Qwen-Image-Bench: From Generation to Creation in Text-to-Image Evaluation

Niantong Li[1], Guangzheng Hu[1], Weixu Qiao[1], Ying Ba[1], Qichen Hong[1], Shijun Shen[1], Jinlin Wang[1], Fan Zhou[1], Jianye Kang[1], Xin Shang[1], Ziyi He[1], Wei Wang[1], Dalin Li[1], Jiahao Li[1], Jie Zhang[1], Kaiyuan Gao[1], Kun Yan[1], Lihan Jiang[1], Ningyuan Tang[1], Shengming Yin[1], Tianhe Wu[1], Xiao Xu[1], Xiaoyue Chen[1], Yuxiang Chen[1], Yan Shu[1], Yanran Zhang[1], Yilei Chen[1], Yixian Xu[1], Zekai Zhang[1], Zhendong Wang[1], Zihao Liu[1], Zikai Zhou[1], Hongzhu Shi[1], Yi Wang[1], Bing Zhao[1], Hu Wei[1], Lin Qu[1], Chenfei Wu[1,*]
[1]Alibaba Group
[*]Corresponding author
{liniantong.lnt, fulai.hr}@alibaba-inc.com

Abstract

Existing Text-to-Image (T2I) benchmarks remain anchored in foundational criteria such as image quality and text-image alignment, failing to capture the nuanced capabilities that professional creators demand in real-world workflows. Meanwhile, evaluation pipelines that rely on an untrained multimodal model as a judge diverge from human standards. We introduce **Qwen-Image-Bench**, a creator-centric benchmark that extends evaluation with two application-driven dimensions (*Real-world Fidelity* and *Creative Generation*), alongside the conventional pillars of *Quality*, *Aesthetics*, and *Alignment*. The benchmark organizes these into a three-level taxonomy (5 pillars, 23 sub-capabilities, 56 verifiable rubrics) and provides 1,000 stratified bilingual prompts. We train a unified Judge Model (**Q-Judger**) on Qwen3.6-27B, supervised by 80 professional annotators under blind labeling and triple-review protocols, that scores every image across all 56 facets with fine-grained, rubric-grounded diagnostics. Empirically, Qwen-Image-Bench reliably distinguishes leading T2I models, achieving the greatest separation on the application-driven dimensions where existing benchmarks provide limited insight.

1 Introduction

Text-to-image (T2I) generation has advanced from producing visually plausible images to supporting photorealistic reconstruction and creative expression (Wang et al., 2025d; Xu et al., 2023; Zhang and Tang, 2024; Betker et al., 2023; Wu et al., 2025a; Z-Image Team et al., 2025; Google, 2025b; Wang et al., 2025c; Stability AI, 2024; Seedream et al., 2025). Driven by web-scale training (Li et al., 2024b), stronger text encoders (Li et al., 2024c; Wang et al., 2025a), and classifier-free guidance (Bradley and Nakkiran, 2024), diffusion-based and unified generative architectures have substantially improved basic prompt following (Stability AI, 2024; Black Forest Labs, 2025a; Google, 2025a; Betker et al., 2023; Wu et al., 2025b; Chen et al., 2025b; Z-Image Team et al., 2025; Jiang et al., 2025). Frontier models like Nano Banana Pro (Google, 2025b) and GPT Image 1.5 (OpenAI, 2025b) signal that the bottleneck has shifted from prompt alignment to *Real-world Fidelity* (faithfully reconstructing physical and cultural reality) and *Creative Generation* (supporting originality and stylistic coherence).

Despite this progress, T2I evaluation has lagged behind. Most benchmarks focus on coarse text-image alignment via CLIP similarity (Bakr et al., 2023; Wu et al., 2023; Pavlov et al., 2023). Recent vision-language judges (TIFA (Hu et al., 2023), GenAI-Bench (Li et al., 2024a)) and preference-based evaluators (ImageReward (Xu et al., 2023), HPS (Wu et al., 2023)) remain tied to semantic correctness. We identify three key limitations: **(1) Saturation on basic axes reduces discriminative power**, as benchmarks compress strong models into indistinguishable score bands (Kamath et al., 2025; Chen et al., 2025a). **(2) *Real-world Fidelity* remains insufficiently characterized**—existing suites rarely separate visual plausibility from genuine physical, cultural, or factual grounding (Meng et al., 2024; Niu et al., 2026; Zhang et al., 2025; Nayak et al., 2025). **(3) *Creative Generation* lacks expert-grounded evaluation**: most benchmarks lack fine-grained taxonomies for aesthetics and originality, and pipelines that rely on an untrained MLLM judge (e.g., UniGenBench++ (Wang et al., 2025b) using Gemini-2.5 as supervision) inherit

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 22760–22787
October 24-29, 2026 ©2026 Association for Computational Linguistics

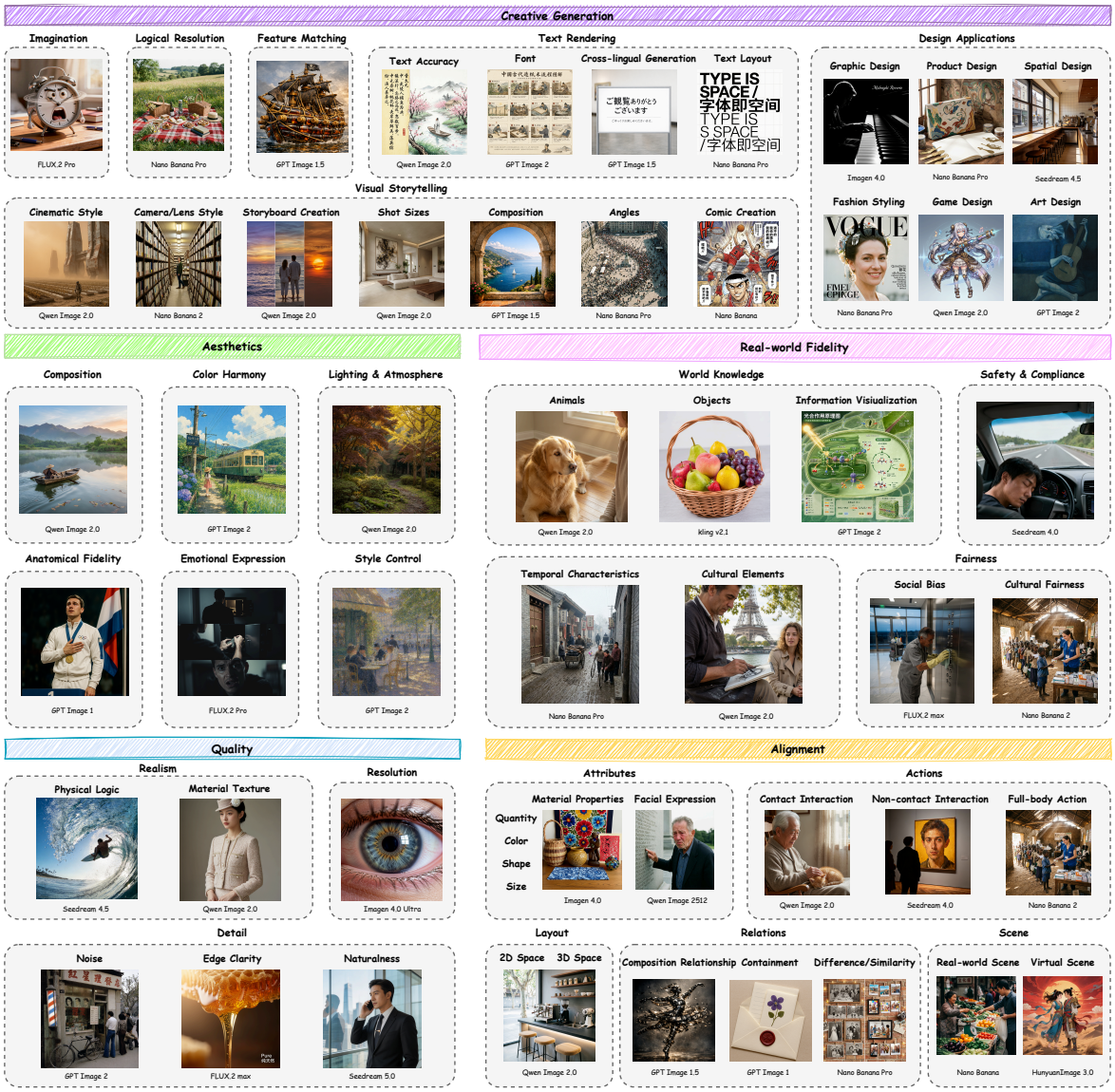


Figure 1: Qwen-Image-Bench Evaluation Dimensions.

systematic biases rather than reflecting professional standards.

To address these gaps, we introduce **Qwen-Image-Bench**, a *creator-centric* benchmark co-developed with professional artists that jointly evaluates *Real-world Fidelity* and *Creative Generation* alongside *Quality*, *Aesthetics*, and *Text-Image Alignment*. As shown in Fig. 1, it comprises 1,000 expert-crafted prompts organized under a three-level taxonomy with balanced coverage.

