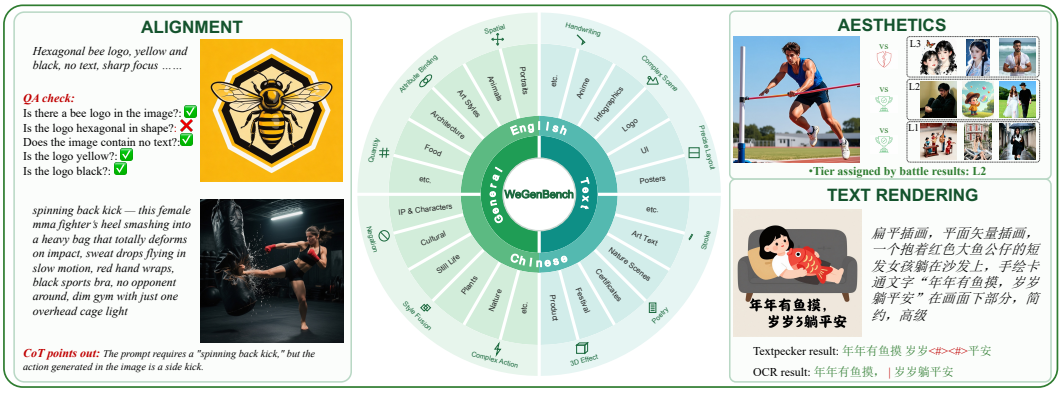
# WEGENBENCH: A MULTIDIMENSIONAL DIAGNOSTIC BENCHMARK TOWARDS TEXT-TO-IMAGE MODEL OPTIMIZATION

**Anonymous authors**
Paper under double-blind review




## ABSTRACT

Recent text-to-image models achieve impressive visual fidelity, yet existing benchmarks often summarize model performance using coarse aggregate scores and provide limited guidance for targeted improvement. We introduce WeGenBench, a bilingual, multidimensional benchmark for diagnosing text-to-image generation capabilities. It contains 4,000 prompts across general image generation and visual text rendering, evenly balanced between Chinese and English. In addition to scenario labels, each prompt is annotated with fine-grained capability tags that capture challenges such as spatial relations, attribute binding, complex actions, negation, typography, and bilingual rendering. Aggregating results by scenario and capability tag reveals localized failure patterns and suggests hypotheses for targeted data curation and post-training. We further develop interpretable vision-language model (VLM)-based evaluation protocols for semantic alignment and aesthetic quality, together with OCR- and VLM-based metrics for visual text rendering. The semantic-alignment and aesthetic-quality protocols provide structured rationales and dimension-level diagnostics, showing stronger or comparable agreement with human judgments than the evaluated baselines on our validation set. Finally, we benchmark a broad collection of recent systems, revealing compositional, linguistic, and typographic bottlenecks among the evaluated models that can be obscured by aggregate evaluation.

## 1 INTRODUCTION

Recent text-to-image (T2I) models have made rapid progress in image fidelity, semantic alignment, and controllability Podell et al. (2023); Wu et al. (2025); Esser et al. (2024); Z-Image Team (2025); Black Forest Labs (2025); Cao et al. (2025). Their evaluation, however, remains fragmented. Broad benchmarks often emphasize aggregate scores across diverse prompts Yu et al. (2022), whereas specialized benchmarks focus on individual capabilities such as compositional alignment Huang et al. (2023) or visual text rendering Tuo et al. (2023). Many widely used resources are also English-centric, providing limited coverage of language-specific challenges such as Chinese character structure, mixed-script rendering, and dense English typography.



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.

We propose **WeGenBench**, a bilingual and multidimensional diagnostic benchmark for T2I generation. It contains 4,000 prompts evenly divided between Chinese and English and between general image generation and visual text rendering. Each prompt has one scenario label and multiple fine-grained capability tags. Shared high-level capability dimensions, implemented through language-specific tag vocabularies, support both broad scenario evaluation and localized failure analysis across semantic alignment, aesthetic quality, and visual text rendering. Tag-level results also support hypotheses about targeted optimization by indicating which capabilities may benefit from additional data or post-training, including negation following, multi-entity attribute binding, complex actions, and layout-aware text rendering. A detailed comparison with prior benchmarks is provided in the supplementary material.

We further develop interpretable evaluation protocols tailored to these dimensions. Semantic alignment is assessed by complementary checklist-based QA and holistic CoT deduction. Aesthetic quality is measured through adaptive level-wise anchor matching with explicit structural constraints. Visual text rendering is evaluated through OCR- and VLM-based metrics that capture spelling, completeness, extraneous or incorrect characters, and readability. These metrics provide structured diagnostic evidence in addition to aggregate scores; for semantic alignment and aesthetic quality, we also validate their agreement with human judgments.

