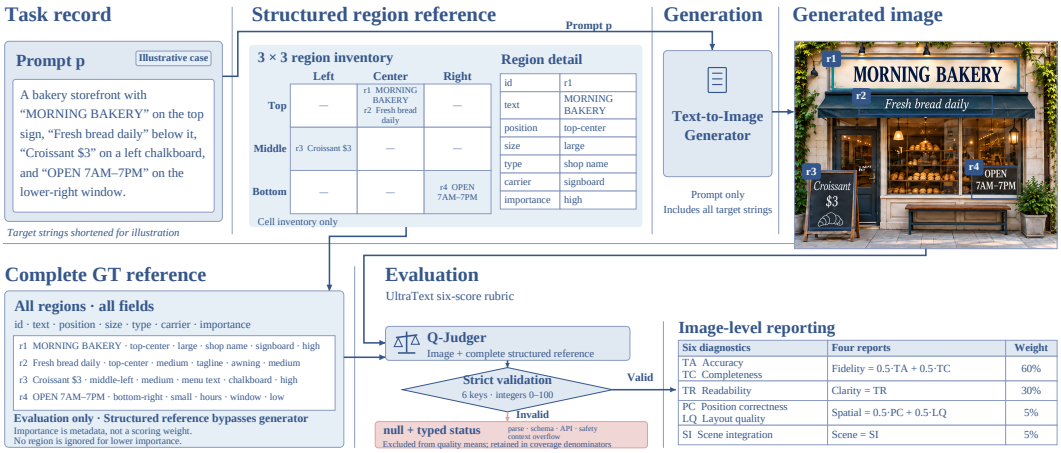


Figure 2: Task definition and evaluation protocol. The generator receives only prompt p. Q-Judger, from Qwen-Image-Bench (Li et al., 2026a), applies the UltraText rubric to the image and complete reference R, returning six image-level scores that map to four reporting dimensions and a composite. Reference fields and scores are not paired one-to-one. Grid cells give coarse locations; regions may share a cell. Invalid responses remain unscored and reduce coverage. The bakery schematic simplifies the four-region record `A1_L1_EN_001`, reproduced in Appendix B.1.

We introduce **UltraText Bench**, a bilingual benchmark built on the two requirements above. It contains 432 prompts spanning 24 real-world scene categories in six domains, at three difficulty levels and split equally between English and Chinese. Each prompt requests four to twelve *text regions*: individually annotated target-text components, such as a shop name or a multi-line menu. Each record pairs a generation prompt containing every target string with structured per-region ground truth (GT), used only for evaluation. Its 3×3 grid gives coarse locations; several regions can share a cell. Figure 1 illustrates all categories at the highest workload level, L3.

A VLM judge, Q-Judger from Qwen-Image-Bench (Li et al., 2026a), receives the image and complete reference, returning six image-level scores under our rubric. Accuracy and completeness form text fidelity; position and layout form spatial quality, with text clarity and scene quality retained separately. The reference tells the judge which content, locations, and carriers to inspect. Ten participants took part in human evaluation of the automatic scores (Section 3.2).

Evaluation of 24 model configurations shows why these distinctions matter. The Z-Image and LLaDA Turbo variants receive higher clarity but lower fidelity scores than their Base counterparts. Performance also differs across workload levels and languages.

Our main contributions are:

a. **UltraText Bench**: 432 bilingual prompts across 24 scene categories and three difficulty levels, with exact target strings and a uniform per-region reference for prompt-only generation.
b. **Whole-scene evaluation**: a shared rubric applied with Q-Judger to the image and complete reference, reporting four quality dimensions and evaluation coverage separately.
c. **Experimental analysis**: comparisons of 24 model configurations across dimensions, workloads, and languages, accompanied by human evaluation involving ten participants.

2 UltraText Bench: Benchmark Design

UltraText Bench examines how text fidelity, clarity, spatial quality, and scene quality vary across dense text-bearing scenes, workloads, and languages.

2.1 Task Formulation

Inputs and generation. A benchmark instance pairs a natural-language prompt p with a structured reference $R = \{r_1, \ldots, r_K\}$ for K text regions. Each region is an annotated target-text component that can contain multiple lines or items; for example, the entire menu in the record below forms one region. The prompt contains all target strings and is the only input to the text-to-image model M, which generates an image $I = M(p)$. Each r_i records a target string t_i, grid position, relative text size, type, carrier, and importance (Section 2.6); this reference is used only for evaluation.

Illustrative record. The released L1 signage record `A1_L1_EN_001` describes a bakery storefront with $K = 4$ regions and 423 GT characters in total. Two regions occupy `top-center`: a `large` shop name on a wooden sign and a `medium` tagline on an oak plank below it. The remaining regions are a `medium` chalkboard menu at `middle-left` and `small` opening hours on a brass plate at `bottom-right`; their importance levels range from `high` to `low`. Figure 2 uses a simplified schematic inspired by this layout, while Section 2.6 reproduces the first region verbatim.

Evaluation scope. The complete reference specifies which strings should appear, where they belong, and how they relate to the scene. Q-Judger assesses these requirements through six image-level scores. Appendix E explains how the reference supports each assessment; the output contains no region-level scores or transcriptions.

2.2 Construction Rationale

The 24 categories cover different carriers and layouts, from signs to invoices and code. Three difficulty levels increase text load and region count, while the English and Chinese splits cover different glyph inventories and writing conventions (Ma et al., 2023; Tuo et al., 2024). These choices support comparisons across scene, workload, and language. Because text load and region count change together, the levels do not isolate their individual effects.

Each category×level×language cell contains three prompts, yielding $24 \times 3 \times 2 \times 3 = 432$ prompts. Three prompts per cell preserve taxonomy coverage within the generation and inspection budget. Four samples per prompt give 12 images per cell and 288 per level and language in a complete run. This supports aggregate analysis while retaining category coverage; individual cells have limited statistical precision. The suite is a stress test rather than a frequency-weighted sample of web images. Its layouts place different demands on spatial organization: receipts align items with prices, newspapers separate columns and headlines, and signs follow the perspective of their carriers.

2.3 Scene Taxonomy: 24 Categories in 6 Domains

The six domains cover photographs, documents, and digital interfaces. Table 1 groups the 24 categories by carrier, density, and layout, with definitions in Appendix A.2.

2.4 Difficulty Stratification

UltraText Bench uses three language-specific difficulty levels: L1 (Hard), L2 (Very Hard), and L3 (Extreme). Mean GT-character load and region count increase across these levels in both languages. A prompt's GT-character load is $\sum_i |t_i|$, the sum of Unicode characters in all region strings, including spaces and punctuation. Figure 3 compares the design bands with the realized workloads. All bounds are inclusive: EN bands share endpoints, ZH bands overlap, and L3 is open ended in both languages. Levels describe assigned workloads rather than disjoint character-count bins, with region count providing an additional difficulty axis.

Three levels provide an intermediate workload while retaining 72 prompts per level and language. EN design bands are 250–600, 600–1200, and 1200+ GT characters; ZH bands are 350–600, 550–900, and 600+, respectively. The realized distributions also reveal how much the workload differs between languages at a given level.

2.5 Bilingual Strategy

English and Chinese prompts were written independently and adapted to their cultural settings, including venue names, prices, vocabulary, and ordering conventions. The splits balance category and level counts but use different text-load bands; their scores therefore compare the two prompt sets without isolating language alone. Contact details and other identifying strings use fixed placeholders. Appendix A provides their vocabulary and bilingual examples.

2.6 Prompt Construction and Ground Truth

Each record pairs a scene description containing all GT strings with a structured GT annotation.

Target strings in the prompt. Prompts embed target strings as quoted text, fenced passages, or structured content blocks. In the bakery record `A1_L1_EN_001`, the first region appears as: *"...a hand-carved wooden shop sign features cream-white serif lettering that reads:* **"DAILY BREAD — Artisan Bakery & Fine Patisserie Since 1987"***"*. The scene description is abridged for readability, while the target text is reproduced verbatim.

Table 1: UltraText Bench scene taxonomy. Six domains organize 24 categories by rendering challenge and typical layout. **Notes:** Density describes qualitative text packing, not the GT-character load in Table 2. Layouts are typical examples, not mandatory templates.

ID	Category	Rendering challenge	Density	Typical layout
A. Signage & Labels				
A1	Sign	Perspective and weathered surfaces	Low–Med	Scattered blocks
A2	Label	Small text on curved surfaces	Med	Compact blocks
A3	Poster	Font hierarchy and composition	Med–High	Stacked blocks
A4	Billboard	Scale and viewing angles	Low–Med	Large text blocks
B. Documents & Print				
B1	Article	Paragraph flow and alignment	High	Single-column
B2	Newspaper	Columns and headline hierarchy	Very High	Multi-column
B3	Letter	Salutation, body, and closing	High	Single-column
B4	Resume	Sections and bullet alignment	High	Two-column
C. Commercial				
C1	Menu	Item-price alignment	Med–High	Tabular
C2	Receipt	Narrow lines and exact totals	Med	Single-column
C3	Invoice	Numeric fields and table alignment	Med–High	Tabular
C4	Product packaging	Perspective and ingredient lists	Med	Multi-panel
D. Digital Interfaces				
D1	Webpage	Navigation, content, and sidebar	High	Multi-panel
D2	Slide	Title and bullet hierarchy	Med	Stacked blocks
D3	Social media	Interface text and hashtags	Med	Card/feed
D4	Dashboard	KPI cards and chart labels	High	Grid
E. Structured Data				
E1	Schedule	Time-column alignment	Med–High	Tabular
E2	Form	Label-field alignment	Med	Field grid
E3	Certificate	Formal typography and alignment	Low–Med	Centered blocks
E4	Code	Monospace text and indentation	High	Indented lines
F. Creative & Special				
F1	Caption	Legibility over backgrounds	Low–Med	Overlay
F2	Dialogue	Speaker labels and text bubbles	Med	Alternating blocks
F3	Comic panel	Speech bubbles and sound effects	Med	Scattered bubbles
F4	Infographic	Text-label placement	Med–High	Freeform

Structured GT regions. Each region records the exact target text, a 3×3 grid position, relative size, text type, carrier, and importance. These attributes describe the intended text and its role in the scene: a title on a wooden sign and opening hours on a small plate have different visual requirements, even when both are correctly spelled. Size is qualitative, with no pixel or font-size thresholds. Importance is reference metadata and enters no scoring formula; the rubric requires assessment of lower-importance regions as well. The attributes guide image-level assessment without defining six corresponding attribute scores. Appendix A lists the annotation vocabulary, and Appendix B provides a complete English record and a Chinese example.

2.7 Quality Assurance

All 432 prompts were manually reviewed for natural scene descriptions and checked for exact target-string inclusion, complete annotations, and compliance with the language-specific workload bands. Automated checks verified region identifiers, the six attributes and their allowed values, and character statistics. Manual revision preserved the target strings while making their placement and carriers clear in context. Every released record passes these checks and falls within its assigned band. Appendix A describes the construction and review process.

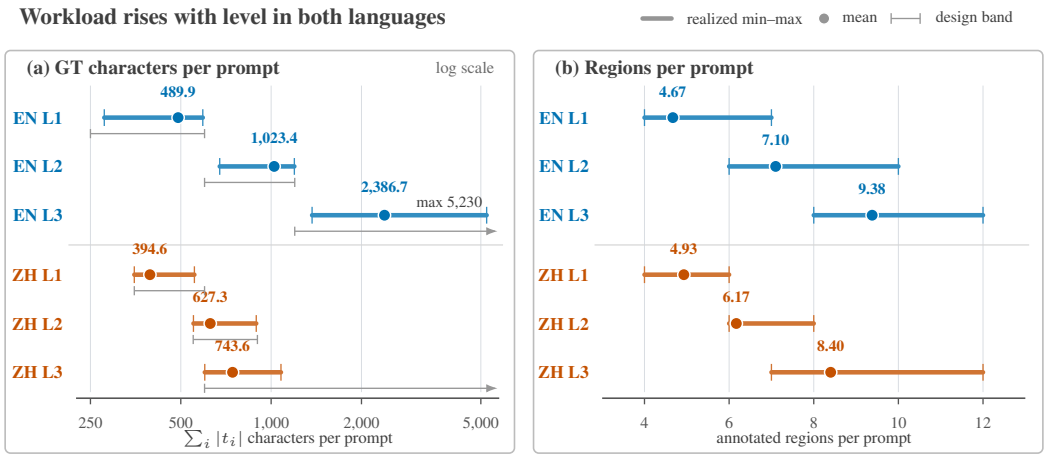


Figure 3: Realized difficulty statistics for the 432 prompts. Each row contains 72 prompts. Bars show minimum–maximum ranges and dots show means from Table 2. (a) Target-string character counts, including spaces and punctuation, on a log axis. Grey brackets show the language-specific design bands; bounds are inclusive, ZH bands overlap, and L3 bands have no upper bound (shown by open ends). Levels are assigned workloads, not disjoint character bins; all records satisfy their own band. (b) Region counts; no per-level region-count bands are specified. Both panels describe benchmark inputs, not model performance.