The taxonomy is designed top-down along professional artistic workflows (*ideation*, *styling*, *refinement*), yielding **5 L1 pillars**, **23 L2 sub-capabilities**, and **56 L3 evaluation facets**. This structure ensures systematic coverage, enables each prompt to exercise multiple facets simultaneously, and localizes model strengths/weaknesses to specific sub-skills. Among the 56 L3 facets, 28 fall under *Real-world Fidelity* and *Creative Generation*, covering world knowledge, design applications, visual storytelling, and text rendering.

Regarding prompts, existing benchmarks are dominated by short prompts (16.43 tokens in HRS-Bench (Bakr et al., 2023), 12.65 in T2I-CompBench (Huang et al., 2025)), while dense-prompt suites like DPG-Bench (Hu et al., 2024) and UniGenBench++ (Wang et al., 2025b) cap around 80–160 tokens. Qwen-Image-Bench balances 500 long and 500 short prompts with bilingual (Chinese/English) versions to probe robustness to both prompt length and language.

For evaluation, we provide a **unified Judge Model** (**Q-Judger**) built on Qwen3.6-27B and trained on dense annotations from 80 professional

experts (photography, directing, fine arts) with at least three independent blind reviews per sample. Unlike MLLM judges that output a single score, Q-Judger produces fine-grained scores across all 56 L3 facets, aggregated bottom-up to L2, L1, and overall scores (Section 3.4). It achieves Spearman $\rho = 0.92$ with human expert rankings. We release both the Judge Model and prompt set for fully offline evaluation.

We evaluate 18 T2I models including GPT Image 2 (OpenAI, 2026), Nano Banana Pro (Google, 2025b), GPT Image 1.5 (OpenAI, 2025b), Seedream 5.0, Imagen 4.0 Ultra (Google, 2025a), Qwen Image 2.0 Pro (Zhao et al., 2026), HunyuanImage 3.0 (Team, 2025), and GLM Image (Z.ai, 2026) (Fig. 3). GPT Image 2 achieves the highest overall score (64.69). Crucially, *Creative Generation* and *Real-world Fidelity* exhibit the largest inter-model variance, exposing capability gaps that are less visible under conventional quality and alignment axes. Five L3 facets (Physical Logic, Anatomical Fidelity, Animals, Objects, Contact Interaction) emerge as systemic ceilings where even the best models score below 44.

2 Related Work

T2I Benchmarks. Early benchmarks such as GenEval (Ghosh et al., 2023) and T2I-CompBench (Huang et al., 2025) focus on evaluating basic capabilities like object-centric properties, typically leveraging detectors and similarity models for scalable evaluation. Recent benchmarks adopt VQA and stronger judge models to move toward verbs and richer scene semantics, like TIFA (Hu et al., 2023) and DSG-1k (Cho et al., 2024). Newer benchmarks further target instruction-following under prompt complexity (TIIF-Bench (Wei et al., 2025)), prompt reliability (Gecko (Wiles et al., 2025)), and world knowledge/reasoning (WISE (Niu et al., 2026), R2I-Bench (Chen et al., 2025c), T2I-ReasonBench (Sun et al., 2026)). Comprehensive frameworks such as UniGenBench++ (Wang et al., 2025b) and OneIG-Bench (Chang et al., 2025) also pursue multi-dimensional coverage, including text rendering and stylization. Despite their breadth, existing benchmarks give less explicit and fine-grained coverage to *Real-world Fidelity* and *Creative Generation*, two capabilities that are central to practical T2I usage.

T2I Evaluation Methods. Early automated evaluation relied on specialized discriminative models as judges, as in GenEval (Ghosh et al., 2023). The field has since largely converged on VQA/VLM-mediated scoring, where prompts are converted into questions and correctness is assessed by VQA consistency: TIFA (Hu et al., 2023) uses LLM-driven question generation, DSG (Cho et al., 2024) enforces structured coverage and consistency constraints, and ConceptMix (Wu et al., 2024) grades concept presence via VLM-based questioning. In parallel, preference-based reward models (e.g., ImageReward (Xu et al., 2023)) and large-scale human preference datasets (e.g., Pick-a-Pic (Kirstain et al., 2023)) aim to capture subjective quality that resists discrete correctness criteria.

However, assessing creativity and aesthetics remains challenging due to their subjective, multi-faceted nature. Existing approaches rely on aesthetic datasets with aggregate ratings (e.g., AVA, Pick-a-Pic), but these are limited in scale and artistic scope. More critically, most benchmarks conflate technical quality (sharpness, artifact absence) with artistic creativity (originality, compositional balance, narrative coherence) into a single score, producing ambiguous signals that poorly reflect expert judgment. Furthermore, none are grounded in actual creative workflows or professional standards.

As a result, even advanced evaluation pipelines remain sensitive to systematic bias and cannot adequately assess both knowledge-consistent realism and authentic creative expression. To address this, we introduce **Qwen-Image-Bench**, a creator-centric benchmark that jointly evaluates *Real-world Fidelity* and *Creative Generation* through an expert-co-designed taxonomy, supported by a unified Judge Model delivering rubric-grounded fine-grained scoring at the third-level facet granularity.

3 Qwen-Image-Bench

Qwen-Image-Bench is built on three tightly coupled components: a hierarchical capability taxonomy that redefines what is measured (Sec.3.1), an expert-in-the-loop prompt factory that turns the taxonomy into executable evaluation prompts (Sec.3.2), and a unified Judge Model supervised by professional annotators that scores every generated image at the finest third-level granularity and propagates those scores up the taxonomy to produce interpretable diagnostics at any level (Sec. 3.3–3.4). We describe each in turn, emphasizing the design

choices that distinguish our framework from prior T2I benchmarks.

3.1 A Hierarchical Capability Taxonomy

Motivation. Existing T2I benchmarks collapse generation ability into a handful of coarse axes, typically image quality, aesthetics, and text-image alignment. As frontier models converge on these fundamentals, flat taxonomies lose discriminative power and, more importantly, lose *explanatory* power: they can rank models but cannot localize where they differ. What is needed is a capability structure that mirrors how professional creators reason about image generation, from the overall purpose of an image, through the sub-capabilities it draws on, down to the concrete visual details a reviewer inspects. Our taxonomy is designed from the creator's perspective (creator-centric): rather than asking *what a researcher can measure*, we ask *what a professional creator actually cares about in their workflow*.

Three-level structure. Qwen-Image-Bench organizes evaluation as a top-down capability tree with three levels of granularity. At the root sit five **first-level pillars**: the three conventional dimensions (*Quality*, *Aesthetics*, and *Alignment*) plus two application-driven pillars, *Real-world Fidelity* and *Creative Generation*, that target the capabilities most valued in modern creator workflows. Each pillar is then expanded by professional artists into a set of **second-level sub-capabilities** reflecting the concrete decisions an artist makes during creation, such as World Knowledge and Fairness under Real-world Fidelity, or Visual Storytelling and Design Applications under Creative Generation. Every sub-capability is further decomposed into **third-level facets**, each defined by a precise rubric that specifies the observable criteria for scoring, making evaluation transparent and reproducible. In total, the taxonomy comprises 5 first-level pillars, 23 second-level sub-capabilities, and 56 third-level evaluation facets (Fig. 1; Tab. 7).

Why the hierarchy matters. This design yields three useful properties. 1) Every facet is *atomic and unambiguous*, so prompts, annotations, and scores share the same vocabulary. 2) Facets *propagate deterministically upward*: each third-level facet belongs to exactly one second-level sub-capability and one first-level pillar, allowing fine-grained scores at all three levels from a single set of images. 3) The hierarchy is also *diagnostic by construction*: a model's strength or weakness can always be traced to an identifiable artistic sub-skill rather than an opaque aggregate.

3.2 Expert-in-the-Loop Prompt Construction

The taxonomy provides what to evaluate; the prompt set determines whether evaluation is realized. Qwen-Image-Bench adopts an **expert-in-the-loop prompt factory** (Fig. 2) in which professional artists are co-authors rather than post-hoc reviewers, defining prompts across **17 application domains** anchored to real creative scenarios.

The factory is a four-stage pipeline: (1) facet-targeted sampling that selects $k \geq 3$ L3 facets per prompt under a balancing constraint; (2) bilingual drafting via LLMs with explicit facet-instantiation notes; (3) expert review, rewrite, and re-verification against rubrics; and (4) length-variant expansion into short/long pairs (thresholds: 70 Chinese characters, 235 English characters). The full formal specification is provided in Appendix A.7.

This pipeline yields **1,000 prompts** (500 short + 500 long), each in Chinese and English, with each prompt exercising 3–5 L1 pillars simultaneously.

3.3 Unified Judge Model

Existing automated evaluators either compress all capabilities into a single score (drowning out subtler dimensions) or train directly on MLLM labels without human supervision (e.g., UniGen-Bench++ (Wang et al., 2025b) using Gemini-2.5), inheriting systematic biases.