Finally, we benchmark a broad collection of open- and closed-source T2I systems. The resulting profiles reveal capability-specific trade-offs that are obscured by aggregate scores: semantic-alignment and exact-text-accuracy scores can diverge substantially, long and compositional prompts remain difficult, and performance differs across languages and typographic settings.

Our contributions are threefold: (1) a 4,000-prompt bilingual benchmark with scenario labels and capability tags that connect model failures to candidate optimization directions; (2) an interpretable evaluation suite for semantic alignment, aesthetic quality, and visual text rendering, with the semantic-alignment and aesthetic-quality metrics validated against human judgments; and (3) a systematic diagnosis of current T2I models across these dimensions.

## 2 WeGenBench

WeGenBench divides its test set evenly between General Image Scenarios and Visual Text Rendering Scenarios. Each category is further split into 1,000 Chinese and 1,000 English prompts, yielding 4,000 prompts in total. The two language subsets were constructed independently rather than by direct translation. Each prompt has one scenario label and multiple capability tags, enabling both application-level evaluation and fine-grained diagnosis.

### 2.1 WeGenBench-General

WeGenBench-General contains 2,000 curated prompts, evenly divided between Chinese and English, for semantic-alignment and aesthetic-quality evaluation. Rather than translating one subset into the other, we construct both languages independently under a shared design objective. The prompts cover 16 non-overlapping scenarios: Portraits & People, Cultural Elements, Art Styles, IP & Iconic Characters, Animals, Landmarks & Architecture, Food & Cuisine, Icons & Logos, Still Life & Commercial, Vehicles & Machinery, Nature & Weather, Counter-Intuitive scenes, Plants, Camera & Composition, Materials & Textures, and Interior & Home. Scenario proportions reflect usage frequency while retaining a minimum number of difficult samples in each category.

The benchmark also stresses stylistic and linguistic diversity. The Chinese and English subsets contain 135 and 143 distinct styles, respectively, combining frequent styles such as illustration and watercolor with long-tail artistic traditions. Chinese prompts average 93.6 characters (median 73), whereas English prompts average 30.8 words (median 27). This range includes both short instructions and long, constraint-dense prompts.

Each prompt receives multiple capability tags covering fine-grained attributes, spatial relations, attribute binding, complex actions, styles, and negation. The Chinese subset uses 21 open-ended tag phrases and averages 2.78 tags per prompt; the English subset uses 13 controlled tags and averages 2.69. Because tags recur across scenarios, tag-level scores can indicate whether a weakness is broad or

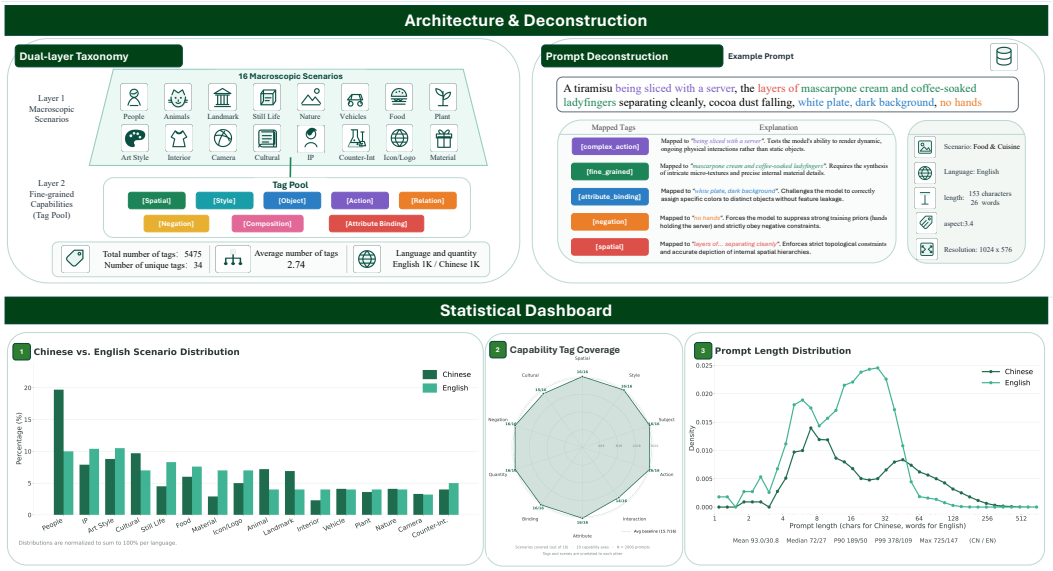



Figure 1: **Overview of WeGenBench-General. Top left:** The taxonomy of 16 scenarios and multidimensional capability tags. **Top right:** An example showing how a prompt is mapped to capability tags, including spatial relations, negation, and attribute binding. **Bottom:** Dataset size, language and prompt-length distributions, scenario distributions, and representative tag coverage across scenarios.

context-specific. Developers can rank weak tags, inspect associated samples, and prioritize candidate training slices or post-training objectives.