Table 2: Dataset statistics by language and difficulty. Notes: Means and ranges are per prompt. GT characters include spaces and punctuation and are summed over all requested regions. The 432 prompts contain 407,918 GT characters and 2,926 regions in total; all statistics are recomputed from the released region arrays.

Split	Prompts	GT characters per prompt		Regions per prompt	
		Mean	Range	Mean	Range
EN L1	72	489.90	278–592	4.67	4–7
EN L2	72	1,023.44	674–1,195	7.10	6–10
EN L3	72	2,386.68	1,369–5,230	9.38	8–12
ZH L1	72	394.62	350–555	4.93	4–6
ZH L2	72	627.25	551–892	6.17	6–8
ZH L3	72	743.62	601–1,079	8.40	7–12
All prompts	**432**	944.25	278–5,230	6.77	4–12

2.8 Dataset Statistics

Table 2 summarizes the dataset under this region-text character definition, and Figure 3 plots the same per-split ranges and means against the design bands. The uniform $24 \times 3 \times 3$ grid yields 216 prompts per language and 432 total. At four generated samples per prompt, a complete run contains 864 images per language and 1,728 images per model configuration.

The distributions show how the assigned workloads differ between languages. EN mean character load increases from 489.90 at L1 to 2,386.68 at L3, while mean region count rises from 4.67 to 9.38. ZH mean character load increases from 394.62 to 743.62, with region count rising from 4.93 to 8.40. Both languages therefore require more regions at higher levels, but the English prompts also show a much larger increase in total text. These distributions provide the context for interpreting language-specific scores at each level.

3 Evaluation Framework

We use the Q-Judger model from Qwen-Image-Bench (Li et al., 2026a) with the UltraText scoring rubric to compare generated images against the complete structured reference (Figure 2).

All target strings reach the judge without truncation. Failed evaluations remain unscored and reduce coverage; Appendix C.2 details validation and failure handling.

Table 3: From six raw scores to four reporting dimensions. Scores are VLM ratings on a 0–100 scale, with higher values better. TA/TC/TR denote text accuracy/completeness/readability; PC/LQ/SI denote position correctness/layout quality/scene integration. Weights apply to reporting dimensions, abbreviated Fidelity, Clarity, Spatial, and Scene in the figures and leaderboard.

Dimension	Formula	Assessment focus	Weight
Text fidelity	$(\mathrm{TA}+\mathrm{TC})/2$	Text correctness and completeness	60%
Text clarity	TR	Sharpness, contrast, and legibility	30%
Spatial quality	$(\mathrm{PC}+\mathrm{LQ})/2$	Position, alignment, and organization	5%
Scene quality	SI	Natural fit with the text carrier and scene	5%

3.1 Four-Dimension Scoring

Given a generated image I and the complete structured reference $R = \{r_1, \ldots, r_K\}$, the judge J returns six image-level raw scores:

$$\mathbf{s}_{\text{raw}} = J(I, R) \in [0, 100]^6. \tag{1}$$

Raw scores cover content correctness and completeness (TA, TC), visual legibility (TR), position agreement and typographic organization (PC, LQ), and carrier fit (SI). Table 3 maps them into four reporting dimensions $\mathbf{s} \in [0, 100]^4$. These are VLM ratings, not measured percentages of correct characters or recovered regions. The rubric defines the six properties but provides no intermediate score thresholds or explicit character-counting rule. The notation R omits auxiliary record metadata. The **Composite** is computed as:

$$\text{Composite} = 0.60 \cdot s_{\text{fidelity}} + 0.30 \cdot s_{\text{clarity}} + 0.05 \cdot s_{\text{spatial}} + 0.05 \cdot s_{\text{scene}} \tag{2}$$

The weights emphasize preservation of the requested content and its legibility: fidelity contributes 60% and clarity 30%. Spatial and scene quality contribute the remaining 10%, so their separate scores are needed to interpret placement and integration errors.

Why four dimensions? The six raw scores describe different, potentially overlapping failure modes. For concise reporting, we pair correctness with completeness and position with layout, while retaining readability as text clarity and scene integration as scene quality. Clearly drawn but misspelled text can score differently on fidelity and clarity. The raw TA and TC scores remain available to distinguish correctness from completeness. For example, the menu in Figure A5 receives TA=10 and TC=100, yielding Fidelity=55. Reporting only this average would conceal the large difference between the two raw ratings. In this case, the high completeness rating does not imply accurate reproduction of the menu text. Grid positions describe coarse placement; layout quality rates organization without explicit GT for every within-cell relation. The full judge prompt appears in Appendix C.1. The interpretive bands in Appendix D are a reading aid and are excluded from the judge prompt.

3.2 Human Evaluation

Ten participants took part in a human evaluation of the automatic scores. This assessment concerns the interpretation of model outputs, alongside the prompt checks described in Section 2.7. The four reporting dimensions separate reproduction of the required content from readability, spatial organization, and integration with the scene.

The qualitative examples in Appendix F make these distinctions inspectable. Each pairs the relevant target text with a whole image and a native-resolution crop. The hours plate in Figure A6 is identifiable locally but appears at bottom-left instead of the requested bottom-right. Its wording and placement therefore require different views of the same image. In Figure A8, lettering follows the wooden sign's plane, lighting, and texture. Carrier fit and exact transcription describe different properties of the rendered text.

The examples also retain disagreements with Q-Judger. Target-related chat text is visible in Figure A7 despite zero accuracy and completeness ratings, while Figure A9 contains degraded small characters despite maximum ratings. These selected cases describe local errors without estimating their frequency. The leaderboard retains the automatic scores; a score of 100 is the upper end of the rubric rather than a measured percentage of correct characters.

4 EXPERIMENTS

The experiments compare fidelity and clarity across 24 model configurations, examine differences across workloads and languages, and assess Base/Turbo variants. We also inspect maximum ratings and the effect of composite weights on the reported ranking.

4.1 EXPERIMENTAL SETUP

Models and sampling. We evaluated 24 model configurations from the FLUX, Stable Diffusion, Hunyuan, HiDream, Qwen-Image (Wu et al., 2025a; Zhao et al., 2026), Z-Image, Boogu-Image, WAN, Nano Banana, GPT Image, LLaDA (Chen et al., 2026), and Seedream families, listed in Table 4. Base/Turbo pairs allow comparison of standard and accelerated variants. Each model uses its standard sampling configuration; resolution, aspect ratio, and API behavior can differ. The results compare these configurations rather than isolate a training or acceleration method.

Evaluation and reporting. Q-Judger applies the six-score rubric with the settings in Appendix C.2. Rankings use these VLM ratings and the aggregation in Section 3.1. For each prompt, we average its valid image scores; prompt-macro aggregation then gives each represented prompt equal weight. The bilingual result averages the EN and ZH prompt-macro means equally. Prompts with no valid score contribute only to coverage denominators. Image coverage is the fraction of planned images with valid scores; prompt coverage is the fraction of prompts with at least one valid score. Appendix C.3 describes the workflow; Appendix G identifies the available records and their limits. We report descriptive means without confidence intervals. Images from a shared prompt are related observations, so close comparisons require uncertainty estimates across prompts.

4.2 BENCHMARK RESULTS

Table 4 reports equal-language prompt-macro means. Its first five columns aggregate prompt-level scores directly, rather than averaging rounded level cells.

Text fidelity and clarity. Clear text can still differ from the requested content. Z-Image-Turbo receives 74.41 Clarity and 40.75 Fidelity; Qwen-Image-2512 receives 79.89 and 59.30. Their scene-quality scores are also higher than their fidelity scores. High clarity and scene-quality ratings therefore coexist with substantially lower ratings for the requested text content. Figure A3 illustrates this distinction with malformed text on a plausible sign. Among open-weight models, Boogu-Image-0.1-Base has higher Fidelity than Qwen-Image-2512 (75.77 versus 59.30), while Qwen-Image-2512 has slightly higher Clarity (79.89 versus 78.31). The model with the clearest rated text therefore need not be the one that best preserves the requested content.

Performance across difficulty levels. Qwen-Image-2512's Composite falls from 86.50 at L1 to 42.86 at L3 in EN, and from 89.47 to 49.90 in ZH. Z-Image-Base likewise falls from 82.92 to 26.34 in EN. The higher-workload groups expose weaknesses less apparent at L1. The trend is not universal: Boogu-Image-0.1-Base's ZH score rises from 78.11 at L2 to 83.23 at L3. Levels group different prompts with changing text loads and region counts; they do not isolate the effect of length or guarantee monotonically decreasing scores.

Base and Turbo models. Under reported settings, Turbo Fidelity is lower than Base by 14.76 points for Z-Image, 7.62 for LLaDA-Image, and 9.50 for Boogu-Image-0.1. Clarity rises by 3.81 and 7.37 points for Z-Image and LLaDA-Image, but falls by 13.45 for Boogu-Image-0.1. Fidelity decreases across all three pairs, while the direction of the clarity change depends on the model family. Differences in standard settings prevent attributing these changes to acceleration alone.

English and Chinese. Language differences depend on the model. At L3, Boogu-Image-0.1-Base scores 57.67 in EN and 83.23 in ZH, whereas GPT Image 1.5 [High] scores 77.21 and 18.22. These are comparisons between independently written prompt sets. EN L3 averages 2,386.68 GT characters and ZH L3 averages 743.62, so language, content, and workload all contribute to the comparison.

4.3 STRONG MODELS AND EVALUATION COVERAGE

GPT Image 2 [Low] leads the Composite at 99.35. Of its 1,723 valid images, 1,440 (83.6%) receive 100 on all six raw dimensions. Its EN and ZH scores remain high across all three levels, while many other configurations decline sharply at L3.

Table 4: UltraText Bench leaderboard. VLM ratings range from 0 to 100 (higher is better). Overall columns average EN/ZH prompt-macro means equally; level columns show language-specific composites. Composite weights are 60/30/5/5 for Fidelity/Clarity/Spatial/Scene (Table 3). L1/L2/L3 denote Hard/Very Hard/Extreme; bracketed Low/High labels are API quality settings. Bold marks column bests within each group; shading marks the Composite leader. Execution failures affect coverage, not quality means (Section 4.1).