Qwen-Image-Bench provides a *single unified Judge Model* (**Q-Judger**) fine-tuned from Qwen3.6-27B, supervised entirely by expert human annotations. Given a prompt-image pair (p, I), Q-Judger scores every L3 facet independently:

$$\{s_c(p, I)\}_{c \in C} = J(p, I, \{r_c\}_{c \in C}), \qquad (1)$$

where r_c is the rubric for facet c. Each score takes one of four values: **0** (Fail), **1** (Pass), **2** (Excel), or **N/A** (inapplicable).

Training data. The model is trained on over **130,000** bilingual prompt-image pairs annotated by **80 professional experts** (art academy graduates in photography, directing, and fine arts). All annotations follow a blind, checklist-driven protocol with at least three independent reviews per sample.

3.4 Evaluation Pipeline and Scoring

At inference, Q-Judger scores each pair (p, I) across all 56 L3 facets. Inapplicable facets receive

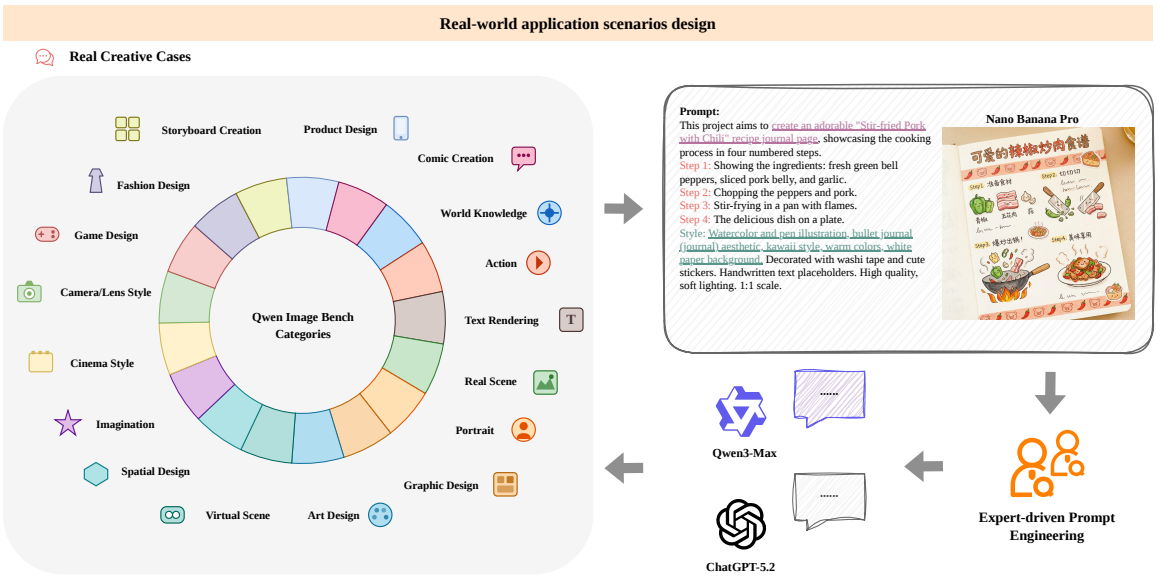


Figure 2: Pipeline for constructing real-world application prompts

N/A and are excluded. Non-N/A scores are normalized to $[0, 100]$:

$$\phi(s) = \begin{cases} 0 & \text{if } s = 0 \\ 60 & \text{if } s = 1 \\ 100 & \text{if } s = 2 \end{cases} \tag{2}$$

Pass maps to 60 (not 50), amplifying the Fail–Pass gap over Pass–Excel, reflecting that unacceptable-vs-acceptable matters more in practice.

Per-sample hierarchical aggregation. Let $\mathcal{C}(p)$ be the active L3 facets for prompt p. Scores aggregate bottom-up:

$$s_b(p, I) = \frac{1}{|\{c \in b \cap \mathcal{C}(p)\}|} \sum_{\substack{c \in b \\ c \in \mathcal{C}(p)}} \phi(s_c(p, I)) \tag{3}$$

$$s_d(p, I) = \frac{1}{|\mathcal{B}_d(p)|} \sum_{b \in \mathcal{B}_d(p)} s_b(p, I) \tag{4}$$

$$s_{\text{overall}}(p, I) = \frac{1}{|\mathcal{D}(p)|} \sum_{d \in \mathcal{D}(p)} s_d(p, I) \tag{5}$$

where $\mathcal{B}_d(p)$ and $\mathcal{D}(p)$ are the active L2 sub-capabilities and L1 pillars for prompt p.

Model-level scores. Benchmark scores average sample-level scores over all activating prompts:

$$S_c = \frac{1}{|\mathcal{T}_c|} \sum_{(p,I) \in \mathcal{T}_c} \phi(s_c(p, I)) \tag{6}$$

$$S_b = \frac{1}{|\mathcal{T}_b|} \sum_{(p,I) \in \mathcal{T}_b} s_b(p, I) \tag{7}$$

$$S_{\text{overall}} = \frac{1}{|\mathcal{P}|} \sum_{(p,I) \in \mathcal{P}} s_{\text{overall}}(p, I) \tag{8}$$

All scores reside in $[0, 100]$. This bottom-up design ensures every aggregate is traceable to specific L3 facets, enabling actionable diagnostics.

4 Experiment

4.1 Implementation Details of Q-Judger

Backbone and framework. The Judge Model is initialized from Qwen3.6-27B and fine-tuned within the MS-SWIFT (Zhao et al., 2024) framework, which we adopt for its mature multimodal fine-tuning support and reproducibility.

Inference. At evaluation time, we use deterministic decoding with `seed` = 42, `repetition_penalty` = 1.05, `top_k` = 1, `top_p` = 1, and `temperature` = 0. We enable the model's thinking mode (`enable_thinking` = True) to allow chain-of-thought reasoning before scoring. The Judge Model receives a structured prompt consisting of a system-level role definition, the generation prompt, the generated image, a scoring rubric, and a dimension-specific evaluation checklist encoding the full L2/L3 hierarchy. The complete prompt templates and all five pillar-level checklists are provided in Appendix B.

Training data. The Judge Model is trained on over **130,000** bilingual expert-annotated prompt-image pairs following the rubric-grounded protocol described in Sec.3.3 (80 art-academy annota-

Table 1: Spearman rank correlation (ρ) between Q-Judger rankings and human expert rankings across five L1 pillars and overall. All correlations are statistically significant ($p < 10^{-4}$, $N = 18$ models).

Dimension	Spearman ρ
Quality	0.89
Aesthetics	0.89
Alignment	0.89
Real-world Fidelity	0.92
Creative Generation	0.92
Overall	**0.92**

tors, blind labeling, at least three independent reviews per sample). A systematic exact, lexical, and semantic audit finds no confirmed overlap between the training prompts and the 1,000 evaluation prompts (Appendix A.1).

Alignment with human experts. On a held-out evaluation set, senior expert reviewers independently rate each image on a 1–10 holistic scale per L1 pillar; models are then ranked by their per-pillar and overall average human scores (detailed results shown in Table 6). The Judge Model achieves strong ranking consistency with these human rankings (as shown in Table 1). Inter-annotator reliability is also substantial (ordinal Krippendorff's $\alpha = 0.705$; ICC$(2, k) = 0.876$); the full protocol is reported in Appendix A.1.

4.2 Comparison with Zero-shot MLLM Judges

To isolate the value of expert-supervised fine-tuning, we compare Q-Judger with four zero-shot

Judge	System-level ρ
Q-Judger (fine-tuned)	**0.924**
Qwen3.6-27B (zero-shot)	0.909
Gemini-3.1-Pro (zero-shot)	0.895
Claude-Opus-4.7 (zero-shot)	0.831
GPT-5.5 (zero-shot)	0.761

Table 2: Controlled comparison with zero-shot MLLM judges. All judges score the same 18,000 images with the same checklist. System-level ρ is computed over the rankings of 18 T2I models.

MLLM judges on the same 18,000 images using the same checklist; only the judge is changed. Table 2 reports system-level Spearman correlation with human rankings. Q-Judger obtains the highest correlation ($\rho = 0.924$), outperforming both its unfine-tuned Qwen3.6-27B backbone (0.909) and three frontier proprietary judges. A direct probe finds no significant Qwen-family favoritism (Qwen-versus-non-Qwen gap $+0.024$, Mann–Whitney $p = 0.54$). These results support Q-Judger's system-level ranking reliability while retaining the practical advantages of fully offline and reproducible evaluation. Full details are provided in Appendix A.1.

Robustness to scoring design. Re-aggregating the same predictions with a linear $0/50/100$ mapping instead of $0/60/100$ preserves the ranking almost perfectly (Spearman $\rho = 0.996$, Kendall $\tau = 0.974$), showing that the observed separation is not an artifact of the nonlinear mapping. Coverage is also broad: all 56 L3 facets are activated, with each prompt receiving scores on 29.5 facets on average (15–46); full results are in Appendix A.1.