We additionally organize prompts into five difficulty levels according to the number of visualizable concepts and the density of capability constraints. L1–L2 cover minimal and basic prompts, L3 combines several concepts with one or two constraints, and L4–L5 contain dense combinations or challenges such as counter-intuitive reasoning, negation, complex interactions, and intricate layouts. L4 and L5 account for 57.3% of WeGenBench-General, making the set a deliberate stress test for current systems.

### 2.2 WeGenBench-Text

WeGenBench-Text contains 2,000 prompts, evenly divided between Chinese and English, across ten practical scenarios: natural scenes, artistic posters, product commercials, logos and charts, user interfaces, anime illustrations, popular-science diagrams, certificates and medals, artistic typography, and festival scenes. The required text averages 21.5 characters for Chinese prompts and 22.4 characters for English prompts, with lengths ranging from short labels to dense, multi-segment layouts.

The benchmark covers Chinese, English, digits, and mixed writing systems, together with handwriting, bold text, three-dimensional effects, and artistic typography. Its capability tags describe precise positioning, controlled line breaks and spacing, and complex visual conditions such as clutter, occlusion, perspective, and strong lighting. Precise positioning occurs in 60.4% of Chinese prompts and 18.4% of English prompts, while precise typography occurs in 22.4% and 31.6%, respectively. Complex scenarios account for roughly one quarter of each language subset.

Prompts are assigned Easy, Medium, Hard, or Extreme difficulty according to text length, segment count, and spatial constraints. Chinese prompts concentrate in the Medium and Hard tiers because short strings can still require complex character structure; English prompts are more dispersed but include a larger Extreme share due to deliberately long strings and compound typographic constraints. Tag-level scores measure text accuracy on samples with layout, handwriting, or bilingual requirements, helping identify candidate needs in character supervision, spatial text planning, and bilingual training data.

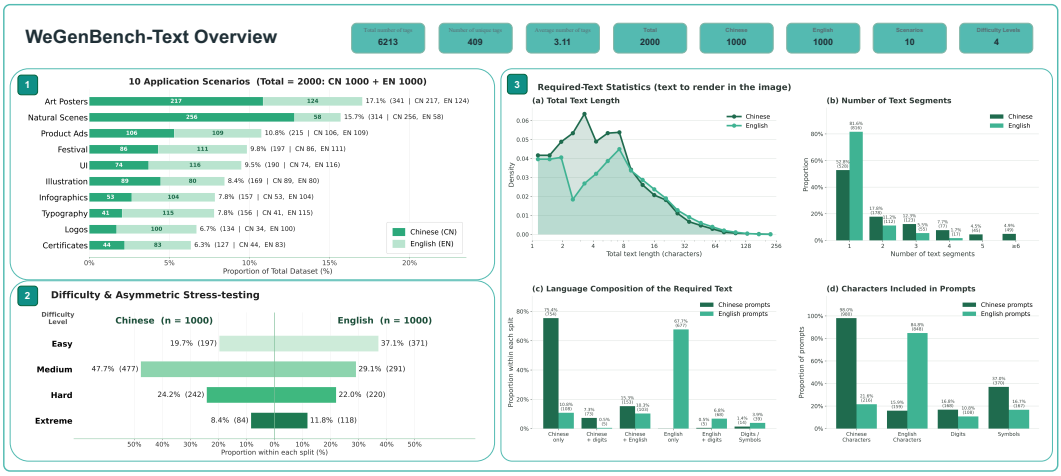



Figure 2: **Overview of WeGenBench-Text.** The dashboard summarizes (1) prompt distributions across ten scenarios, (2) difficulty distributions for Chinese and English prompts, and (3) properties of the required rendered text, including language, length, and style.