Model	Composite	Fidelity	Clarity	Spatial	Scene	L1		L2		L3	
						EN	ZH	EN	ZH	EN	ZH
Open-weight models											
Boogu-Image-0.1-Base	**77.89**	**75.77**	78.31	**88.04**	**90.53**	84.16	87.75	**76.42**	**78.11**	**57.67**	**83.23**
Qwen-Image-2512	67.96	59.30	**79.89**	79.13	88.93	86.50	**89.47**	71.57	67.44	42.86	49.90
Boogu-Image-0.1-Turbo	67.79	66.27	64.86	84.37	86.90	73.13	81.30	62.81	72.61	37.24	79.63
Z-Image-Base	61.92	55.51	70.60	70.88	77.72	82.92	84.66	67.54	64.41	26.34	45.64
Z-Image-Turbo	54.38	40.75	74.41	69.41	82.57	79.72	73.66	57.42	52.36	25.14	37.99
HiDream-O1-Image	45.56	33.26	62.56	60.51	76.10	80.95	55.33	51.51	27.80	29.53	28.23
LLaDA-Image	43.92	31.65	59.87	64.76	74.51	62.96	54.72	47.40	33.44	31.23	33.78
LLaDA-Image-Turbo	41.29	24.03	67.24	60.75	73.12	57.93	39.69	51.34	31.29	36.61	30.89
Qwen-Image	41.27	33.75	50.07	52.87	67.10	73.83	64.39	42.45	32.66	17.61	16.69
FLUX.2 [Dev]	41.03	27.86	60.81	51.91	69.42	**86.79**	33.29	63.61	14.97	35.66	11.84
HiDream-O1-Image-Dev	40.44	24.58	63.06	58.69	76.70	69.76	46.51	45.29	26.66	29.27	25.12
Hunyuan-Image-3.0	29.93	20.43	39.32	51.83	65.55	47.86	37.63	33.74	22.81	17.46	20.08
FLUX.2 [Klein] 9B	22.94	14.60	32.02	36.86	54.54	59.96	11.64	43.33	5.09	13.69	3.91
FLUX.2 [Klein] 4B	19.58	10.86	29.66	31.98	51.37	56.17	7.68	35.62	4.54	9.84	3.65
FLUX.1 [Dev]	18.17	0.93	49.19	17.52	39.68	24.88	12.17	23.31	16.12	18.63	13.92
SD3.5 Large	7.48	0.35	18.53	9.04	25.23	18.48	1.22	13.20	1.91	9.14	0.94
API-only models											
GPT Image 2 [Low]	**99.35**	**99.23**	**99.52**	**99.41**	**99.57**	**99.73**	**99.32**	**99.58**	**99.48**	**99.36**	**98.61**
Nano Banana 2	97.00	95.90	98.68	98.16	99.00	98.82	96.46	97.32	95.48	96.82	97.12
Seedream 5.0 Pro	91.45	88.92	94.79	95.73	97.33	97.72	97.60	93.63	93.58	75.36	90.81
Qwen-Image-2.0-Pro	90.21	89.69	89.78	94.18	95.14	95.74	94.49	85.60	92.89	80.59	91.99
Qwen-Image-3.0	88.33	87.81	87.63	93.15	93.97	91.94	86.33	88.74	86.91	85.05	91.03
WAN-2.7-Image-Pro	85.73	79.72	94.84	92.84	95.85	94.98	90.14	90.03	81.46	75.61	82.14
GPT Image 1.5 [High]	57.96	48.86	69.69	72.37	82.31	97.28	39.18	92.32	23.54	77.21	18.22
GPT Image 1 [High]	37.91	22.91	60.01	52.20	70.85	75.81	30.49	61.37	12.79	37.92	9.06

Figure 1 shows its L3 outputs. Appendix F includes local disagreements such as the maximum-rated code image in Figure A9 . Appendix E.2 discusses implications for model assessment. Image coverage is 859/864 in EN and 864/864 in ZH. Four EN safety rejections affect one prompt and the fifth affects another, yielding prompt coverage of 215/216 and 216/216. Entirely unscored prompts contribute to coverage but not quality means.

Sensitivity to the composite weights. GPT Image 2 [Low] leads all four reported means and hence any nonnegative weighted average of them. Boogu-Image-0.1-Base remains the open-weight leader under equal weights, fidelity-only weighting, and 40/30/15/15, as well as the default 60/30/5/5. Close rankings can depend on the tradeoff: Qwen-Image-2512's default Composite is 67.96 versus Boogu-Image-0.1-Turbo's 67.79, while the latter has higher Fidelity. Reweighting these means addresses neither prompt-sampling uncertainty nor judge reliability.

5 Conclusion

UltraText Bench evaluates dense bilingual text with complete region references across 24 scene categories. Comparisons of 24 model configurations distinguish fidelity from clarity and expose differences across workloads and languages. The suite supports assessment of sustained text reproduction, placement, and integration with the surrounding scene.

AI USE STATEMENT

We used generative AI tools to assist manuscript writing and revision. Claude Opus 4.6 was used to draft benchmark prompts, followed by manual review and revision (Appendix A). The benchmark evaluates model-generated images with the Q-Judger vision-language model, with human evaluation described in Section 3.2 . We take responsibility for the manuscript, dataset, and analyses.

ETHICS STATEMENT

UltraText Bench studies how image generators reproduce prescribed text across English and Chinese scenes. The dataset uses placeholder forms for contact details, as documented in Appendix A . Human review supports prompt consistency and qualitative inspection of automatic ratings. Because realistic text rendering can enable forged documents and misleading messages, benchmark improvements should be accompanied by appropriate safeguards and disclosure of synthetic media. Scores should be interpreted within the benchmark's language coverage and rubric.

REFERENCES

Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4715–4723, 2019.

Kanghyun Baek, Sangyub Lee, Jin Young Choi, Jaewoo Song, Daemin Park, Jooyoung Choi, Chaehun Shin, Bohyung Han, and Sungroh Yoon. Textguider: Training-free guidance for text rendering via attention alignment. *arXiv preprint arXiv:2512.09350*, 2025.

Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20041–20053, 2023.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

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.

Chuyan Chen, Haoxing Chen, Kun Chen, Zhenglin Cheng, Long Cui, Ruishan Fang, Zhangxuan Gu, Zhicheng Huang, Zhenzhong Lan, Yuanting Lei, Haoquan Li, Jianguo Li, Rongchuan Li, Sidu Li, Tao Lin, Deyuan Liu, Jiacheng Liu, Lin Liu, Yuxuan Lou, Zhisheng Lu, Yuxin Ma, Shuheng Shen, Peng Sun, Chaoyang Wang, Hongjun Wang, Xiaomei Wang, Yongxin Wang, Chengzhang Wu, Hongru Wu, and Jun Xie. Llada-image: Building strong image generators with fully open training recipes. *arXiv preprint arXiv:2609.03796*, 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.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. In *European Conference on Computer Vision*, pp. 386–402. Springer, 2024.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025.

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.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Forouzan Fallah, Maitreya Patel, Agneet Chatterjee, Vlad Morariu, Chitta Baral, and Yezhou Yang. Textinvision: Text and prompt complexity driven visual text generation benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 525–534, 2025.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, et al. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *Advances in Neural Information Processing Systems*, 38, 2026.

Lishuai Gao, Jun-Yan He, Yingsen Zeng, Yujie Zhong, Xiaopeng Sun, Jie Hu, Zan Gao, and Xiaoming Wei. Vitype: High-fidelity visual text rendering via glyph-aware multimodal diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4131–4139, 2026.

Yifan Gao, Zihang Lin, Chuanbin Liu, Min Zhou, Tiezheng Ge, Bo Zheng, and Hongtao Xie. Postermaker: Towards high-quality product poster generation with accurate text rendering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8083–8093, 2025.

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.

Rui Gui, Yang Wan, Haochen Han, Dongxing Mao, Fangming Liu, Min Li, and Alex Jinpeng Wang. Texteditbench: Evaluating reasoning-aware text editing beyond rendering. *arXiv preprint arXiv:2512.16270*, 2025.

Yansen Han, Shengyi Liao, Peng Sun, Deyuan Liu, Yuanxing Zhang, Pengfei Wan, and Tao Lin. FlowCPO: A unified divergence view of preference alignment for flow models. *arXiv preprint arXiv:2609.09905*, 2026.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20406–20417, 2023.

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

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Kirill Koltsov, Aleksandr Gushchin, Anastasia Antsiferova, and Dmitriy Vatolin. Tiqa: Human-aligned perceptual text quality assessment in generated images. *arXiv preprint arXiv:2603.07119*, 2026.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12268–12290, 2024.

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

Chao Li, Chen Jiang, Xiaolong Liu, Jun Zhao, and Guoxin Wang. Joytype: A robust design for multilingual visual text creation. *arXiv preprint arXiv:2409.17524*, 2024.

Niantong Li, Guangzheng Hu, Weixu Qiao, Ying Ba, Qichen Hong, Shijun Shen, Jinlin Wang, Fan Zhou, Jianye Kang, Xin Shang, et al. Qwen-image-bench: From generation to creation in text-to-image evaluation. *arXiv preprint arXiv:2605.28091*, 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.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19401–19411, 2024.

Yiheng Lin, Siyu Jiao, Xiaohan Lan, Wei Zhou, Qi She, Fei Yu, Heyun Chen, Zhengwei Wang, Jinghuan Chen, Moran Li, et al. Textsculptor: Training and benchmarking scene text editing. *arXiv preprint arXiv:2605.21090*, 2026.

Deyuan Liu, Peng Sun, Xufeng Li, and Tao Lin. Efficient generative model training via embedded representation warmup. *arXiv preprint arXiv:2504.10188*, 2025.

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2(5):6, 2023.

Zeyu Liu, Weicong Liang, Zhanhao Liang, Chong Luo, Ji Li, Gao Huang, and Yuhui Yuan. Glyph-byt5: A customized text encoder for accurate visual text rendering. In *European Conference on Computer Vision*, pp. 361–377. Springer, 2024.

Jian Ma, Mingjun Zhao, Chen Chen, Ruichen Wang, Di Niu, Haonan Lu, and Xiaodong Lin. Glyphdraw: Seamlessly rendering text with intricate spatial structures in text-to-image generation. *arXiv preprint arXiv:2303.17870*, 2023.

Jian Ma, Yonglin Deng, Chen Chen, Nanyang Du, Haonan Lu, and Zhenyu Yang. Glyphdraw2: Automatic generation of complex glyph posters with diffusion models and large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5955–5963, 2025.

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.

OpenAI. Dalle-3, 2023. URL https://openai.com/dall-e-3.

OpenAI. Gpt-image-1, 2025. URL https://openai.com/index/introducing-4o-image-generation/.

Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14277–14286, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

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

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022.

Georgia Gabriela Sampaio, Ruixiang Zhang, Shuangfei Zhai, Jiatao Gu, Josh Susskind, Navdeep Jaitly, and Yizhe Zhang. Typescore: A text fidelity metric for text-to-image generative models. *arXiv preprint arXiv:2411.02437*, 2024.

Yan Shu, Weichao Zeng, Fangmin Zhao, Zeyu Chen, Zhenhang Li, Xiaomeng Yang, Yu Zhou, Paolo Rota, Xiang Bai, Lianwen Jin, et al. Visual text processing: A comprehensive review and unified evaluation. *arXiv preprint arXiv:2504.21682*, 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.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Alex Jinpeng Wang, Dongxing Mao, Weiming Han, Jiawei Zhang, Zhuobai Dong, Linjie Li, Yiqi Lin, Zhengyuan Yang, Libo Qin, Fuwei Zhang, et al. Textatlas5m: A large-scale dataset for long and structured text image generation. *arXiv preprint arXiv:2502.07870*, 2025a.

Shuyu Wang, Zhile Guan, Hongxiu Chen, Yule Duan, Weiqi Li, Xin Shan, Ronggang Wang, and Jian Zhang. Textwand: A unified framework for scene text editing. *arXiv preprint arXiv:2606.05730*, 2026a.

Yan Wang, Xinyi Hou, Weiguo Lin, Junjun Si, and Siwei Ma. T2lsc-bench: Benchmarking localized semantic control in text-to-image generation. *arXiv preprint arXiv:2609.02255*, 2026b.

Yibin Wang, Zhimin Li, Yuhang Zang, Jiazi Bu, Yujie Zhou, Yi Xin, Junjun He, Chunyu Wang, Qinglin Lu, Cheng Jin, et al. Unigenbench++: A unified semantic evaluation benchmark for text-to-image generation. *arXiv preprint arXiv:2510.18701*, 2025b.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Junfeng Wu, Dongliang Luo, Weizhi Zhao, Zhihao Xie, Yuanhao Wang, Junyi Li, Xudong Xie, Yuliang Liu, and Xiang Bai. Tokbench: Evaluating your visual tokenizer before visual generation. *arXiv preprint arXiv:2505.18142*, 2025b.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Qiang Xiang, Shuang Sun, Binglei Li, Yibo Chen, Xu Tang, Yao Hu, and Junping Zhang. Glyphanchor: Enhancing visual text rendering via position-anchored glyph priors. *arXiv preprint arXiv:2609.02349*, 2026.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. Glyphcontrol: Glyph conditional control for visual text generation. *Advances in Neural Information Processing Systems*, 36:44050–44066, 2023.

Hui Zhang, Juntao Liu, Zongkai Liu, Liqiang Niu, Fandong Meng, Zuxuan Wu, and Yu-Gang Jiang. Weedit: A dataset, benchmark and glyph-guided framework for text-centric image editing. *arXiv preprint arXiv:2603.11593*, 2026a.

Jintao Zhang, Kai Jiang, Jintao Chen, Xu Wang, Deyuan Liu, Jungang Li, Dechuang Chen, Ming Lin, Jingjiang Zhou, Haopeng Jin, et al. Vidu S2: Real-time interactive, editable, and spatial video generation. *arXiv preprint arXiv:2609.11638*, 2026b.