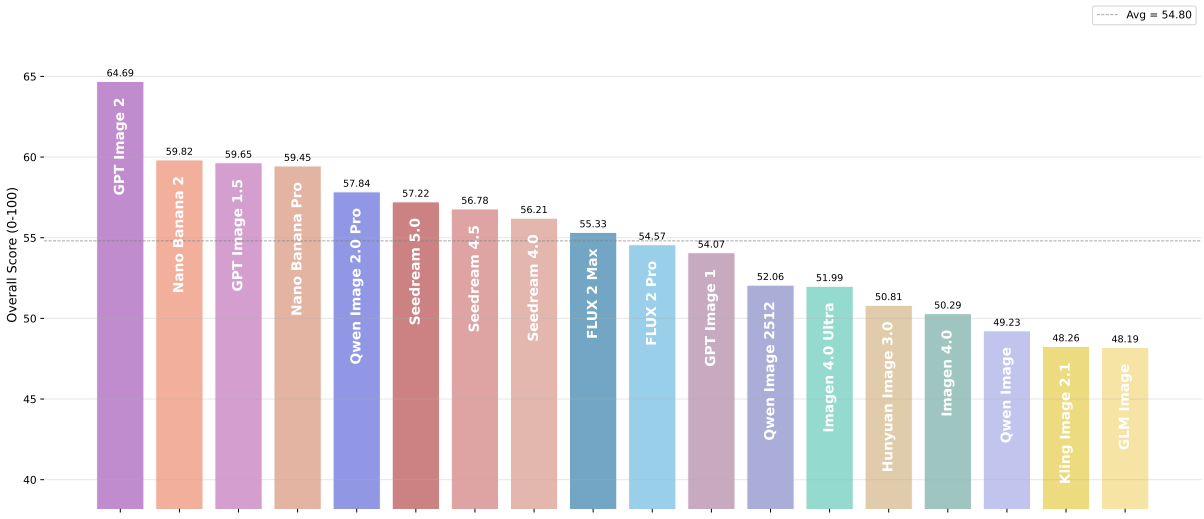


Figure 3: Overall score ranking of 18 T2I models on Qwen-Image-Bench. The dashed line marks the cross-model mean.

Table 3: Overall performance of 18 T2I models on Qwen-Image-Bench. Scores are on a $[0, 100]$ scale, aggregated bottom-up from 56 L3 facets through the three-level taxonomy (Sec. 3.4). Models are sorted by overall score. The best score in each column is **bolded**.

Model	Evaluation Dimension					Overall
	Quality	*Aesthetics*	*Alignment*	*Real-world Fidelity*	*Creative Generation*	
GPT Image 2	**58.65**	**67.53**	**65.85**	**57.38**	**75.23**	**64.69**
Nano Banana 2.0	54.77	61.08	62.40	54.28	67.05	59.82
GPT Image 1.5	55.14	60.88	61.72	53.95	66.35	59.65
Nano Banana Pro	55.67	60.26	61.25	54.07	66.23	59.45
Qwen Image 2.0 Pro	54.39	58.67	59.28	51.83	64.94	57.84
Seedream 5.0	52.55	58.40	58.90	51.92	65.29	57.22
Seedream 4.5	54.41	58.72	57.31	51.69	60.64	56.78
Seedream 4.0	54.01	58.81	56.64	51.05	58.15	56.21
FLUX 2 Max	53.64	56.85	57.35	49.35	56.50	55.33
FLUX 2 Pro	52.30	56.94	57.01	47.29	56.18	54.57
GPT Image 1	52.34	55.09	56.28	48.14	55.78	54.07
Qwen Image 2512	51.76	54.74	52.72	47.00	50.19	52.06
Imagen 4.0 Ultra	50.90	54.25	54.02	45.59	51.14	51.99
HunyuanImage 3.0	50.35	53.57	52.00	44.31	49.12	50.81
Imagen 4.0	50.16	52.68	51.64	44.84	47.94	50.29
Qwen Image	48.44	52.25	50.72	43.16	47.30	49.23
Kling Image 2.1	49.11	50.15	49.18	44.74	44.67	48.26
GLM Image	49.26	50.64	47.90	44.69	45.23	48.19

4.3 Benchmarked T2I Models

We evaluate 18 representative T2I models on Qwen-Image-Bench, spanning a broad spectrum of frontier T2I models: GPT Image 2 (OpenAI, 2026), GPT Image 1.5 (OpenAI, 2025b), GPT Image 1 (OpenAI, 2025a), Nano Banana 2.0 (Google, 2026), Nano Banana Pro (Google, 2025b), Seedream 5.0, Seedream 4.5 (ByteDance Seed, 2025), Seedream 4.0 (Seedream et al., 2025), Imagen 4.0 Ultra (Google, 2025a), Imagen 4.0 (Google, 2025a), FLUX 2 Max (Black Forest Labs, 2025b), FLUX 2 Pro (Black Forest Labs, 2025a), Qwen Image 2.0 Pro (Zhao et al., 2026), Qwen Image 2512 (Wu et al., 2025a), Qwen Image (Wu et al., 2025a), HunyuanImage 3.0 (Team, 2025), GLM Image (Z.ai, 2026), and Kling Image 2.1 (Kling AI, 2025).

For every model, we generate images for all 1,000 Qwen-Image-Bench prompts, then evaluate all prompt-image pairs with the Judge Model, which produces a structured score vector of 56 third-level facet scores per pair. These fine-grained scores are aggregated bottom-up following the protocol in Sec.3.4 to yield first-level pillar scores and the overall benchmark score.

4.4 Overall Performance on Qwen-Image-Bench

Tab. 3 reports the performance of all 18 T2I models along the five first-level pillars and the aggregated overall score. We present the results following a top-down analysis: overall ranking, fine-grained capability profiles, variance-driven discriminability analysis, sub-capability rankings, per-pillar rankings, industry-wide capability landscape, and cross-tier gap analysis.

Overall Ranking. Fig. 3 presents the overall ranking. GPT Image 2 leads by nearly 5 points over the second-ranked Nano Banana 2.0, with GPT Image 1.5 and Nano Banana Pro following closely to form a tightly clustered second tier. Qwen Image 2.0 Pro ranks fifth overall, heading the third tier. GLM Image sits at the bottom, yielding a 16.5-point spread from the leader and demonstrating the benchmark's effective discriminative power across the full model spectrum. Based on the two-decimal scores reported in Table 3, the tiers are T1 ($s \geq 64$), T2 ($59 \leq s < 64$), T3 ($56 \leq s < 59$), T4 ($54 \leq s < 56$), and T5 ($s < 54$). Notably, GPT Image 2 achieves the highest score on all five L1 pillars simultaneously, forming a dominant profile with no discernible weakness, a rarity among the 18 evaluated models.

4.5 Discriminative Power of New Dimensions

A key design goal of Qwen-Image-Bench is to introduce evaluation dimensions that expose capability gaps invisible to previous benchmarks. Figure 4 quantifies this by measuring inter-model score variance (the higher the variance, the more a dimension differentiates models) across all three taxonomy levels.

L3 variance pinpoints the sharpest frontiers. At the finest granularity, Text Accuracy (under Cre-

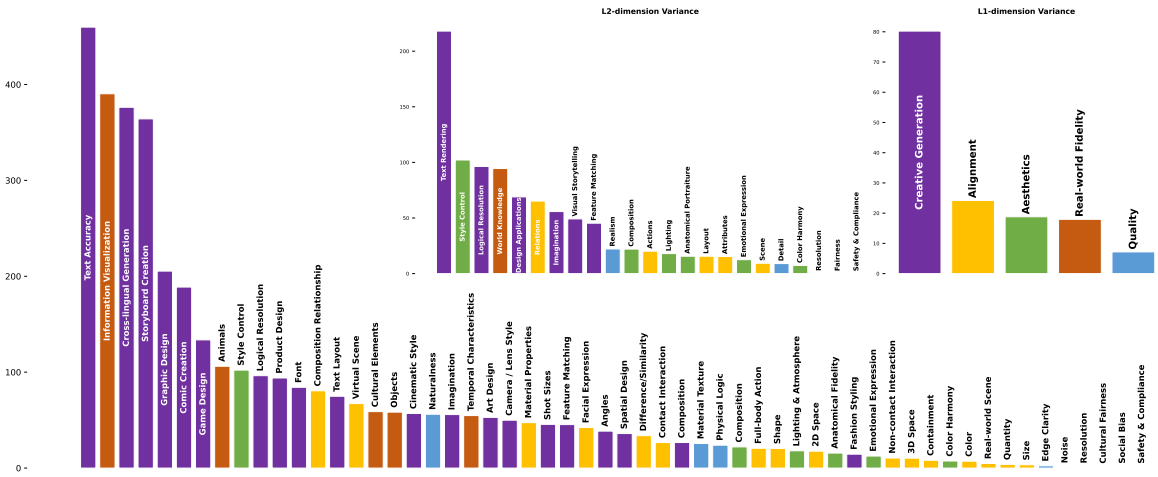


Figure 4: Inter-model score variance across the three levels of the taxonomy. Main plot: L3-level variance (56 facets); insets: L2-level (23 sub-capabilities) and L1-level (5 pillars).

ative Generation) is the single most discriminative facet. Information Visualization (under Real-world Fidelity) and Cross-lingual Generation (under Creative Generation) rank second and third. Of the 15 highest-variance L3 facets, 12 belong to Creative Generation or Real-world Fidelity (e.g., Storyboard Creation, Graphic Design, Cross-lingual Generation, Game Design), dimensions that jointly test creative imagination, logical reasoning, and execution precision.