# 3 Evaluation Metrics

## 3.1 Semantic Alignment

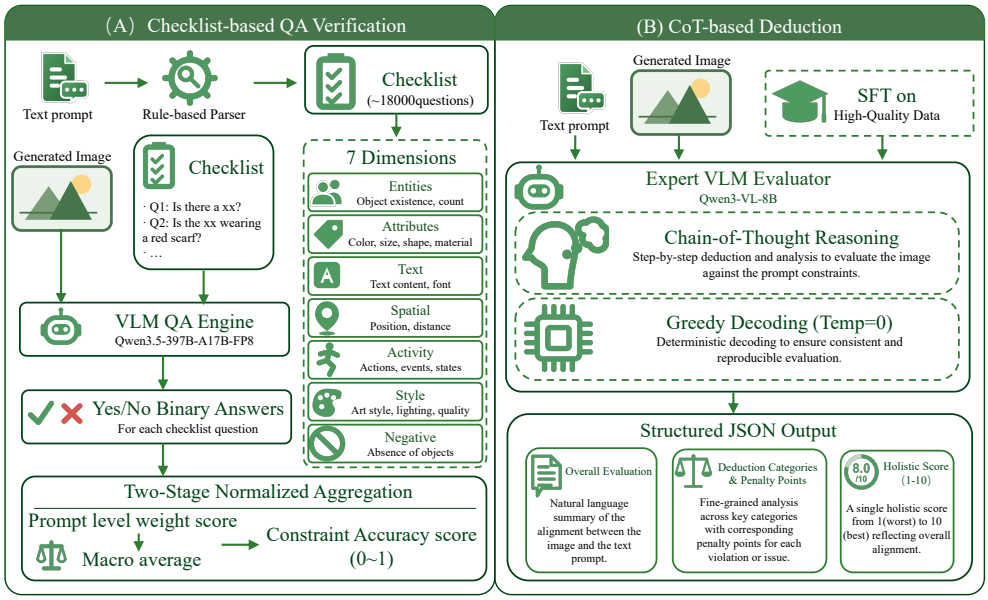



Figure 3: **Dual-Track Semantic-Alignment Evaluation Framework. (A) Checklist-based QA Verification** parses the prompt into a multidimensional checklist spanning seven semantic dimensions. A VLM answers each binary question based on the generated image, producing a Constraint Accuracy score through two-stage normalized aggregation. **(B) CoT-based Deduction** uses an expert VLM evaluator fine-tuned on human-annotated reviews to produce a structured chain-of-thought-based assessment. With greedy decoding, it outputs an overall evaluation, category-specific deductions and penalties, and a holistic score from 1 to 10.

We use two independent and complementary VLM-based metrics. **Checklist-based QA** decomposes each prompt into weighted binary questions over seven dimensions: entities, attributes, text constraints, spatial relations, activity, visual style, and negative constraints. Nearly 18,000 fixed

questions are evaluated independently, normalized within each prompt, and macro-averaged across prompts. This produces stable dimension- and tag-level diagnostics, although isolated checks may miss global inconsistencies.

**CoT-based Deduction** instead evaluates the image holistically under a unified deduction rubric. A VLM fine-tuned on human-annotated expert reviews outputs a structured assessment, category-specific penalties, and a score from 1 to 10 using deterministic decoding. Candidate checkpoints are selected against human judgments using $\pm 1$ accuracy, category F1, and deduction MAE. Additional protocol and alignment-pipeline details are given in the supplementary material.

### 3.2 Aesthetic Quality

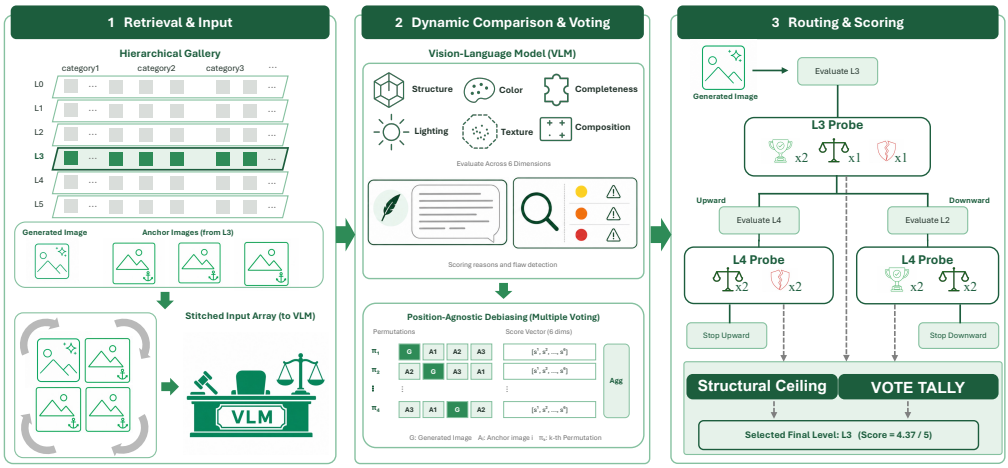



Figure 4: **Adaptive Level-wise Anchor-based Match Grading for aesthetic quality.** At each probed level, one input and three retrieved anchors are jointly evaluated under four positional permutations. Weighted results across levels and six dimensions produce the final score, subject to a structural ceiling.

Our **Adaptive Level-wise Anchor-based Match Grading** jointly compares each generated image with three anchors under four positional permutations. Starting at L3, it adaptively probes higher or lower levels and aggregates structure, texture, lighting, color, composition, and completeness across all probed levels into a continuous score. A structural ceiling limits scores when anatomical or physical defects are detected; additional methodological details appear in the supplementary material.