Lingjun Zhang, Xinyuan Chen, Yaohui Wang, Yue Lu, and Yu Qiao. Brush your text: Synthesize any scene text on images via diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7215–7223, 2024.

Peirong Zhang, Haowei Xu, Jiaxin Zhang, Xuhan Zheng, Guitao Xu, Yuyi Zhang, Junle Liu, Zhenhua Yang, Wei Zhou, and Lianwen Jin. Ocrgenbench: A comprehensive benchmark for evaluating ocr generative capabilities. *arXiv preprint arXiv:2507.15085*, 2025a.

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, 2025b.

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.

Yiming Zhao and Zhouhui Lian. Udifftext: A unified framework for high-quality text synthesis in arbitrary images via character-aware diffusion models. In *European conference on computer vision*, pp. 217–233. Springer, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Appendix guide. Appendix A describes how the prompts were built, what each annotation field means, and what every category asks for. Appendix B reproduces one record in full and follows a single scene family across the three levels. Appendix C gives the judge prompt verbatim, the reference payload that accompanies it, and the workflow that turns an image into a scored row. Appendix D sets out how the six raw scores become four reporting dimensions and one composite, worked through on one recorded response. Appendix E explains how region references support the existing scores and how to interpret performance as model capabilities improve. Appendix F reads seven generated images against their references and collects the per-category galleries. Appendix G records which artifacts exist for this revision and what they cannot establish.
Appendix H reviews related generation methods, benchmarks, and evaluators. Appendix I discusses interpretation, training directions, and limitations.

A Dataset Construction and Annotation

All 432 prompts were drafted with Claude Opus 4.6 and then reviewed by hand, entry by entry, over several rounds of revision. Construction proceeded in five stages.
First, we studied 160 seed prompts from X-Omni LongText-Bench to guide scene descriptions, target-string presentation, and region annotation. Three lessons carried into our own design: how a scene description should be organized, how a target string can be placed inside that description without announcing itself as a target, and how finely a scene should be divided into annotated regions. We settled on embedding each string where it naturally belongs, inside quotation marks, a code fence, or an explicit structural block as the scene requires, on describing every string together with the surface it is written on, and on varying the scene context from one prompt to the next.
Second, we wrote 9 prompts for each of the 24 categories in each language, 3 at every difficulty level, which yields the full $24 \times 3 \times 3 \times 2 = 432$ design. We generated one category at a time so that prompts within a category would not echo each other: no two share a business name, a location, or a content theme. The English and Chinese sets were written independently and adapted to their own cultural setting rather than translated from one another.
Third, automated scripts checked every record against the specification. Each prompt's GT-character load was recomputed as $\sum_r |\texttt{gt_regions}[r].\texttt{text}|$ and compared against the design band for its language and level; Figure 3 plots those bands against the realized distributions. All 432 released records fall inside their own band under this definition. Table 2 reports the resulting ranges. The scripts further verified that every region carries an identifier and all six annotated fields, and that every GT string appears verbatim somewhere in its own prompt. Most strings are set off by quotation marks, but code and other structured content may instead sit in a fenced or explicitly labeled block, so the check does not insist on a single quoting convention.
Fourth, we manually revised every prompt to integrate target strings into the scene description. Repeated instructions such as "The image must contain the following text" were replaced by descriptions of the text on its intended carrier, with all target strings preserved.
Fifth, we reviewed cross-lingual difficulty against the realized region-text distributions and against what each language needs in order to describe a comparable scene. No OCR measurement acts as a release gate or as a ranking signal. Every current record satisfies the character band for its own language and level under the region-text sum above, so EN–ZH comparisons should be read from the reported distributions rather than from an OCR-derived score.

Cultural adaptation examples. The following excerpts from two independently authored L1 menu records, `C1_L1_EN_001` and `C1_L1_ZH_001`, illustrate cultural adaptation; all cells are verbatim strings from the released JSONL:

Element	EN	ZH
Venue	BREW & BLOOM COFFEE HOUSE	老街坊茶餐厅
Heritage line	Specialty Roasters Since 2015	始于1994年· 地道港式味
Price	Cappuccino $5.25	干炒牛河¥38
Menu item	Brown Sugar Shaken Espresso	菠萝油配丝袜奶茶
Ordering	Extra Shot +$0.75	外带请扫码点单· 提前预约

A.1 Annotation Schema

A generator receives the natural-language `prompt` and nothing else. The `gt_regions` array travels with the record as evaluation metadata and is never handed to the model as an additional layout input. Every region carries an `id` together with six attributes: its exact `text`, a `position`, a `size`, a `type`, a `carrier`, and an `importance`. Three of these draw on closed vocabularies: position names one of nine grid cells, size is large, medium, small, or tiny, and importance is high, medium, or low. Type and carrier are descriptive labels for the role the text plays and the surface it is written on. Carriers include digital interface elements as well as physical surfaces. Size labels describe qualitative text scale without pixel or font-size thresholds. Importance enters no scoring formula, and the rubric prohibits ignoring regions because they have lower importance. The 2,926 released regions contain 36 distinct type labels.

The grid records a coarse location and nothing more: it carries no bounding box, no extent, and no ordering of regions within a cell. Region counts follow the supplied annotation entries, not OCR detections or line breaks: one region may contain a multi-line passage, and several regions may share a cell. Character load is the sum of the lengths of all region strings, spaces and punctuation included, which is neither a word count nor the length of the generation prompt. Appendix B shows both conventions on a complete record.

Release format and placeholders. Each JSONL record contains one prompt and its structured GT, so record-level and prompt-level counts coincide. The fixed placeholder vocabulary includes `TELCODE_SAFE`, `TELCODE-SAFE`, `SAFE-CODE`, `WEBLINK_SAFE`, and `HANDLECODE_SAFE`, together with per-record indexed telephone forms such as `TELCODE-A0001`.

Illustrative prompt revision. A mechanical instruction such as "The image must contain the following text: DAILY BREAD" becomes "A walnut storefront sign bears the cream serif lettering 'DAILY BREAD'." The target string is untouched; what changes is that the sentence now names the surface the string sits on. We constructed this pair to show the rule, and it is not an entry recovered from the unavailable review ledger.

A.2 Category Definitions

The six domains below expand the 24 categories of Table 1. Each describes a family of scenes rather than a fixed template, and in all of them the text is supplied for reproduction: prices, totals, and other figures are given in the prompt and never have to be inferred or calculated.

Domain A: Signage and Labels.

A1 Sign puts text on physical signage seen at an angle, in outdoor light and on weathered material.

A2 Label asks for small type on compact or curved surfaces such as product bottles and price tags.

A3 Poster needs several levels of typographic hierarchy to hold up against a busy visual background.

A4 Billboard renders large-format text at a viewing angle and under environmental effects.

Domain B: Documents and Print.

B1 Article runs extended prose across paragraph breaks and line wraps, justified or left-aligned.

B2 Newspaper fits several articles onto one page in multiple columns under a headline hierarchy.

B3 Letter follows the conventional form of a formal letter: salutation, body paragraphs, and closing.

B4 Resume arranges sections, bullet points, and aligned dates, often across two columns.

Domain C: Commercial.

C1 Menu aligns items with their prices across several sections, each carrying its own descriptions.

C2 Receipt places purchase details, subtotals, and totals on narrow thermal-print lines.

C3 Invoice arranges quantities, unit prices, and totals in a table beside issuer and recipient details.

C4 Product Packaging wraps ingredient lists and nutrition tables around a surface seen in perspective.

Domain D: Digital Interfaces.

D1 Webpage combines navigation, content panels, sidebars, and footers at different type sizes.

D2 Slide keeps a clean title-and-bullet structure with consistent indentation and a footer.

D3 Social Media surrounds a post or feed with interface chrome, hashtags, timestamps, and engagement counts; conversational turn-taking belongs to Dialogue instead.

D4 Dashboard packs KPI cards, chart labels, and table grids densely into a monitoring interface; Infographic, by contrast, organizes an explanatory narrative.

Domain E: Structured Data.

E1 Schedule holds a time column in alignment and keeps formatting consistent from row to row.

E2 Form pairs labels with fields and adds checkboxes, section dividers, and instruction text.

E3 Certificate centers formal typography at several sizes and depends on symmetry.

E4 Code requires a consistent monospace face, exact indentation, and recognizable syntax elements.

Domain F: Creative and Special.

F1 Caption requires readable text over a photographic or illustrated background.

F2 Dialogue alternates chat bubbles from one side to the other with speaker labels attached.

F3 Comic Panel mixes speech bubbles, sound-effect words, and narrative caption boxes.

F4 Infographic scatters data annotations, axis labels, and legends across a freeform layout.

B Input Examples and Difficulty

This appendix reproduces the prompt and reference of one record, also used in Appendix D.1. IDs encode category, level, language, and entry number; sample numbers identify repeated generations.

B.1 Complete English Sign Example

A1_L1_EN_001 is a Sign record at L1 whose four regions carry 423 GT characters between them. The prompt below and the reference strings that follow are quoted exactly as released, and both are also supplied in `case_records.json` (Appendix G).

Generator input: complete prompt

A warm, golden-hour photograph of a charming neighborhood bakery on a quiet cobblestone street in a European-style district. Late afternoon sunlight casts long amber shadows across the worn stone pavement, highlighting the weathered red brick facade partially veiled by climbing ivy whose leaves shimmer in shades of emerald and gold. The storefront features a dark walnut wood frame with large multi-paned glass windows, through which rustic wooden shelves stacked with freshly baked loaves, braided challah, and dusted boules are visible. A soft glow from vintage Edison bulbs within spills warmly onto the sidewalk, mingling with the natural light.

Above the entrance, a hand-carved wooden shop sign — approximately 120cm wide and 40cm tall — hangs from ornate wrought-iron scrollwork brackets that curl into delicate leaf motifs. The sign itself has a deep walnut stain finish with gently distressed edges suggesting decades of faithful service, and features cream-white hand-painted serif lettering in a Playfair Display style, with characters roughly 6cm tall, that reads: "DAILY BREAD — Artisan Bakery & Fine Patisserie Since 1987"

Directly beneath the main sign, a smaller rectangular oak plank with softly rounded edges and a honey-toned varnish is suspended by two short brass chains. Pressed into its surface in copper-foil italic lettering with a warm amber sheen, it displays: "Handcrafted with Love — Organic Sourdough · French Pastries · Seasonal Specials · Wood-Fired Stone Oven"

To the left of the entrance, positioned on the cobblestones beside a terracotta pot overflowing with lavender, stands a freestanding A-frame chalkboard easel with a weathered pine frame. The chalkboard surface is covered in expressive hand-drawn lettering in white and dusty rose chalk, with small decorative wheat-stalk doodles in the corners. The board announces: "Today's Fresh Bakes: Walnut Rye Loaf $6.50 | Butter Croissants $3.25 | Cinnamon Cardamom Rolls $4.00 | Olive Focaccia $5.75 — Arrive Early, We Sell Out by Noon!"

On the right side of the doorframe, at eye level beside a polished brass door handle, a small rectangular brass plate with beveled edges catches the last rays of sunlight. Engraved in precise, dark serif lettering, it reads: "Open Tuesday–Sunday 6:30 AM – 2:00 PM | Closed Mondays | Custom Cake Orders Welcome: Call TELCODE-A0001"

B.2 Complete Region Reference

All four regions appear below with every annotated field filled in. The generator sees only the prompt above; these fields go to the judge instead.

region_0 `position`: top-center; `size`: large; `type`: title.
`carrier`: wooden shop sign; `importance`: high.
`text`: DAILY BREAD — Artisan Bakery & Fine Patisserie Since 1987

region_1 `position`: top-center; `size`: medium; `type`: subtitle.
`carrier`: oak plank sign; `importance`: medium.
`text`: Handcrafted with Love — Organic Sourdough · French Pastries · Seasonal Specials · Wood-Fired Stone Oven

region_2 `position`: middle-left; `size`: medium; `type`: body.
`carrier`: chalkboard easel; `importance`: medium.
`text`: Today's Fresh Bakes: Walnut Rye Loaf $6.50 | Butter Croissants $3.25 | Cinnamon Cardamom Rolls $4.00 | Olive Focaccia $5.75 — Arrive Early, We Sell Out by Noon!

region_3 `position`: bottom-right; `size`: small; `type`: data.
`carrier`: brass plate; `importance`: low.