L2 variance confirms application-driven dimensions dominate. Rolling up to the second level, the highest-variance L2 sub-capability is Text Rendering (under Creative Generation), followed by Style Control (under Aesthetics), Logical Resolution (under Creative Generation), and World Knowledge (under Real-world Fidelity). Among the top six L2 dimensions by variance, four belong to the two application-driven pillars introduced by our benchmark.

L1 variance reveals where differentiation lies. At the pillar level, Creative Generation variance is over 11× that of Quality and over 4× that of Aesthetics. The low variance on Quality indicates that basic image quality has become a "table-stakes" capability, while Creative Generation, the pillar most unique to our creator-centric design, is precisely where models diverge most sharply.

Taken together, the variance analysis provides quantitative evidence that the two application-driven pillars introduced by Qwen-Image-Bench, namely Creative Generation and Real-world Fidelity, expose capability gaps that are less visible under conventional quality and alignment axes. Models achieve near-consensus on basic quality and alignment (low variance), but diverge sharply on knowledge-grounded and creative tasks. These dimensions increase the benchmark's diagnostic resolution and highlight concrete directions for future model improvement.

4.6 Additional Analyses

Beyond the overall ranking and variance analyses presented above, we conduct extensive fine-grained investigations at multiple granularities, with full results reported in Appendix A:

- **Fine-grained capability profiles** (Appendix A.2, Fig. 5): The L3 radar chart reveals that GPT Image 2 forms a uniquely full heart-shaped polygon, while all models share a universal indentation at five "ceiling" facets (Anatomical Fidelity, Physical Logic, Objects, Animals, Contact Interaction), exposing an industry-wide inability to handle fine-grained physical and biological structure regardless of overall ranking. Mid-tier models exhibit asymmetric deformation—strong on visual style but collapsing on execution precision—revealing a fundamental "imagination vs. execution" gap within Creative Generation.
- **Sub-capability and facet-level rankings** (Appendix A.3, Fig. 6): Text Rendering exhibits a sharp technology cliff between rank 7 and rank 8 (7-point drop), indicating a binary divide between models with and without dedicated text capabilities. Game Design shows the most extreme threshold effect (91.8 for the leader, below 66 from rank 7 onward), charac-

teristic of tasks requiring both visual imagination and domain-specific knowledge simultaneously. Overall ranking alone masks significant specialization differences: FLUX 2 Pro is competitive on Text Rendering but drops sharply on World Knowledge.

- **Per-pillar model rankings** (Appendix A.4, Fig. 7): Model ordering shifts dramatically across pillars, demonstrating that no single ranking captures the full picture. Creative Generation produces the largest spread (30.6 points) and most dramatic re-ordering. A notable version-evolution trade-off emerges: Seedream 5.0 surpasses 4.5 on Creative Generation by 4+ points but *regresses* on Quality and Aesthetics, a trade-off only visible through multi-pillar evaluation.
- **Industry-wide capability landscape** (Appendix A.5, Fig. 8): The field exhibits a clear **perception-to-cognition frontier**—facets requiring only learned visual statistics (Camera/Lens Style, Art Design, Cinematic Style, all above 66) are effectively solved, while facets requiring implicit reasoning about physics, biology, or 3D structure collapse uniformly below 35. The 32-point gap between highest and lowest cross-model means within Creative Generation alone confirms this as the defining structural feature of current T2I capabilities.
- **Cross-tier gap analysis** (Appendix A.6, Tab. 5): Creative Generation dominates tier separation at every boundary (+8.68 T1–T2, +4.29 T2–T3), roughly $1.8\times$ the gap on conventional pillars, while Quality gap is only 1.35—confirming basic quality as table-stakes. The upgrade path analysis reveals a "language-understanding-first, visual-execution-next" pattern: Qwen Image 2.0 Pro already surpasses T2 on language-intensive facets (Text Accuracy +11.4, Storyboard +10.2) but lags on visual-execution facets (Game Design, Objects), suggesting a concrete two-phase improvement strategy.

The L3-level and L2-level score heatmaps (Appendix, Figs. 10–9) provide complementary visualizations at the individual-model level.

5 Conclusion

We presented **Qwen-Image-Bench**, a creator-centric benchmark with a three-level taxonomy (5 pillars, 23 sub-capabilities, 56 facets) co-designed with professional artists, and a unified Judge Model trained on 130K+ expert-labeled samples achieving system-level Spearman $\rho = 0.924$ with human rankings. The benchmark cleanly separates 18 T2I models into five tiers with a 16.5-point spread, exposing a *perception-to-cognition frontier*: five facets (Physical Logic, Anatomical Fidelity, Animals, Objects, Contact Interaction) collapse uniformly across all models (best < 44), revealing that current models have mastered surface reproduction but lack implicit world knowledge. We plan to evolve Qwen-Image-Bench into a living benchmark with dynamic prompt refresh, extension to video/editing, and a continuously updated public leaderboard.

Limitations

Our benchmark currently evaluates only single-image generation and does not cover multi-turn editing, video synthesis, or interactive workflows. The Judge Model, while achieving strong system-level human correlation ($\rho = 0.924$), is trained on annotations from Chinese-nationality experts across global art academies, which may introduce cultural bias in aesthetic judgments. Additionally, the 1,000-prompt evaluation set, though carefully balanced, cannot exhaustively cover all real-world creative scenarios; we will continuously update the prompt set as model capabilities evolve.

References

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

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. 2023. Improving image generation with better captions.

Black Forest Labs. 2025a. "FLUX.2". Accessed: 2026-05-14.

Black Forest Labs. 2025b. "FLUX.2 [max]". Accessed: 2026-05-14.

Arwen Bradley and Preetum Nakkiran. 2024. Classifier-free guidance is a predictor-corrector. *Preprint*, arXiv:2408.09000.

ByteDance Seed. 2025. "seedream 4.5". Accessed: 2026-05-14.

Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. 2025. OneIG-Bench: Omni-dimensional nuanced evaluation for image generation. *Advances in Neural Information Processing Systems*, 38.

Jiahui Chen, Candace Ross, Reyhane Askari-Hemmat, Koustuv Sinha, Melissa Hall, Michal Drozdzal, and Adriana Romero-Soriano. 2025a. Multi-modal language models as text-to-image model evaluators. *Preprint*, arXiv:2505.00759.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. 2025b. BLIP3-o: A family of fully open unified multimodal models-architecture, training and dataset. *Preprint*, arXiv:2505.09568.

Kaijie Chen, Zihao Lin, Zhiyang Xu, Ying Shen, Yuguang Yao, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. 2025c. R2I-Bench: Benchmarking reasoning-driven text-to-image generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12606–12641.

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

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

Google. 2025a. "imagen 4.0-ultra". Accessed: 2026-05-14.

Google. 2025b. "nano banana pro". Accessed: 2026-05-14.

Google. 2026. "nano banana 2". Accessed: 2026-05-14.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. 2024. Ella: Equip diffusion models with llm for enhanced semantic alignment. *Preprint*, arXiv:2403.05135.

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

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2025. T2I-CompBench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3563–3579.

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. 2025. T2I-R1: Reinforcing image generation with collaborative semantic-level and token-level cot. *Advances in Neural Information Processing Systems*, 38:39856–39890.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. 2025. GenEval 2: Addressing benchmark drift in text-to-image evaluation. *Preprint*, arXiv:2512.16853.

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

Kling AI. 2025. "kling image 2.1". Accessed: 2026-05-14. Online image editing service for style transfer and artistic rendering.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. 2024a. GenAI-Bench: Evaluating and improving compositional text-to-visual generation. *Preprint*, arXiv:2406.13743.

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

Yanyu Li, Xian Liu, Anil Kag, Ju Hu, Yerlan Idelbayev, Dhritiman Sagar, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. 2024c. Textcraftor: Your text encoder can be image quality controller. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7985–7995.

Fanqing Meng, Wenqi Shao, Lixin Luo, Yahong Wang, Yiran Chen, Quanfeng Lu, Yue Yang, Tianshuo Yang, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024. PhyBench: A physical commonsense benchmark for evaluating text-to-image models. *Preprint*, arXiv:2406.11802.

Shravan Nayak, Mehar Bhatia, Xiaofeng Zhang, Verena Rieser, Lisa Anne Hendricks, Sjoerd Van Steenkiste, Yash Goyal, Karolina Stańczak, and Aishwarya Agrawal. 2025. Culturalframes: Assessing cultural expectation alignment in text-to-image models and evaluation metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20918–20953.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Fanqing Meng, Kunpeng Ning, Bin Zhu, and Li Yuan. 2026. WISE: A world knowledge-informed semantic evaluation for text-to-image generation. In *Proceedings of the 43rd International Conference on Machine Learning*.

OpenAI. 2025a. GPT Image 1 Model. Accessed: 2026-08-28.