### 3.3 Visual Text Rendering

Given target segments GT and recognized segments Pred, we first use normalized edit distance (NED) and minimum-cost assignment to align multi-segment text despite OCR splitting or ordering differences. We then report five complementary metrics: **Edit Similarity** ($\mathrm{Sim}_{\mathrm{edit}} = 1 - \mathrm{NED}(\mathrm{GT}_{\mathrm{joined}}, \mathrm{Pred}_{\mathrm{joined}})$) measures global spelling agreement after minimum-cost concatenation. **Sentence Accuracy** ($\mathrm{Acc}_{\mathrm{sen}}$) applies Hungarian segment matching and reports the fraction of target segments matched exactly, providing a strict measure of complete sentence rendering. **Quality/Readability** ($\mathrm{Score}_{\mathrm{qua}} = 1 - N_{\mathrm{abnormal}}/N_{\mathrm{pred}}$) uses TextPecker Zhu et al. (2026) to measure visual legibility independently of semantic correctness. **GNED** aligns character tokens flexibly and computes one minus the unmatched-token ratio, capturing character agreement while tolerating segment reordering. Finally, **character Precision, Recall, and F1** derive true positives, false positives, and false negatives from matches, substitutions, insertions, and deletions after edit alignment. Precision is primarily sensitive to extraneous characters, Recall to omissions, and substitutions affect both; F1 summarizes the two. Together, these metrics separate readable but misspelled text from exact rendering and diagnose whether errors arise from additions, substitutions, or omissions. Equations and additional implementation details are provided in the supplementary material.

# 4 EXPERIMENTS

## 4.1 VALIDATION OF EVALUATION METRICS

### 4.1.1 HUMAN GROUND TRUTH CONSTRUCTION AND CORRELATION ANALYSIS

For each of the 2,000 WeGenBench-General prompts, we sample outputs from two different models and ask human experts to evaluate aesthetics and semantic alignment independently. Aesthetic judgments use forced win/loss choices, while semantic-alignment judgments additionally allow ties.

We compare automated metrics with these judgments at both pair and model levels. Pair AUC and Decisive Accuracy measure agreement on decisive comparisons; Spearman $\rho$, Kendall $\tau$-b, and MARD compare model rankings obtained from mean per-image scores with the human leaderboard. Detailed annotation instructions, tie handling, and aggregation procedures are provided in the supplementary material.

Table 1: Validation of evaluation metrics against human judgments. Pair AUC and Decisive Accuracy measure agreement on decisive, non-tied comparisons; model-level $\rho$, $\tau$-b, and MARD measure agreement with the human leaderboard. Among the evaluated methods, the best proposed metrics obtain the highest displayed correlations and lowest MARD. Best results in each category are highlighted in **bold**.

| Method | Pair AUC ↑ | Decisive Acc. ↑ | Model $\rho$ ↑ | Model $\tau$-b ↑ | Model MARD ↓ |
|---|---|---|---|---|---|
| **Aesthetic Quality & Preference** | | | | | |
| ImageReward Xu et al. (2023) | 0.51 | 0.52 | 0.05 | -0.01 | 5.25 |
| HPSv2 Wu et al. (2023) | 0.54 | 0.53 | 0.24 | 0.23 | 4.88 |
| HPSv3 Ma et al. (2025) | 0.60 | 0.57 | 0.24 | 0.19 | 4.75 |
| **Anchor-based Match (Ours)** | **0.63** | **0.61** | **0.83** | **0.68** | **2.06** |
| **Semantic Alignment** | | | | | |
| FG-CLIP 2 Xie et al. (2025) | 0.63 | 0.58 | 0.58 | 0.44 | 2.81 |
| VQAScore Lin et al. (2024) | 0.70 | 0.65 | 0.87 | 0.71 | 1.88 |
| **Checklist-based QA (Ours)** | 0.73 | 0.66 | 0.87 | 0.71 | 1.75 |
| **CoT-based Deduction (Ours)** | **0.75** | **0.71** | **0.91** | **0.78** | **1.25** |

As shown in Tab. 1, the proposed methods achieve higher agreement with human judgments than the evaluated baselines on this validation set. For aesthetic evaluation, **Anchor-based Match Grading** reaches a model-level Spearman $\rho$ of $0.83$, compared with $0.24$ for HPSv3 and HPSv2, and reduces MARD to $2.06$. For Semantic Alignment, **CoT-based Deduction** obtains the highest values across the five reported measures, including a model-level Spearman $\rho$ of $0.91$ and Pair AUC of $0.75$. **Checklist-based QA** also improves Pair AUC and MARD over VQAScore. These results support the proposed metrics as interpretable evaluators that correlate with human judgments on our validation set.

### 4.1.2 QUALITATIVE VALIDATION OF METRICS

Figure 5 presents cases consistent with the higher measured agreement of our metrics with human judgments.