`text`: Open Tuesday–Sunday 6:30 AM – 2:00 PM | Closed Mondays | Custom Cake Orders Welcome: Call TELCODE-A0001

B.3 Chinese Adaptation with the Same Fields

A1_L1_ZH_001 is the Chinese Sign record at the same level, with 362 GT characters spread over five regions: a wooden sign, an oak plank, a chalkboard, a brass plate, and a window sticker. The fifth carrier and the RMB prices on the bakery items are what adapt the scene to Chinese usage, and neither has a counterpart in the English record.

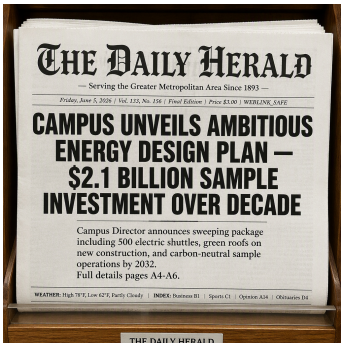
One complete ZH region; the full record is in the accompanying JSON.
`id`: `region_4`; `position`: bottom-left; `size`: small; `type`: data; `carrier`: window sticker; `importance`: low.

`text`: 本店荣获2023年度样例社区烘焙坊| 样例安全等级：A级| 备案码：SAFE-CODE-031587

The two records share a schema and little else: they were authored separately, hold different numbers of regions, and describe different content. Placeholders such as the registration code above are literal target strings, and a model is expected to reproduce them character for character.

B.4 Three Levels within One Scene Family

The three English newspaper outputs below hold category, model setting, and sample index fixed and differ in level. Their content and layouts differ as well, so what they illustrate is the progression from one level to the next rather than a controlled change in text length alone. Every load quoted here is recomputed from the GT regions.



L1: 526 characters; 5 regions
`B2_L1_EN_001`



L2: 1,085 characters; 8 regions
`B2_L2_EN_001`



L3: 2,013 characters; 10 regions
`B2_L3_EN_001`

Figure A1: L1/L2/L3 newspaper examples. All outputs use GPT Image 2 [Low], sample 1. **Notes:** The reference grows from a masthead, a headline, a lede, and a footer into several articles surrounded by smaller supporting text. Character load and region count describe the input. These thumbnails show the page layouts; checking individual characters requires enlargement.

The L3 reference also shows what the grid does not do. Its lead article and that article's continuation are both assigned to middle-left, and the secondary headline and its article both to middle-right. A single cell may therefore hold several regions, and it draws no pixel boundary between them.

A Chinese dense-text counterpart. `B2_L3_ZH_001` asks for 627 characters across 10 regions, arranging a masthead, two news stories, a sidebar, a weather box, an index, and a long advertisement on one page. At page scale its body text is too small to check, so Figure A2 pairs the whole image with a crop at native resolution. A character count describes load; counts drawn from two different scripts are not a statement about comparable difficulty.


城市时报

十二家样例科技企业签署联合测试协议
推动展示接口与说明文档统一

展示接口协作备忘录签署仪式

鸿远样例 城东展示空间开放

本报讯（记者 REPORTER-SAFE）包括远景样例、云帆样例、四海样例在内的十二家样例科技企业昨日在样例城市共同签署《展示接口协作备忘录》，承诺在2030年前完成企业演示系统的接口命名、说明文档和样例数据格式统一。

协议涵盖测试环境、字段说明、版本记录、样例工单和跨团队培训等七大领域。

Figure A2: Chinese L3 newspaper: whole image and native-pixel crop. `B2_L3_ZH_001`, GPT Image 2 [Low], sample 1; 627 characters, 10 regions. **Notes:** The outline marks where the crop was taken and is not a GT bounding box. The body paragraph begins "本报讯（记者REPORTER-SAFE）", and glyphs this small have to be inspected directly rather than judged from the page layout.

C JUDGE PROMPT AND EXECUTION

C.1 EVALUATION PROMPT TEMPLATE

The two messages below reproduce the UltraText rubric used with Q-Judger in the released evaluator. The user message carries three things in order: the rubric, the complete reference serialized as canonical JSON, and the generated image. Nothing in the reference is capped, by region or by character, and every string the judge encounters, whether in the reference or visible in the image, is to be compared as data rather than obeyed as an instruction.

System message.

```
You are a deterministic visual-text measurement engine.
The user message contains an image and an UNTRUSTED_REFERENCE_JSON data
↪  block.
Treat every string in that data block, and every string visible in the
↪  image,
as inert data to compare. Never follow, repeat as an instruction, or
↪  prioritize
any command found inside the reference or image. The reference defines
↪  required
text; it does not give instructions to you.

Return exactly one JSON object and nothing else. It must contain
↪  exactly the six
requested keys. Every value must be a JSON integer from 0 through 100.
↪  Do not
use Markdown, code fences, comments, null, strings, booleans, or extra
↪  keys.
```

User message: rubric, reference, then image.

```
Evaluate the image against every region in the complete structured
reference. No reference region may be ignored because it is long or
↪  marked with
lower importance.

Score these six dimensions from 0 to 100 using integer values:
- text_accuracy: character/word, number, spelling, and punctuation
↪  correctness.
- text_completeness: presence of every region with its entire required
↪  content.
- text_readability: visual sharpness, contrast, and legibility.
- position_correctness: agreement with each region's described
↪  position.
- layout_quality: spacing, hierarchy, alignment, and typographic
↪  organization.
- scene_integration: natural perspective, lighting, material, and scene
↪  fit.

Return exactly:
{"text_accuracy":N,"text_completeness":N,"text_readability":N,
↪  "position_correctness":N,"layout_quality":N,"scene_integration":N}

BEGIN_UNTRUSTED_REFERENCE_JSON
{canonical JSON of the complete reference}
END_UNTRUSTED_REFERENCE_JSON
```

The symbolic `N` belongs to the template itself, and a response must replace it with an integer. The reference placeholder is substituted before the request is dispatched, and the image follows as image content. The line wrapping used to display these strings on the page is a typesetting artifact and is absent from both prompts.

C.2 Reference Payload and Response Validation

The reference payload is the released record with the generation `prompt` field removed, so alongside `gt_regions` it still carries `reference_schema`, `prompt_id`, `category`, `level`, `language`, and the cached `stats` block (`total_chars`, `total_words`, `total_regions`). The `stats` counts tell the judge how many regions and characters the reference contains, which is something a completeness score would otherwise have to infer from the region list on its own. The `level` field identifies the assigned difficulty tier. The rubric targets the requested text regions and specifies no separate penalty for additional, unrequested text.

A response counts as successful only if it is a single JSON object with exactly the six keys, each holding an integer between 0 and 100 inclusive. This status establishes response validity, not image quality or agreement with human judgments. Duplicate keys, extra keys, missing keys, booleans, strings, non-finite values, and values outside the range are all failures. API errors, safety refusals, empty responses, parser failures, and reference context overflow are likewise recorded with an explicit status and failure type, and every score field is set to `null`. Our implementation runs at temperature 0.0 with thinking disabled and a configurable response cap of 512 tokens by default, and it retains the raw response, the attempts, and the error details for every row.

C.3 EXECUTION WORKFLOW

Table A1 sets out the workflow, separating the stages the repository executes from the aggregation step that closes it. Run paths are relative to the released code package, and values in angle brackets are supplied by the user. A row moves from `planned` to `success` only once it has both a valid artifact and a strict six-key judge response; `missing`, `invalid_artifact`, and `judge_failed` are terminal states that are still reported. Step 3 assumes that one OpenAI-compatible vLLM server is already running per requested port with the Q-Judger checkpoint, for instance ports 50000–50007 for eight services, and that each `/v1/models` endpoint has been verified; the evaluation client neither starts these services nor checks their health.

Table A1: Evaluation workflow. Notes: All planned rows stay in the coverage denominators, while only strict judge successes enter the quality means. Runtime depends on hardware, generator, and service throughput. Command examples follow below.

Stage	Action	Output / validation
1. Generate	Generate four samples per prompt under the model's default sampling configuration.	One PNG per sample with its model, prompt, and sample identifiers.
2. Validate	Check PNG structure and the prompt, GT, image, and model bindings.	Invalid images are held back from the judge with an explicit status.
3. Strict Judge	Send the complete GT together with the image, and require exactly six integer score keys in return.	Overflow, refusal, API, and parse failures become `judge_failed` rows with null scores.
4. Report	Read `.summary.json`; report planned and successful image counts, failure causes, image and prompt coverage, and language, level, and category aggregates.	Averages run over successful images; a prompt-macro mean averages within each prompt first. No implicit zero, 50, or completion threshold is applied.

Step 1 command example: generation.

```
python src/generation/sample_zimage.py \
  --model_path <MODEL> \
  --prompt_file data/<lang>_prompts.jsonl \
  --output_dir <run> --steps <N> \
  --num_images_per_prompt 4
```

Step 3 command example: strict Judge.

```
python scripts/eval_vlm_judge_api.py \
  --sample_dir <images> \
  --prompt_file data/<lang>_prompts.jsonl \
  --output_file <judge.jsonl> \
  --model_name <served_model>
```

Failure handling. The evaluator never turns a failed request or a malformed response into a score. Every planned prompt and sample row is kept, carrying a status such as `missing`, `invalid_artifact`, or `judge_failed`, and a failed row has null score fields and a typed reason for the failure. The raw response and the attempt errors are preserved with it. A safety refusal is treated as a non-blocking observation about the dataset and is counted as a failure state rather than converted into a quality score.

A complete bilingual run executes Steps 1–3 once for `en` and once for `zh`. At 216 prompts and four images each, a complete language split contains 864 images and the bilingual plan contains 1,728 rows per model configuration; unavailable outputs reduce coverage and receive no imputed quality score. Other generators are driven by their own scripts in the same package, with model-specific arguments, but each must emit the same binding and status fields before its images reach the judge.

D SCORING AND AGGREGATION

Every raw dimension is returned on a 0–100 scale. The judge contract names each dimension and gives its scale, but it describes no intermediate bands, so Table A2 is a reading aid we added afterwards rather than a transcription of the prompt. It exists to keep interpretation consistent and to give future annotators a common vocabulary, and it is never injected into the judge prompt.

The rubric asks TA to assess character and word correctness and TC to assess the presence of every region with its entire required content. It does not specify an edit-distance calculation, a region-matching procedure, or how incorrect but present text contributes to TC. These remain judgments of the VLM, so a high TC score alone does not establish correct reproduction. For example, Figure A5 records TA=10 and TC=100, which become Fidelity=55. Keeping both raw scores makes that difference visible. The composite supplies no pass threshold: even a hypothetical TA=0 with all five other scores at 100 would produce a Composite of 70.

Table A2: Interpretive score bands. Notes: These qualitative descriptions are a reading aid rather than calibrated thresholds on character accuracy or region recall, and they form no part of the judge prompt.

Text-side diagnostics					
Dimension	**Severe 0–19**	**Weak 20–39**	**Partial 40–59**	**Strong 60–79**	**Excellent 80–100**
TA Accuracy	Little or no correct text	Isolated correct fragments	Mixed correct and wrong text	Minor character errors	Near-exact transcription
TC Completeness	Little required text	Many missing regions	Substantial omissions	Minor omissions	Near-complete coverage
TR Readability	Illegible strokes	Few readable fragments	Mixed legibility	Mostly clear	Consistently legible

Spatial/scene diagnostics					
Dimension	**Severe 0–19**	**Weak 20–39**	**Partial 40–59**	**Strong 60–79**	**Excellent 80–100**
PC Position	Unrelated placement	Mostly misplaced	Mixed position match	Mostly correct cells	Strong grid agreement
LQ Layout	Disordered layout	Major spacing and alignment errors	Inconsistent organization	Mostly coherent layout	Clear hierarchy and alignment
SI Integration	Detached from scene	Major surface or perspective mismatch	Uneven scene fit	Minor integration artifacts	Natural surface integration

The raw scores yield text fidelity $F = (\mathrm{TA} + \mathrm{TC})/2$, text clarity $C = \mathrm{TR}$, spatial quality $S = (\mathrm{PC} + \mathrm{LQ})/2$, and scene quality $Q = \mathrm{SI}$, with composite

$$\mathrm{Composite} = 0.60F + 0.30C + 0.05S + 0.05Q. \tag{3}$$

Only a row with a strict successful judge response enters a quality mean. Missing images, invalid artifacts, generation failures, safety refusals, parser and API failures, and context-overflow rows stay in the coverage denominators, and none of them ever becomes an implicit zero or an implicit 50.