OpenAI. 2025b. "gpt-image-1.5". Accessed: 2026-02-04.

OpenAI. 2026. GPT Image 2 Model. Accessed: 2026-08-28.

Igor Pavlov, Artyom Ivanov, and Stanislav Stafievskiy. 2023. Text-to-image benchmark: A benchmark for generative models. Version 0.1.0; Accessed: 2026-02-04.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, Xiaowen Jian, Huafeng Kuang, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, Wei Liu, and 31 others. 2025. Seedream 4.0: Toward next-generation multimodal image generation. *Preprint*, arXiv:2509.20427.

Stability AI. 2024. Stable Diffusion 3: Research paper.

Kaiyue Sun, Rongyao Fang, Chengqi Duan, Xian Liu, Aoxue Li, and Xihui Liu. 2026. T2I-ReasonBench: Benchmarking reasoning-informed text-to-image generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 8919–8944, San Diego, California, United States. Association for Computational Linguistics.

Tencent Hunyuan Foundation Model Team. 2025. HunyuanImage 3.0 technical report. *Preprint*, arXiv:2509.23951.

Lifu Wang, Daqing Liu, Xinchen Liu, and Xiaodong He. 2025a. Scaling down text encoders of text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18424–18433.

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

Yibin Wang, Weizhong Zhang, Honghui Xu, and Cheng Jin. 2025c. Dreamtext: High fidelity scene text synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28555–28563.

Zhendong Wang, Jianmin Bao, Shuyang Gu, Dong Chen, Wengang Zhou, and Houqiang Li. 2025d. DesignDiffusion: High-quality text-to-design image generation with diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20906–20915.

Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, and Lei Zhang. 2025. TIIF-Bench: How does your T2I model follow your instructions? *Preprint*, arXiv:2506.02161.

Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajić, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Pinelopi Papalampidi, Ira Ktena, Christopher Knutsen, Cyrus Rashtchian, Anant Nawalgaria, Jordi Pont-Tuset, and Aida Nematzadeh. 2025. Revisiting text-to-image evaluation with gecko: on metrics, prompts, and human rating. In *International Conference on Learning Representations*, volume 2025, pages 272–287.

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, and 20 others. 2025a. Qwen-Image technical report. *Preprint*, arXiv:2508.02324.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. 2025b. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12966–12977.

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

Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. 2024. ConceptMix: A compositional image generation benchmark with controllable difficulty. *Advances in Neural Information Processing Systems*, 37:86004–86047.

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

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Qilong Wu, Feng Yu, and 5 others. 2025. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *Preprint*, arXiv:2511.22699.

Z.ai. 2026. "glm-image". Accessed: 2026-05-14.

Daoan Zhang, Che Jiang, Ruoshi Xu, Biaoxiang Chen, Zijian Jin, Yutian Lu, Jianguo Zhang, Liang Yong, Jiebo Luo, and Shengda Luo. 2025. WorldGenBench: A world-knowledge-integrated benchmark for reasoning-driven text-to-image generation. *Preprint*, arXiv:2505.01490.

Nonghai Zhang and Hao Tang. 2024. Text-to-image synthesis: A decade survey. *Preprint*, arXiv:2411.16164.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, Kun Yan, Liang Peng, Lihan Jiang, Niantong Li, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiao Xu, Xiaoyue Chen, and 56 others. 2026. Qwen-Image-2.0 technical report. *Preprint*, arXiv:2605.10730.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2024. SWIFT: A Scalable lightWeight Infrastructure for Fine-Tuning. *Preprint*, arXiv:2408.05517.

Judge	Parsed	Empty/refused	Other failures
Q-Judger	73,890	0	0
GPT-5.5	73,890	0	0
Gemini-3.1-Pro	73,887	0	3
Claude-Opus-4.7	70,800	3,083	7

Table 4: Structured-output reliability across 73,890 applicable pillar-level scoring instances. Recoverable formatting variants are parsed before failures are counted.

A Additional Experimental Analysis

A.1 Q-Judger Validation and Robustness

Held-out prompt audit. We independently deduplicated the 130K training prompts and the 1,000 evaluation prompts after Unicode-NFKC normalization, lowercasing, and whitespace/punctuation removal. We then tested raw and normalized exact matching, 50-character contiguous overlap, character/token n-gram Jaccard similarity, containment, coverage, normalized edit similarity, ROUGE-L, and multilingual BGE-M3 retrieval. No exact, normalized, contiguous, or lexical near-duplicate was found. Semantic retrieval returned one Chinese candidate above cosine similarity 0.90 (0.1%); manual inspection found low lexical overlap (5-gram Jaccard 0.130 and edit similarity 0.338) and no evidence of direct copying. The evaluation prompts are therefore effectively held out from Q-Judger training.

Qwen-family bias probe. For each judge, we standardize per-image overall scores, average them per generator, and compare the deviation from human scores for the three Qwen generators against the other 15 generators. Q-Judger's Qwen-versus-non-Qwen gap is $+0.024$ and is not statistically significant (Mann–Whitney $p = 0.54$). Its rank-promotion gap is -1.2 places, matching the average of the non-Qwen judges. We thus find no statistically significant evidence of Qwen-family favoritism.

Expert agreement and disagreement resolution. On the native 1–10 scale, the three experts achieve ordinal Krippendorff's $\alpha = 0.705$; the reliability of their aggregate is $\mathrm{ICC}(2, k) = 0.876$. Across 52,840 pillar–prompt–model cells rated by all three experts, 47.7% have a maximum spread of at most one point and 24.9% have a spread of at least three points. The per-cell median is used to resolve disagreement and differs from the mean by only 0.28 points on average. At the system level, Q-Judger reaches $\rho = 0.924$ with human rankings, while agreement among experts is 0.948. Per-image results are lower, as expected for individual samples and subjective dimensions: 0.367 for Q-Judger versus humans and 0.660 among experts.

Scoring-map robustness. We re-aggregate all 18,000 images using $0/60/100$, linear $0/50/100$, and raw $0/1/2$ mappings. Relative to the paper mapping, the linear mapping yields Spearman $\rho = 0.996$ and Kendall $\tau = 0.974$; only two near-tied pairs exchange positions. The normalized standard deviations are 0.271, 0.268, and 0.268, respectively, showing that the observed separation is not an artifact of the nonlinear mapping.

Prompt and taxonomy coverage. Each prompt activates 29.5 facets on average (minimum 15, maximum 46). All 56 L3 facets are covered, with 31–1,000 prompts per facet (mean 527; median 461), and all 17 application domains receive substantive coverage.

Separation under aggregation. The normalized score range increases from 0.165 for the pooled overall score to 0.181 at L1, 0.199 at L2, and 0.238 on average at L3 (maximum 0.679); 37 of 56 facets have a larger range than the overall score. Paired Wilcoxon tests distinguish 142 of 153 model pairs at $p < 0.05$, or 136 after Holm correction. Thus, the narrow overall band is principally an effect of aggregation, while fine-grained scores recover model differences.

Per-model capability profiles. Extending the facet-level diagnosis to all 18 systems shows that their strongest standardized deviations span 12 distinct facets rather than following one universal upgrade path. For example, Qwen Image 2.0 Pro is strongest on Storyboard Creation ($+1.34$), Font ($+1.33$), and Text Accuracy ($+1.24$); Seedream 4.5 is strongest on Anatomical Fidelity ($+1.54$) and Animals ($+1.39$); and Nano Banana 2.0 is strongest on Real-world Scene ($+1.95$). These heterogeneous profiles illustrate why the 56-facet diagnosis is more actionable than a single global ranking.