For *Semantic Alignment* (Fig. 5, right), FG-CLIP can reward broad visual–text similarity without sufficiently penalizing semantic inconsistencies in the illustrated examples. CoT-based Deduction assigns lower scores when the generated content conflicts with the prompt.

For *Aesthetic Quality* (Fig. 5, left), ImageReward can assign high scores to visually appealing images despite structural inconsistencies, such as the extra figure in the first example and the character deformation in the second. Anchor-based Match Grading assigns lower scores to these outputs, more closely matching the human judgments shown in the figure.




Figure 5: **Qualitative validation of evaluation metrics.** The 1×4 grid illustrates cases in which baseline scores differ from human judgments. **Left (Aesthetic Quality):** ImageReward Xu et al. (2023) does not reproduce the displayed human preference ordering for images with structural inconsistencies, whereas Anchor-based Match Grading does. **Right (Semantic Alignment):** FG-CLIP Xie et al. (2025) assigns high scores despite semantic inconsistencies, whereas CoT-based Deduction assigns lower scores to these errors.

### 4.2 Benchmarking Existing Models

We benchmark representative open-source and commercial T2I systems using the official default configurations for open-source models. Qwen3.5-397B-A17B-FP8 serves as the QA and aesthetic evaluator. Due to access and generation constraints, the sample set obtained from commercial APIs is incomplete; therefore, the reported averages are computed over the available generations.

Semantic Alignment and Aesthetic Quality are evaluated on WeGenBench-General, while Visual Text Rendering is evaluated on WeGenBench-Text. Additional model, evaluator-prompt, configuration, and implementation details are provided in the supplementary material.

#### 4.2.1 Quantitative Analysis

Table 2 summarizes model performance using the semantic-alignment, aesthetic-quality, and visual-text-rendering metrics defined in Sec. 3. Detailed metric-specific analyses are provided in the supplementary material.

Table 2 shows that no single system leads uniformly across all dimensions. GPT-Image-2 obtains the highest CoT and AM scores and ties Nano-banana-2 for the highest displayed QA score, while Seedream-4.5 achieves the strongest overall visual-text-rendering results. Among open-source configurations, HunyuanImage variants are strong in Semantic Alignment, whereas GLM-Image is particularly competitive in visual text rendering. Tag-level analyses further reveal patterns shared by many evaluated models and capability-specific differences between systems.

**Shared bottlenecks.** Strong scene-level alignment does not imply accurate visual text. Several models obtain Quality/Readability scores above 0.90 while remaining substantially weaker on exact matching, and even the best systems remain below 0.55 character-level F1 on English user-interface prompts. This separation shows that producing clear text-like shapes and producing the requested characters are distinct capabilities.

Prompt complexity remains another broad bottleneck. Short prompts score approximately 1.0–1.7 points above the corresponding model averages, whereas multi-subject interactions, complex actions, precise quantities, counting, and negation consistently reduce performance. Figure 6 illustrates attribute-assignment failures under multiple subjects and a case in which a learned visual prior competes with an explicit color instruction.

The bilingual subsets also expose asymmetric behavior. Chinese prompts generally receive lower semantic-alignment scores than English prompts, while many models perform slightly better on Chinese visual text in this benchmark, possibly because the Chinese target strings are shorter and more standardized. Additional results by capability tag, together with discussion of the underlying VLM and OCR evaluators, are provided in the supplementary material.

Table 2: Benchmarking results for text-to-image generation models on WeGenBench across Aesthetic Quality, Semantic Alignment, and Visual Text Rendering. AM: Anchor-based Match Grading; QA: Checklist-based QA; CoT: CoT-based Deduction.