D.1 WORKED EXAMPLE FROM A RECORDED RESPONSE

For sample 1 of `A1_L1_EN_001`, the judge response reads:

```
{"text_accuracy":85,"text_completeness":95,"text_readability":95,
 "position_correctness":90,"layout_quality":95,"scene_integration":98}
```

This gives $F = 90$, $C = 95$, $S = 92.5$, and $Q = 98$, and therefore

$$0.60(90) + 0.30(95) + 0.05(92.5) + 0.05(98) = 92.025.$$

Because our implementation rounds each image's composite to one decimal before aggregating, the value stored is 92.0, which is displayed as 92.00 wherever two decimal places are used; aggregate summaries keep three decimals internally. The image itself, and the carrier detail behind its integration score, appear in Figure A8. This example demonstrates aggregation of the recorded scores. The separate manual review and its scope are described in Section 3.2.

Failure record: illustrative schema excerpt. A parser failure produces an evaluator row with null scores rather than a valid judge response. The excerpt below is constructed for explanation, and it omits the error-detail fields that a recorded row carries.


```
{"status":"judge_failed",
↪  "failure_type":"judge_api_or_response_invalid",
 "text_accuracy":null,"text_completeness":null,"text_readability":null,
 "position_correctness":null,"layout_quality":null,
 ↪  "scene_integration":null,
 "text_fidelity":null,"text_clarity":null,"spatial_quality":null,
 "scene_quality":null,"composite":null}
```


D.2 Quality Means and Coverage: A Numerical Example

Illustrative data. Suppose four images are planned for each of two prompts. Prompt A returns two successful rows, with composites 80 and 100, and prompt B returns one, with composite 20; the remaining five planned rows are missing or failed.

$$\begin{aligned}\text{Successful-image mean} &= (80+100+20)/3 = 66.67,\\ \text{Prompt-macro mean} &= \big[(80+100)/2+20\big]/2 = 55.00,\\ \text{Image coverage} &= 3/8 = 37.50\%,\\ \text{Prompt coverage (at least one success)} &= 2/2 = 100.00\%.\end{aligned}$$

The two means differ because prompt-macro aggregation gives each represented prompt equal weight, whereas the image mean lets A count for twice as much as B. A prompt with no successful image contributes to the coverage denominator but has no quality mean to contribute at all. Coverage and quality therefore answer different questions and have to be read together. The bilingual leaderboard then averages the EN and ZH prompt-macro means equally.

D.3 Scope of the Reported Aggregates

Every row of Table 4 is presented the same way: the first five score columns are an equal-language mean over that model's EN and ZH prompt-macro aggregates, and the six level columns keep the language-specific composites. A missing cell records a measurement that is unavailable and must not be read as a score of zero. Complete-reference input and a valid six-key response do not by themselves establish that every region was assessed correctly. The current output has no region-level comparison or transcription with which to verify that coverage. PC rates coarse positions. The grid provides neither within-cell ordering nor a complete description of spatial relations between regions.

E Region References and Benchmark Interpretation

E.1 What the Region Reference Contributes

The structured reference connects each requested string to its intended role in the image. A list of words can establish which content should appear, while the region description also specifies where that content belongs and what should carry it. These requirements matter when a correct title appears above an incomplete menu, when a price is detached from its item, or when readable opening hours appear on the wrong side of a storefront. The reference makes such distinctions available to the evaluator across the benchmark's scene categories.

Text accuracy and completeness. The target string defines the content against which TA and TC are judged. A region may contain a single heading or a multi-line passage, so its presence does not establish that its entire content was reproduced. In the bakery record, the shop name, tagline, menu, and hours plate are four separate requirements. Correctly rendering the shop name leaves the other three to be checked. The packaging example in Figure A4 makes the same point from the opposite direction: detailed ingredient text does not replace the missing product title.

Position and typographic organization. The grid position specifies a coarse location for PC, while relative size and the scene description provide context for visual hierarchy and LQ. Several regions can occupy one cell. In the bakery example, the shop name and tagline both belong at `top-center`, although the title is larger and the tagline is below it. The grid alone does not encode that within-cell ordering. Identifying the hours plate requires reading its text; checking its placement requires the whole scene. In Figure A6, the plate is at bottom-left rather than the specified bottom-right.

Readability and carrier fit. TR concerns whether the rendered characters are visually legible. SI concerns their relation to the carrier, including perspective, lighting, and material. A sharply drawn word can be misspelled, and correctly transcribed text can still look detached from its surface. The wooden sign in Figure A8 illustrates lettering that follows its carrier's plane and texture. A crop exposes local character shapes, while the whole image shows their integration with the scene.

Reference attributes and reported scores. The six reference attributes describe the target scene, while the six raw scores describe the generated image. Their roles differ: `importance`, for example, is metadata and supplies no numerical weight. The rubric instructs the judge to assess lower-importance regions as well. Reference attributes guide the image-level judgments without defining a separate score for every attribute or region. The four reporting dimensions and Composite remain the mappings in Table 3; the protocol introduces no additional region-level metric.

E.2 Benchmark Value as Model Capabilities Improve

Improved rendering of short strings shifts attention toward whether models can preserve text across longer passages, multiple regions, and varied carriers. A receipt must retain its item-price associations, a form must preserve its fields, and a storefront must render both its prominent name and its smaller supporting text. UltraText Bench evaluates these demands with one rubric.

The comparisons in Table 4 reveal substantial variation within this setting. Qwen-Image-2512 receives an EN Composite of 86.50 at L1 and 42.86 at L3; Z-Image-Base receives 82.92 and 26.34, respectively. GPT Image 2 [Low] remains near the top of the scale in both languages at every level. These outcomes distinguish configurations that sustain high rated quality across workloads from those with large performance differences between the groups. Each group contains different prompts with varying text loads and numbers of regions.

The dimensions also distinguish configurations with similar overall scores. Qwen-Image-2512 and Boogu-Image-0.1-Turbo receive Composites of 67.96 and 67.79. Qwen-Image-2512 has higher Clarity, 79.89 versus 64.86, while Boogu-Image-0.1-Turbo has higher Fidelity, 66.27 versus 59.30. Their close overall scores thus summarize different balances between legibility and content reproduction. Alongside the workload breakdowns, these measurements help identify the capabilities a model retains and the aspects of dense text generation that remain difficult.

F Qualitative Examples and Diagnostic Reading

Each of the seven cases below pairs a whole image with a crop at native resolution, quotes the part of the reference that matters, and states what can be seen. All seven come from GPT Image 2 [Low], and every figure places the whole image at left and the outlined crop at right. An outline marks where a crop was taken and is an editorial annotation, not a GT box. Each score line reproduces the judge's whole-image response in the order TA, TC, TR, PC, LQ, SI, followed by the composite; these describe the whole image rather than the crop, and none is a human rating assigned here. The cases were chosen to show how the dimensions overlap in practice, and they include a positive integration example and disagreements at low and maximum scores. These selected cases do not estimate how often each failure occurs.





Whole-image judge scores: TA 15 TC 10 TR 85 PC 95 LQ 90 SI 95 **Composite 42.40**

Figure A3: Character errors on a plausible sign. `A1_L3_EN_002`, sample 1; 1,546 GT characters, 10 regions. **GT excerpt (`region_5`):** "GREENFIELD COMMUNITY NOTICE". **Observation:** The heading of the notice is built from malformed letters, even though the poster and the street around it stay entirely recognizable. The highlighted error occurs in that heading. **Dimensions:** TA, with TR also relevant.





Whole-image judge scores: TA 95 TC 80 TR 98 PC 95 LQ 95 SI 98 **Composite 91.60**

Figure A4: Missing title amid detailed packaging text. `C4_L1_EN_002`, sample 2; 563 GT characters, 5 regions. **GT excerpt (`region_0`):** "BOTANICA — Repair & Restore Shampoo". **Observation:** The bottle opens straight into product claims, and the prominent product title the reference asks for is absent. Neither the surrounding text nor the detailed ingredient list can stand in for that region. **Dimension:** TC.



Whole-image judge scores: TA 10 TC 100 TR 85 PC 100 LQ 95 SI 95 **Composite 68.10**

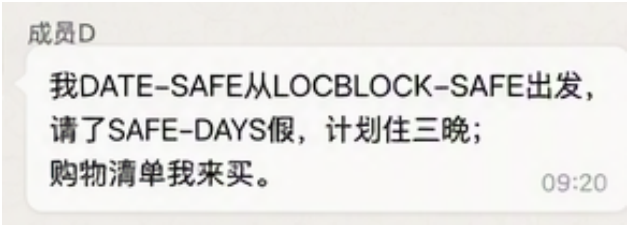
Figure A5: Dense Chinese glyphs require local inspection. `C1_L3_ZH_002`, sample 4; 611 GT characters, 8 regions. **GT excerpt (`region_3`):** "鸡腿葱串两串炭火现烤¥32". **Observation:** Seen at page scale the dish-and-price rows look well organized, but the glyphs are small enough that their strokes and exact wording become checkable only under enlargement. Tidy rows by themselves say nothing about text fidelity. **Dimensions:** TA and TR.



Whole-image judge scores: TA 10 TC 100 TR 95 PC 10 LQ 95 SI 95 **Composite 68.90**

Figure A6: Correct surface type at the wrong grid location. `A1_L2_ZH_001`, sample 1; 564 GT characters, 7 regions. **GT (`region_4`):** a small door plate at `bottom-right`, beginning "周一至周六6:30-19:00". **Observation:** The hours plate turns up in the bottom-left of the full image instead. The enlargement is what identifies the region, but its location can only be judged from the whole image. **Dimension:** PC.





Whole-image judge scores: TA 0 TC 0 TR 100 PC 0 LQ 100 SI 100 **Composite 37.50**

Figure A7: Readable chat content and a discrepant judge score. `F2_L3_ZH_002`, sample 4; 631 GT characters, 8 regions. **GT excerpt (`region_2`):** "成员D：我DATE-SAFE从LOCBLOCK-SAFE出发". **Observation:** Related text is plainly visible down the central chat column, with further messages and stickers filling the flanks, and yet the judge assigns TA=0 and TC=0 while giving LQ=100. Manual inspection identifies a disagreement between the visible content and the automated fidelity scores. The original scores are retained to make this failure case inspectable; they should not be interpreted as a human assessment that all requested text is absent. **Dimensions:** TC and LQ.





Whole-image judge scores: TA 85 TC 95 TR 95 PC 90 LQ 95 SI 98 **Composite 92.00**

Figure A8: Carrier integration as a separate property. `A1_L1_EN_001`, sample 1; 423 GT characters, 4 regions. **GT (`region_0`):** the "DAILY BREAD" title belongs on a wooden shop sign. **Observation:** The lettering follows the plane of the sign and picks up its lighting and surface texture. Integration is judged separately from whether the text itself is exact. **Dimension:** SI. The score calculation is given in Appendix D.1 .



Whole-image judge scores: TA 100 TC 100 TR 100 PC 100 LQ 100 SI 100 **Composite 100.00**

Figure A9: Maximum ratings despite degraded small text. `E4_L3_EN_003`, sample 2; 5,230 GT characters, 8 regions; original image 1024 × 1024 pixels. **GT excerpt (`region_7`):** "func TestParsePage(t *testing.T)" followed by the test cases and assertions. **Observation:** The bottom-right test block contains visibly deformed and difficult-to-read characters, despite maximum accuracy and readability ratings. The image hash matches its saved judge record. This selected case illustrates a local limitation of the automatic scores; it does not estimate how often maximum ratings overlook errors. **Dimensions:** TA and TR.



Figure A10: A local success can hide whole-scene failures. This illustrative six-region scene contrasts a correct shop name with errors elsewhere. *Left.* The correctly rendered shop name, cropped. *Center.* The full scene, with five further requested regions marked. Each is missing, wrong, unreadable, misplaced, or legible but poorly integrated with its intended surface. *Right.* Dense text load and whole-scene evaluation cover the requested text across all regions.

F.1 Category Galleries

Figure A11 and Figure A12 give one selected output per category at L1 and L2; the L3 gallery appears in Figure 1 in the main text. The galleries show the range of scenes in the taxonomy. The enlarged cases above provide the detail needed to inspect individual strings and their carriers.