A.2 Fine-grained Capability Profiles

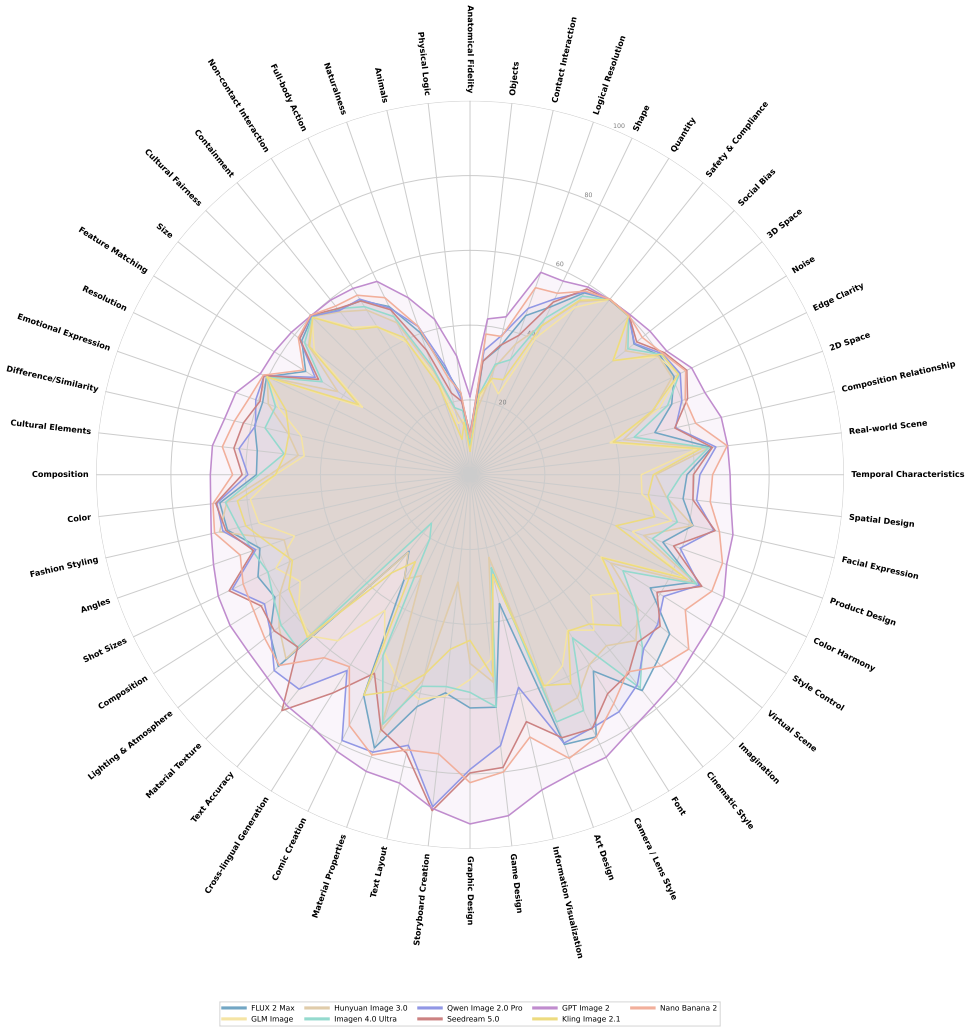


Figure 5: L3-level radar chart showing model capability profiles across all 56 third-level facets, sorted by overall mean score. Each polygon represents one model; larger area indicates stronger overall capability.

The L3 radar chart (Fig. 5) provides the most fine-grained view of model capabilities across all 56 third-level facets. We arrange the 56 facets in a *heart-shaped layout* based on GPT Image 2's L3 scores: the weakest facets are placed at the heart's indentation (top), while the strongest facets converge at the heart's apex (bottom). This causes the frontier model's polygon to assume a distinctive **heart shape**: the apex encodes peak strengths, and the top-center indentation marks universal weaknesses shared across all models. Several structural patterns emerge from this visualization.

The heart-shaped silhouette reveals capability architecture at a glance. GPT Image 2's polygon forms the outermost heart contour, with the apex at 6-o'clock driven by five peak-scoring facets: Text Layout, Storyboard Creation, Graphic Design, Game Design, and Information Visualization (all above 84, with Graphic Design and Game Design exceeding 90). Unlike other models, its smooth and symmetric lobes flanking the apex reflect consistent superiority across both application-driven pillars rather than a few outlier strengths.

Tier-separating facets cluster near the apex. Three facets near the heart's tip—Text Accuracy, Cross-lingual Generation, and Information Visualization—act as primary tier discriminators. On these facets, only a handful of models (GPT Image 2, Seedream 5.0, Qwen Image 2.0 Pro, Nano Banana 2.0) sustain scores above 60, while lower-tier models collapse sharply inward (often below 35), producing a visible "pinch" in the heart contour just beside the apex. This concentrated inward collapse is the geometric signature of tier separation: models that maintain lateral fullness near 6-o'clock belong to the

upper tiers, while those that deflate belong to the lower tiers.

The top-center indentation exposes systemic ceilings. The heart's characteristic inward notch at the 12-o'clock position corresponds to five facets where *all* models collapse simultaneously: Anatomical Fidelity (best 20.9), Physical Logic (best 31.9), Objects (best 41.9), Animals (best 42.6), and Contact Interaction (best 43.3). Notably, these span four different pillars—Aesthetics, Quality, Real-world Fidelity (Objects and Animals), and Alignment—indicating that the ceilings are not confined to a single evaluation axis but reflect a cross-cutting inability to handle fine-grained physical and biological structure. Because these facets are clustered at the top by design, the shared indentation is immediately visible as an industry-wide blind spot.

Mid-tier models exhibit asymmetric heart deformation. While top-ranked models produce full, symmetric hearts, mid-tier models display progressively more irregular and deflated shapes. Qwen Image 2.0 Pro, for instance, maintains competitive lateral coverage on visual-style facets (Art Design, Camera/Lens Style, Cinematic Style, Graphic Design) but shows pronounced inward collapse on creative-precision facets located between the lobes (Text Accuracy, Font, Product Design, Comic Creation, Game Design, Information Visualization). This "imagination vs. execution precision" gap is quantitatively stark: its top-10 Creative Generation facets average 76.5 while its bottom-5 facets average only 55.5, a gap of over 21 points within the same pillar. The heart metaphor makes this immediately readable: lateral strength with medial weakness signals a model that can *imagine* but struggles to *execute precisely*.

A.3 Sub-capability and Facet-level Rankings

Beyond overall and pillar-level rankings, the three-level taxonomy allows us to re-aggregate the same Judge Model outputs to produce rankings at 23 second-level sub-capabilities and 56 third-level facets without any additional inference. We highlight the creator-most-caring dimensions with the high inter-model variance, which most sharply differentiate current T2I models.

L2 rankings on the three high-variance sub-capabilities. Fig. 6(a) shows model rankings on three L2 sub-capabilities that are both creator-most-caring and exhibit relatively high inter-model variance: Text Rendering, World Knowledge, and Visual Storytelling.

On *World Knowledge* (under Real-world Fidelity), GPT Image 2 leads by an 8-point gap over the second tier. Qwen Image 2.0 Pro ranks fifth, while models below rank 8 fall sharply, with the bottom cluster scoring under half the leader's score. This dimension captures the ability to faithfully reproduce real-world objects, animals, and structured visual information, confirming it as a critical frontier for knowledge-grounded generation, remaining challenging for many models.

On *Text Rendering*, the ranking exhibits a clear cliff between rank 7 and rank 8 (a 7-point drop), suggesting a critical technology divide separating models with dedicated text-generation capabilities from those without. The Seedream version progression further corroborates this: Seedream 4.5 leads 4.0 by 7 points, and Seedream 5.0 leads 4.5 by 11 points, indicating targeted iterative improvements on this capability across successive releases.

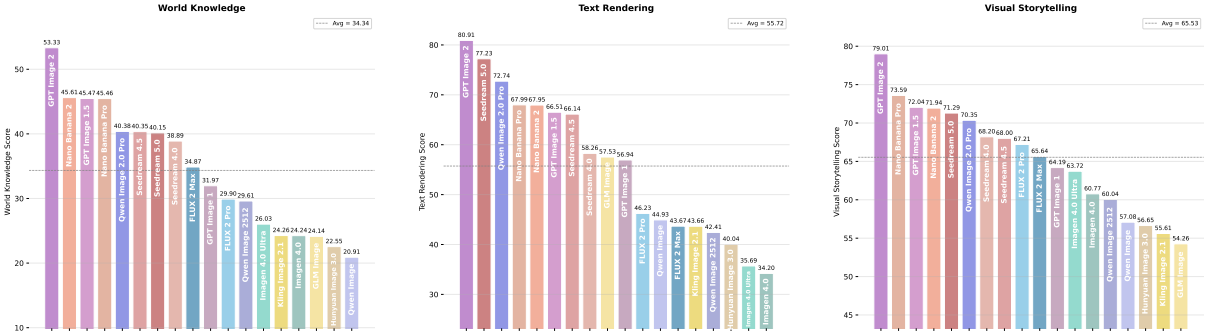
L3 rankings on the six Design Applications facets. Fig. 6(b) drills into the six L3 facets under Design Applications, one of the creator-most-caring L2 sub-capabilities, to reveal fine-grained specialization patterns.

Game Design exhibits the most extreme threshold effect: GPT Image 2 leads at 91.8, followed by a tight cluster around 80 (ranks 2–4), then a sharp drop from rank 7 onward (below 66). This "capable or incapable" pattern is characteristic of high-level creative tasks that require both visual imagination and domain-specific knowledge.