| | Semantic Alignment | | Aesthetic Quality | Visual Text Rendering | | | | |
|---|---|---|---|---|---|---|---|---|
| **Models** | QA ↑ | CoT ↑ | AM ↑ | $Sim_{edit}$ ↑ | $Acc_{sen}$ ↑ | $Score_{qua}$ ↑ | GNED ↑ | $Char_{F1}$ ↑ |
| *Open-Source Models* | | | | | | | | |
| GLM-Image Z.ai (2026) | 0.86 | 6.23 | 2.58 | 0.73 | 0.56 | 0.92 | 0.74 | 0.78 |
| FLUX.2-dev Black Forest Labs (2025) | 0.93 | 7.91 | 2.69 | 0.51 | 0.38 | 0.76 | 0.55 | 0.58 |
| Qwen-Image Wu et al. (2025) | 0.91 | 7.44 | 2.56 | 0.45 | 0.33 | 0.61 | 0.43 | 0.46 |
| Qwen-Image-2512 Wu et al. (2025) | 0.90 | 7.43 | 2.70 | 0.62 | 0.43 | 0.72 | 0.57 | 0.59 |
| Z-Image Z-Image Team (2025) | 0.92 | 7.65 | 2.44 | 0.50 | 0.53 | 0.78 | 0.52 | 0.55 |
| Z-Image-Turbo Z-Image Team (2025) | 0.90 | 7.26 | 2.67 | 0.67 | 0.50 | 0.87 | 0.69 | 0.73 |
| HunyuanImage-3.0-Inst. (image) Cao et al. (2025) | 0.95 | 8.26 | 2.65 | 0.63 | 0.56 | 0.88 | 0.66 | 0.69 |
| HunyuanImage-3.0-Inst. (think_re.) Cao et al. (2025) | 0.95 | 8.32 | 2.74 | 0.56 | 0.56 | 0.90 | 0.57 | 0.63 |
| HunyuanImage-3.0-Inst.-D (image) Cao et al. (2025) | 0.93 | 8.00 | 2.62 | 0.66 | 0.50 | 0.85 | 0.67 | 0.70 |
| HunyuanImage-3.0-Inst.-D (think_re.) Cao et al. (2025) | 0.94 | 8.08 | 2.70 | 0.57 | 0.54 | 0.89 | 0.57 | 0.63 |
| Ernie-Image-Turbo (pe) ERNIE Team (2026) | 0.92 | 7.72 | 2.59 | 0.60 | 0.56 | 0.90 | 0.61 | 0.65 |
| Ernie-Image-Turbo (no-PE) ERNIE Team (2026) | 0.92 | 7.87 | 2.65 | 0.70 | 0.56 | 0.89 | 0.71 | 0.75 |
| SenseNova-U1-8B-MoT Diao et al. (2026) | 0.92 | 7.72 | 2.38 | 0.68 | 0.52 | 0.90 | 0.71 | 0.74 |
| SenseNova-U1-8B-MoT-think Diao et al. (2026) | 0.93 | 7.81 | 2.41 | 0.68 | 0.51 | 0.88 | 0.70 | 0.72 |
| HiDream-O1 Cai et al. (2026) | 0.87 | 6.11 | 2.30 | 0.49 | 0.54 | 0.78 | 0.56 | 0.57 |
| *Closed-Source Models* | | | | | | | | |
| GPT-Image-2 OpenAI (2026) | **0.97** | **8.95** | **2.84** | 0.50 | 0.57 | 0.88 | 0.52 | 0.56 |
| Seedream-4.5 ByteDance Seed (2026) | 0.96 | 8.74 | 2.81 | **0.74** | **0.59** | **0.93** | **0.76** | **0.79** |
| Nano-banana-2 Google (2026) | 0.97 | 8.71 | 2.77 | 0.38 | 0.55 | 0.86 | 0.45 | 0.49 |

*Note:* Automatic Prompt Enhancement (APE) may introduce unprompted text elements and thereby reduce strict visual-text-rendering scores. This effect may influence commercial systems and reasoning-enhanced open-source configurations.



GPT-Image-2 Seedream-4.5

*Five Mexican mariachi musicians in black and silver charro suits playing instruments on a sunlit plaza, one with a gold trumpet, one with a white guitar, violin, vihuela, and guitarrón, orange marigold petals scattered on the cobblestones*

GPT-Image-2 Seedream-4.5

*Strings deforming around the ball — a white tennis ball strikes a racket's sweet spot, yellow felt fibers visible, motion lines from the swing, dark court background*

Figure 6: **Illustrative bottlenecks observed in leading T2I models. Left:** A prompt requesting five Mexican musicians, each holding a distinct instrument, illustrates a multi-subject attribute-binding failure: GPT-Image-2 and Seedream-4.5 misassign or omit some instruments. **Right:** For a prompt requiring a *white* tennis ball, GPT-Image-2 produces a yellow-green ball, a behavior consistent with a learned visual prior competing with the explicit instruction.

The supplementary material reports the complete L0–L5 aesthetic distributions, character-level Precision/Recall analysis, and tag-level capability results. The aesthetic distributions show that similar means can conceal different rates of high-quality outputs and severe failures: GPT-Image-2 and Seedream-4.5 place a larger share of samples in the upper levels, while Qwen-Image-2512 exhibits a comparatively high ceiling among open-source systems. Character-level decomposition further separates omissions from extraneous or incorrect characters; Seedream-4.5 and GLM-Image maintain relatively balanced Precision and Recall, whereas Nano-banana-2 has much lower Precision than Recall. These differences demonstrate why aggregate scores alone are insufficient for model diagnosis.

Tag-level results reveal similarly distinct profiles. FLUX.2-dev and GPT-Image-2 are comparatively balanced on spatial relations and attribute binding, while most systems weaken on complex actions.