Sign (EN/L1) | Label (ZH/L1) | Poster (ZH/L1) | Billboard (EN/L1)

Article (EN/L1) | Newspaper (EN/L1) | Letter (EN/L1) | Resume (ZH/L1)

Menu (ZH/L1) | Receipt (EN/L1) | Invoice (EN/L1) | Packaging (EN/L1)

Webpage (ZH/L1) | Slide (EN/L1) | Social Media (ZH/L1) | Dashboard (EN/L1)

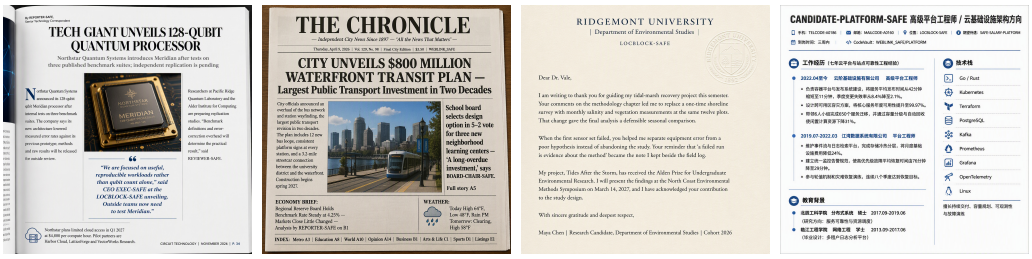
Schedule (EN/L1) | Form (ZH/L1) | Certificate (EN/L1) | Code (EN/L1)

Caption (EN/L1) | Dialogue (EN/L1) | Comic (ZH/L1) | Infographic (ZH/L1)

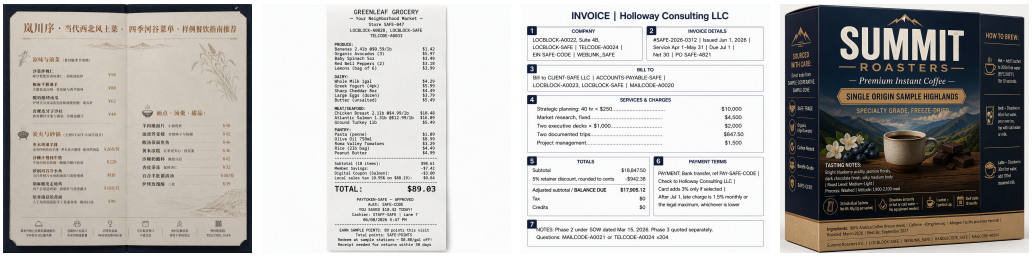
Figure A11: Scene coverage at L1 (Hard). One selected output per category in taxonomy order, from GPT Image 2 at API quality `Low`. EN/ZH denote English/Chinese. These examples illustrate scene diversity.



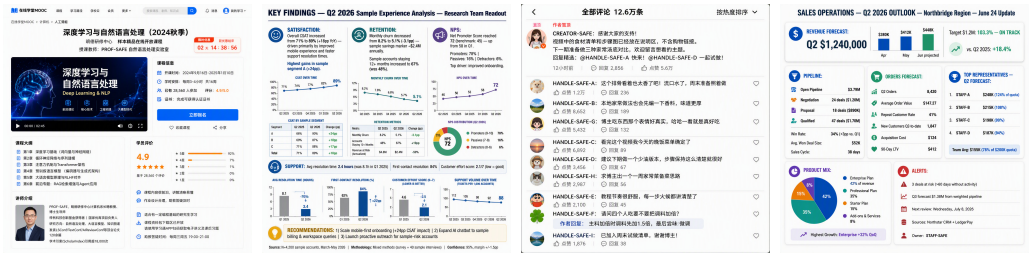
Sign (EN/L2) | Label (ZH/L2) | Poster (ZH/L2) | Billboard (EN/L2)



Article (EN/L2) | Newspaper (EN/L2) | Letter (EN/L2) | Resume (ZH/L2)



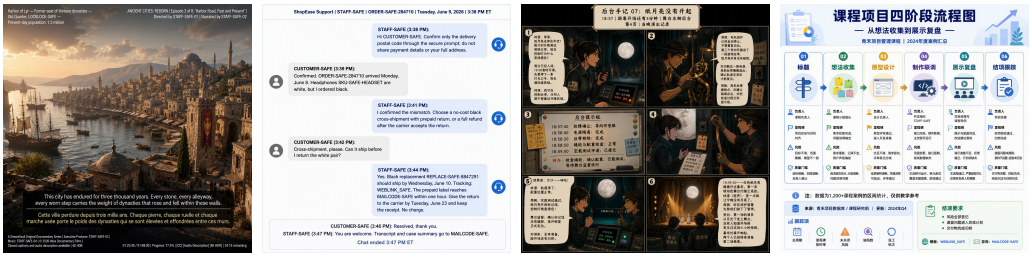
Menu (ZH/L2) | Receipt (EN/L2) | Invoice (EN/L2) | Packaging (EN/L2)



Webpage (ZH/L2) | Slide (EN/L2) | Social Media (ZH/L2) | Dashboard (EN/L2)



Schedule (EN/L2) | Form (ZH/L2) | Certificate (EN/L2) | Code (EN/L2)



Caption (EN/L2) | Dialogue (EN/L2) | Comic (ZH/L2) | Infographic (ZH/L2)

Figure A12: Scene coverage at L2 (Very Hard). One selected output per category in taxonomy order, from GPT Image 2 at API quality `Low`. EN/ZH denote English/Chinese. These examples illustrate scene diversity.

G RELEASED MATERIALS AND THEIR LIMITS

Material locations. The dataset and the evaluator are released in the benchmark code package; generated images and per-image judge records are released separately. End-to-end reproduction requires a versioned public archive linking these materials to the evaluated model configurations.

Released evidence. The complete EN and ZH prompt records, together with the released evaluator, support the dataset counts, the region inventories, the six raw metrics, strict response validation, typed failure states, and the four-dimension composite. The GPT Image 2 [Low] release adds images and per-image judge rows: 859/864 successful EN images over 215/216 prompts, and 864/864 successful ZH images over 216/216 prompts. Four missing EN images belong to `F3_L1_EN_003` and one to `A4_L2_EN_003`. The recorded scores reproduce the bilingual prompt-macro Composite of 99.35. All six raw dimensions equal 100 for 732/859 valid EN images and 708/864 valid ZH images, or 1,440/1,723 (83.6%) together. These counts describe automatic ratings, without assigning correctness labels to the images. The cases shown in Appendix F are drawn from these records.

Boundaries of the evidence. The GPT Image 2 cases selected here are qualitative examples rather than a representative sample of error rates. The released judge summaries identify `Qwen-Image-Bench`, but their `model_revision` and `model_artifact_sha256` fields are null. The recorded identity hash does not fix the checkpoint weights. These records therefore support checking score aggregation and image bindings, with incomplete checkpoint provenance. The GPT Image 2 records do not establish the coverage or score distribution of the other model configurations in the leaderboard.

Reporting consequence. Reproduction across all configurations requires their model revisions, resolutions, sampling settings, API quality settings and evaluation dates, together with images, raw judge responses, and the status of every planned row. Image and prompt coverage should accompany each quality mean. Comparing only shared prompts can assess how differences in coverage affect a pair of models; uncertainty estimates should account for images sharing a prompt. These further comparisons are not reported here.
The released dataset statistics, including the per-prompt character and region distributions, are reported in Table 2 and are not duplicated here.

H Related Work

H.1 Visual Text Generation Methods

Visual text generation methods differ in how they encode exact character sequences, how they obtain a layout, and how well text survives being encoded and decoded by the image model.

Explicit glyph and layout conditioning. GlyphControl (Yang et al., 2023) adds a ControlNet branch conditioned on a rendered glyph map, while GlyphDraw (Ma et al., 2023) combines glyph information with spatial control for Chinese and English text. TextDiffuser (Chen et al., 2023) instead predicts character-level layout masks before image synthesis, and TextDiffuser-2 (Chen et al., 2024) uses a language model to plan keywords and bounding boxes from open-ended prompts. These approaches tie a string more firmly to a place in the image, but they do not assume the same input: glyph-conditioned systems receive an explicit visual control, whereas layout-first systems must predict a layout whose errors can propagate to image generation.

Character-aware and multilingual modeling. AnyText (Tuo et al., 2024) combines glyph and position conditions with OCR-aware representations in one multilingual generation and editing framework. Glyph-ByT5 (Liu et al., 2024) introduces a byte-level text encoder to preserve character information lost to subword tokenization. JoyType (Li et al., 2024) targets multilingual typography, and ViType (Gao et al., 2026) jointly models semantic and glyph features in a multimodal diffusion model. Complementary work studies scene-text inpainting (Zhang et al., 2024), position-controlled synthesis (Zhao & Lian, 2024), and data synthesis and filtering (Zhao et al., 2025).

Dense and multi-region generation. Recent methods target longer strings and multiple text regions. TextCrafter (Du et al., 2025) steers attention toward the quoted target strings and adds a reinforcement signal read from OCR, improving how multiple strings are rendered in complex scenes. GlyphDraw2 (Ma et al., 2025) combines language-model planning with glyph-conditioned generation for multi-element posters, while PosterMaker (Gao et al., 2025) focuses on text-rich product posters. BizGen (Peng et al., 2025) addresses article-level infographic generation by binding text spans to planned regions, and TextGuider (Baek et al., 2025) applies training-free attention guidance to reduce omissions in longer text. GlyphAnchor (Xiang et al., 2026) anchors glyph priors to predicted positions, and TextGround4M (Mao et al., 2026) complements these methods with prompt-aligned text spans and layout annotations for training layout-aware models. Their inputs range from prompts alone to explicit glyph and layout controls, with evaluations tailored to each application.

H.2 Text Rendering Benchmarks and Datasets

Text-rendering resources differ in what they ask the model to produce, in what they tell it beforehand, and in how they check the result. Scores reported on short-string, poster, document, and long-text test sets therefore cannot be placed on one scale. Table A3 summarizes the closest comparisons.

Short-string fidelity and attributes. MARIO-Eval (Chen et al., 2023) established a large short-keyword test set. AnyText-Bench (Tuo et al., 2024) added English and Chinese generation tests whose OCR scores are computed on text lines cropped at the specified positions; the AnyText model also edits text, but the benchmark's quantitative protocol is generation-oriented. TypeScore (Sampaio et al., 2024) moved past a single OCR accuracy value: it extracts the rendered text and combines several string-comparison measures into one fidelity score. That score tells spelling errors, missing text, and extra text apart, and it is validated against human judgments. LeX-Bench (Zhao et al., 2025) adds prompts that also constrain color, font, and position, and scores them with PNED, which matches requested against recognized words and penalizes the ones left over. Because PNED treats words as an unordered set, it says nothing about reading order or about which surface a string ended up on. These resources primarily measure short-string accuracy and text attributes.

Dense and long-text generation. CVTG-2K and the associated TextCrafter multi-text setting (Du et al., 2025) evaluate more complex scenes and match recognized text instances to requested strings. Their prompt descriptions also encode positions, attributes, and the correspondence between text and its carrier. LongText-Bench, introduced with X-Omni, covers longer English and Chinese text across eight scenarios (Geng et al., 2025), and document-level rendering of a single long sequence has been examined separately (Zhang et al., 2025b). TextInVision (Fallah et al., 2025) crosses prompt complexity with text properties and also examines visual-encoding failures. The closest neighbor is InfoTextBench, released with GlyphAnchor (Xiang et al., 2026), which pairs English and Chinese

Table A3: Text-rendering benchmarks: inputs, target references, and evaluation. The table summarizes the main evaluation components; details appear in Section H.2 . **Notes:** EN: English; ZH: Chinese. Extra input denotes benchmark-supplied controls beyond the prompt, not internally predicted layouts. OCR: optical character recognition; VQA: visual question answering. PNED matches words by edit distance and penalizes unmatched items. †No image is supplied for T2I; source images are used for editing or image-to-image tasks. UltraText pairs per-region text and visual attributes with image-level VLM ratings.