HunyuanImage variants are strong among open-source models on complex-action alignment. For visual text, Seedream-4.5 and GLM-Image perform strongly on handwriting, complex layouts, and bilingual rendering, whereas models with high general-semantic scores can still struggle on these subsets. This separation connects each observed weakness to a narrower capability group instead of treating it as a generic model-quality failure.

#### 4.2.2 Qualitative Findings

Matched-prompt comparisons further show how tag-level scores correspond to localized failures. The supplementary material provides examples covering attribute binding, cultural elements, complex layout and handwriting, bilingual rendering, negation, and style fusion.

**Attribute binding.** In one representative case, GPT-Image-2 and Hunyuan preserve the requested materials of a concrete sneaker and a neon-orange yarn sneaker, whereas GLM-Image and Z-Image-Turbo transfer color or material across entities. The *Attribute Binding* tag localizes this failure and suggests targeted supervision on multi-entity prompts with explicitly bound attributes.

**Typography and bilingual rendering.** The qualitative comparisons reinforce the tag-level results. Seedream-4.5 and GLM-Image more accurately preserve multi-line handwritten text and mixed Chinese–English signage, while weaker outputs contain omissions, pseudo-characters, or unprompted text. These cases separate layout-sensitive text accuracy from general scene understanding and point to candidate improvements in spatial text planning and bilingual character supervision.

**Style and constraint composition.** Additional examples show that satisfying one requirement does not guarantee successful composition of several requirements. In style-fusion prompts, some models preserve character identity but lose the requested medium, while others apply the style so strongly that identity is weakened. Negation prompts reveal a related limitation: mentioning a forbidden concept can still influence the output. Together, these cases illustrate how capability tags turn qualitative failures into concrete groups for targeted inspection; the complete image comparisons are provided in the supplementary material.

**Diagnostic implications.** These examples complement the aggregate benchmark scores in two ways. First, they distinguish errors that would otherwise be grouped together: an incorrect output may arise from attribute transfer, omitted characters, layout violations, failure to combine writing systems, or loss of a requested style. Second, grouping such cases by capability tag makes it possible to inspect whether a failure recurs across scenarios and models. A consistently weak tag does not by itself establish the appropriate intervention, but it narrows the evidence that developers need to examine. For example, attribute-binding failures motivate testing additional multi-entity supervision, while layout and bilingual errors motivate controlled studies of spatial text planning and mixed-script training data. The supplementary case studies provide the corresponding images and more detailed model-by-model comparisons.

Additional generated-image galleries are provided in the supplementary material.

## 5 Conclusion

We introduced WeGenBench, a 4,000-prompt bilingual benchmark for diagnosing T2I systems across semantic alignment, aesthetic quality, and visual text rendering. Its scenario-and-tag taxonomy localizes failures in attribute binding, negation, complex actions, layout, and bilingual rendering, providing evidence for candidate data-curation and post-training directions. These diagnostic differences do not establish which intervention will succeed, and absolute scores remain subject to VLM and OCR recognition and calibration limits. Future work will expand language and cultural coverage and test tag-motivated interventions through controlled ablations.

### AI USE STATEMENT

Generative AI systems from the GPT and Gemini model families were used during benchmark construction to generate and rewrite a subset of candidate test cases and to assist with preliminary annotation. Human annotators screened every retained test case and finalized its labels. These tools were also used to assist with literature retrieval and discovery, drafting portions of the implementation code and manuscript, polishing the manuscript, and producing or editing figures. The authors manually reviewed and, where necessary, revised every AI-assisted output before inclusion.

The exact model versions and the proportion of AI-assisted test cases were not systematically recorded during the early stages of dataset construction. Generative AI neither replaced final human selection and verification nor independently determined the reported conclusions. The authors take full responsibility for the benchmark, experiments, claims, and final content of this work.

## REFERENCES

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

ByteDance Seed. "seedream-4.5", 2026. URL https://seed.bytedance.com/en/seedream4_5. Accessed: 2026-06-16.

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

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025. URL https://arxiv.org/abs/2509.23951.

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

ERNIE Team. Ernie-image technical report, 2026. URL https://arxiv.org/abs/2605.25347.

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

Google. "nano banana 2", 2026. URL https://deepmind.google/models/gemini-image/flash/. Accessed: 2026-06-16.

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.

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

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15086–15095, 2025.

OpenAI. "gpt-image-2", 2026. URL https://openai.com/index/introducing-chatgpt-images-2-0/. Accessed: 2026-06-16.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. *arXiv preprint arXiv:2311.03054*, 2023.

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

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.

Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Ji Ao, Dawei Leng, and Yuhui Yin. Fg-clip 2: A bilingual fine-grained vision-language alignment model. *arXiv preprint arXiv:2510.10921*, 2025.

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.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

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

Z.ai. "glm-image", 2026. URL https://github.com/zai-org/GLM-Image. Accessed: 2026-06-16.

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