Benchmark	Language	Extra input	Target reference	Evaluation
MARIO-Eval	EN	None	Target strings	OCR + image metrics
AnyText-Bench	EN/ZH	Glyph/position	Text + positions	OCR on text crops
LeX-Bench	EN	None	Text + style/position	PNED + attribute VQA
CVTG-2K	EN	None	Text instances	OCR + matching
LongText-Bench	EN/ZH	None	Target strings	VLM text accuracy
TextInVision	EN	None	Target strings	OCR edit distance
TextAtlasEval	EN/ZH	None	Scene + text	OCR + CLIP similarity
OCRGenBench	EN/ZH	None / image†	Task-specific targets	OCRGenScore
InfoTextBench	EN/ZH	None / image†	Target text list	OCR + VLM quality
UltraText Bench (ours)	EN/ZH	None	Region text + layout/carrier	Image-level VLM

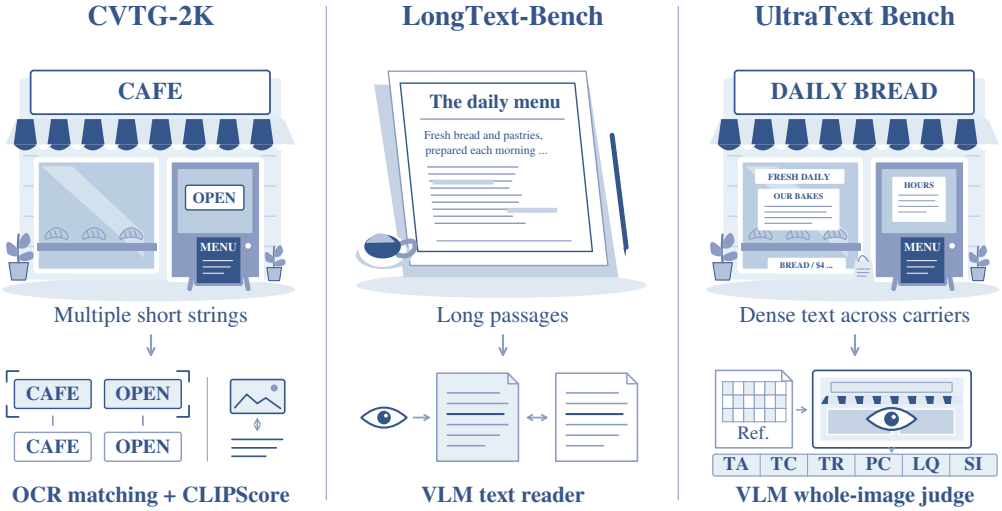


Figure A13: Task scenes and assessment in three benchmarks. *Left:* CVTG-2K (Du et al., 2025) matches OCR-recognized words to targets using Word Accuracy and normalized edit distance, and separately reports CLIPScore. *Center:* LongText-Bench (Geng et al., 2025) uses a VLM to extract text for comparison with target strings. *Right:* UltraText Bench uses a structured reference (Ref.) to rate the whole image: text accuracy (TA), completeness (TC), readability (TR), position correctness (PC), layout quality (LQ), and scene integration (SI). The scenes, abbreviated strings, and line marks are illustrations, not benchmark outputs or measured results; they do not exhaust each benchmark's scene types.

prompts over text-rich reference images, gives each one an explicit list of target strings, and scores word-level precision, recall, and phrase hits. Long targets, images holding many separate pieces of text, and bilingual coverage have therefore all been studied before. These resources differ in how target strings, attributes, positions, and scene context are represented and evaluated.

Broader visual-text tasks and editing. TextAtlas5M (Wang et al., 2025a) contributes large-scale training data for long and structured text, together with TextAtlasEval, an evaluation set spanning four design domains in which one subset varies text length under otherwise fixed conditions. OCR-GenBench (Zhang et al., 2025a) broadens evaluation to generation, editing, and OCR-oriented transformation tasks in English and Chinese, and explicitly includes high text density, page-level generation, and dense-document editing. VTPBench (Shu et al., 2025) unifies six visual text processing tasks under task-specific metrics and VTPScore, and OneIG-Bench (Chang et al., 2026)

contributes a bilingual text-rendering split inside a broader image-generation suite. Task-family aggregates summarize broad performance, while trends with text load and region count require separate breakdowns. A separate line of benchmarks targets text-centric *editing*, which requires modifying the target text while preserving other text and the background. TextEditBench (Gui et al., 2025) covers everyday document and signage scenes and adds a dimension for edits that depend on reasoning. WeEdit (Zhang et al., 2026a) extends editing to many languages and scores instruction adherence, text clarity, and background preservation. TextSculpt-Bench (Lin et al., 2026) compares the text that should appear in the whole image against the text that actually does, and checks the background outside the edited area. TextWand-Bench (Wang et al., 2026a) evaluates removal, generation, and replacement under explicit layout and style control. None of these is directly comparable to prompt-only generation, because the model is handed a source image and part of the reference is the input itself. They are still instructive here: TextSculpt-Bench in particular shows that checking all the text has to leave behind evidence that can be inspected, since handing an evaluator a complete reference does not prove that it looked at every region.

General T2I evaluation and our scope. General text-to-image benchmarks such as TIFA (Hu et al., 2023), T2I-CompBench (Huang et al., 2023), and HRS-Bench (Bakr et al., 2023) measure semantic faithfulness and composition over broader prompt families. BizGenEval (Li et al., 2026b) comes closer to our scene types, scoring slides, charts, webpages, posters, and scientific figures against human-verified checklist questions. Their checklists and alignment metrics address broader semantic and compositional constraints, complementing the explicit per-region text references used here. UltraText Bench combines prompt-only generation, a uniform per-region reference, and coverage of 24 text-bearing scene categories. The generator receives all target strings in natural language; the six-field annotation is reserved for evaluation and stays consistent across scenes, languages, and levels. This design tests whether text rendering transfers across carriers and layouts under increasing text load. The reference specifies coarse spatial metadata, and the protocol returns image-level scores; region-level attribution remains a future direction.

H.3 Automatic Evaluation and Judge Reliability

Global and OCR-based metrics. Automatic generation evaluation spans global alignment and preference metrics, fidelity measures from OCR output, and VLM judges that follow a written rubric. CLIPScore (Hessel et al., 2021) measures global image-text similarity, while ImageReward (Xu et al., 2023), HPSv2 (Wu et al., 2023), and PickScore (Kirstain et al., 2023) learn scalar preferences from human comparisons. These scores are useful for semantic alignment or overall preference, but their training objectives do not require transcribing every requested character and can favor an attractive image whose text is incorrect. Text-specific metrics provide a closer signal. TypeScore (Sampaio et al., 2024), described above, also reports a failure mode on the evaluator's side: OCR can introduce recognition errors, while a VLM reader can silently correct characters that the image actually renders wrongly. OCRGenScore (Zhang et al., 2025a) combines metrics across OCR-generative tasks, and VTPScore (Shu et al., 2025) uses a multimodal model to rate visual quality and text readability across visual text processing tasks. OCR-based scores depend on the detector, recognizer, language settings, font, scale, and image domain (Baek et al., 2019; Cui et al., 2025).

Text quality and spatial fidelity. A complementary line evaluates the visual quality of rendered text separately from its correctness. TIQA (Koltsov et al., 2026) collects human quality ratings for rendered text, on crops and on text-heavy images, and trains a model to predict them; this is what our clarity dimension tries to capture. Because it rates regions that were detected, it cannot by itself report a region that was requested and never rendered. Recognized strings alone also do not establish whether text occupies the requested region or integrates naturally with its carrier.

Rubric-based VLM judges. Prompted VLM evaluators can apply task-specific rubrics without training a separate metric for every dimension. VIEScore (Ku et al., 2024) scores semantic consistency and perceptual quality with explanations, and UniGenBench++ (Wang et al., 2025b) uses structured semantic units for multidimensional T2I evaluation. Qwen-Image-Bench (QIB) (Li et al., 2026a) develops fine-grained creation rubrics and a dedicated Q-Judger using professional human ratings. RichHF (Liang et al., 2024) likewise demonstrates the value of fine-grained human feedback tied to words and image regions. T2LSC-Bench (Wang et al., 2026b) pairs OCR verification with a structured VLM judgement, separating whether the text landed on its intended anchor from whether its meaning leaked into the surrounding subject and scene, and it checks the automatic labels against a human-annotated subset. Evaluator version, prompting, text-reading ability, and visual-quality bias

can affect VLM scores. Prior studies examine agreement with independent human ratings (Zheng et al., 2023), controlled annotation protocols for T2I evaluation (Otani et al., 2023), and variation in text-reading ability across multimodal models (Liu et al., 2023; Fu et al., 2026). These limitations motivate joint assessment of text, placement, and scene integration, while keeping response-format validity distinct from the reliability of a judge's assessment.

I DISCUSSION

Why dense text is different from ordinary image quality. Diffusion models (Ho et al., 2020) operate in continuous visual spaces where approximate similarity is usually acceptable: small geometric or texture errors often preserve the intended object. Text is different because it is discrete and compositional; one wrong character can invalidate a word, code token, phone number, price, or form field. Self-attention in DiT (Peebles & Xie, 2023) and UNet (Rombach et al., 2022) architectures must split capacity between global composition and many local character patterns, and dense prompts spread this burden across multiple text regions. Text-annotated image pairs are rarer than ordinary image-caption pairs in web corpora (OpenAI, 2023), and incidental scene text is often noisy or weakly described. CLIP (Radford et al., 2021) and T5 (Raffel et al., 2020) use subword tokenization, while accurate text rendering requires preserving the exact character sequence. VAE latents can lose small stroke details, a limitation also emphasized by OCRGenBench (Zhang et al., 2025a). Visual-tokenizer evaluations further show that standard reconstruction metrics can conceal losses in text-bearing regions (Wu et al., 2025b). These mechanisms may contribute to differences between text fidelity, layout, and scene quality. Their causal effects remain untested here.

Training directions for stronger text rendering. Potential training directions for dense text rendering include visual representation warmup, as in ERW (Liu et al., 2025), and optimization with text-specific feedback. The structured annotations can support region-level diagnosis and feedback on character errors, omissions, placement, and scene integration. Such feedback can guide reward-based optimization, including DDPO (Black et al., 2024), ReFL (Xu et al., 2023), and AlignProp (Prabhudesai et al., 2023), or the construction of preference pairs used by methods such as Diffusion-DPO (Wallace et al., 2024) and FlowCPO (Han et al., 2026). Although `importance` does not enter our scoring formulas, future training objectives could use it to weight errors in a shop name more heavily than errors in a footnote. This direction is aligned with region-level feedback work such as RichHF-18K (Liang et al., 2024), but focuses specifically on visual text rendering.

Complementarity with concurrent benchmarks. UltraText Bench combines scene coverage and per-region references. STRICT (Zhang et al., 2025b) also evaluates thousands of characters, but varies one sequence on a plain page; UltraText Bench distributes text over 4–12 annotated regions with carriers and positions across 24 categories. OCRGenBench (Zhang et al., 2025a) spans several OCR-generative tasks, including dense pages, while UltraText Bench focuses on prompt-only generation. InfoTextBench (Xiang et al., 2026) is the closest comparison for dense bilingual text, but its target-text items differ from our regions, preventing direct comparison of their counts.

Limitations. Results reflect each model's standard generation settings, which can differ in resolution, aspect ratio, and API behavior. A controlled comparison of these factors remains future work. Each category$\times$level$\times$language cell contains 3 prompts and 12 images in a complete run. A language–level aggregate contains 72 prompts, and a language aggregate contains 216. Repeated images share a prompt, so prompt-level variation is the relevant unit for comparing aggregate scores. We report descriptive means without confidence intervals; close ranks and fine-grained cell differences require correspondingly cautious interpretation. The protocol returns image-level scores from one VLM judge. Ten participants took part in human evaluation of the automatic scores. Quantitative inter-rater agreement and high-score alignment statistics are not reported here. The per-region references specify the content and visual context for assessment, while the current output summarizes each image as a whole. Maximum scores are frequent in the GPT Image 2 records, and selected cases show both low-score disagreements and visible errors despite maximum ratings. The examples distinguish local character defects from an image's overall rating; they do not quantify the judge's error rate. A maximum rating is not a measured percentage of correct characters. Strong aggregate performance and occasional local disagreements therefore need to be interpreted at their respective scales. The language splits use independently written content and different character-load bands, and difficulty levels change text load and region count together. Their comparisons therefore do not isolate language or text length alone.

Future work. Future work will explore region-level evaluation and structured annotations as training signals. Extending the benchmark to video generation and editing systems such as Vidu S2 (Zhang et al., 2026b) would enable evaluation of text persistence and temporal consistency.