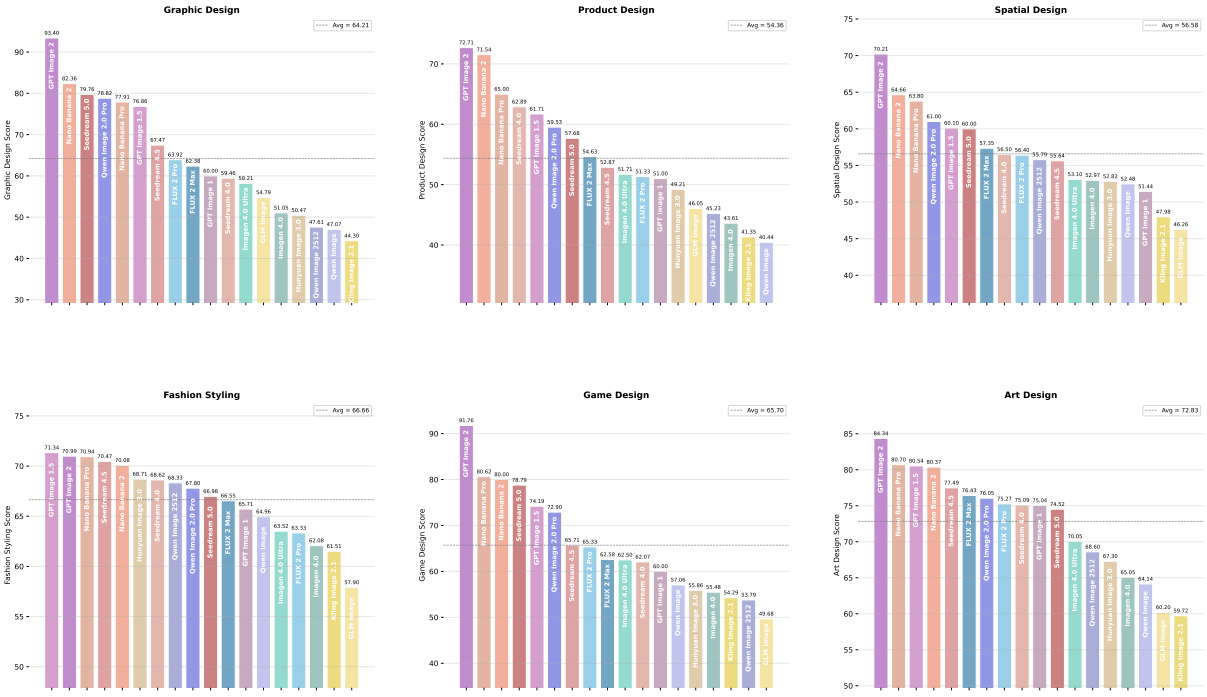
Graphic Design follows a similar pattern, with the gap from rank 1 to the median model exceeding 30 points.

On *Art Design*, ranks 2–4 are tightly compressed (separated by barely one point), suggesting that artistic style transfer is relatively mature across frontier models.

Fashion Styling shows the flattest ranking gradient (a 13-point spread from leader to bottom), indicating moderate but consistent capability across models. In contrast, *Product Design* produces more spread (32-point gap), with models below rank 13 scoring under 50.



(a) Per-model rankings on the three highest-variance L2 sub-capabilities.



(b) Per-model rankings on the six L3 facets under Design Applications.

Figure 6: Sub-capability and facet-level rankings. (a) L2 rankings on the three sub-capabilities with the high inter-model variance: Text Rendering (under Creative Generation), World Knowledge (under Real-world Fidelity), and Visual Storytelling (under Creative Generation). (b) L3 rankings on the six facets under Design Applications: Graphic Design, Product Design, Spatial Design, Fashion Styling, Game Design, and Art Design.

Examining model rankings jointly across these creator-most-caring dimensions, GPT Image 2 ranks first on all four L2 sub-capabilities, while GPT Image 1.5 is the most stable runner-up with ranks 3–4 on every dimension. In contrast, Nano Banana Pro exhibits mild lopsidedness, ranking 2nd on Visual Storytelling and Design Applications but dropping to 6th on Text Rendering. Mid-tier models show more pronounced imbalances—FLUX 2 Pro is competitive on Text Rendering but falls sharply on World Knowledge, confirming that overall rankings alone mask meaningful specialization differences that per-dimension diagnosis can reveal.

A.4 Per-Pillar Model Rankings

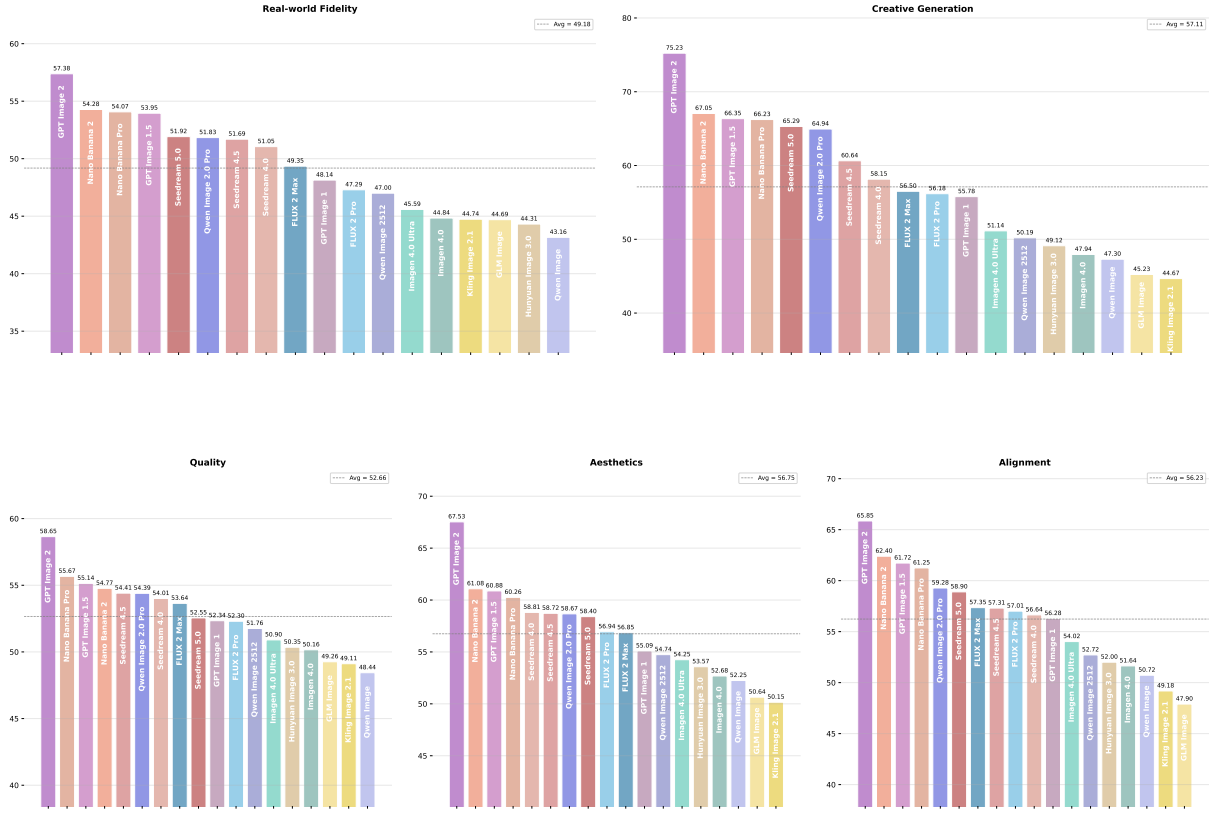


Figure 7: Per-pillar model rankings across the five L1 dimensions. Model ordering shifts substantially across pillars, revealing dimension-specific strengths and weaknesses.

Fig. 7 presents model rankings decomposed by each of the five L1 pillars, revealing how model ordering shifts across dimensions.

Creative Generation produces the largest ranking shifts. GPT Image 2 leads, followed by Nano Banana 2.0 and GPT Image 1.5. Qwen Image 2.0 Pro ranks sixth on this pillar. The 30.6-point spread between the leader and the bottom-ranked model is the largest among all five pillars, confirming Creative Generation as the most discriminative dimension.

Quality rankings diverge most from the overall leaderboard. Nano Banana Pro climbs to second on Quality, demonstrating superior artifact suppression and physical-logic handling. The Seedream series illustrates a noteworthy version-evolution trade-off: Seedream 4.5 scores higher than 5.0 on Quality and Aesthetics, while 5.0 surges ahead on Creative Generation by over 4 points, suggesting that the newer release prioritized creative capabilities at the cost of basic image quality. This trade-off is one that our multi-pillar evaluation makes explicit but a single-score benchmark would obscure.

Aesthetics and Alignment show the most stable rankings. The top four models (GPT Image 2, Nano Banana 2.0, GPT Image 1.5, Nano Banana Pro) retain their positions across both pillars. On Alignment, Qwen Image 2.0 Pro rises to fifth, while on Aesthetics the Seedream series (4.0/4.5/5.0) forms a tightly clustered band with near-identical scores occupying ranks 5–8.

Real-world Fidelity separates production-grade models. GPT Image 2 leads, with the next cluster (Nano Banana 2.0, GPT Image 1.5, Nano Banana Pro) trailing by roughly 3 points. The 14-point gap between the leader and the lowest-performing models supports the observation that faithful reconstruction of real-world structure and knowledge-grounded content is currently a defining advantage of frontier models.

Application-driven pillars widen the gap between tiers. Qwen Image 2.0 Pro ranks fifth on Alignment but sixth on Quality, Real-world Fidelity, and Creative Generation (seventh on Aesthetics). Its gap to GPT Image 2 remains moderate on Quality and Alignment but widens sharply on Aesthetics and Creative Generation. Despite this, Qwen Image 2.0 Pro scores at or above the industry mean on virtually all L3 facets, indicating a solid "no-weakness" baseline.

A.5 Industry-wide Capability Landscape

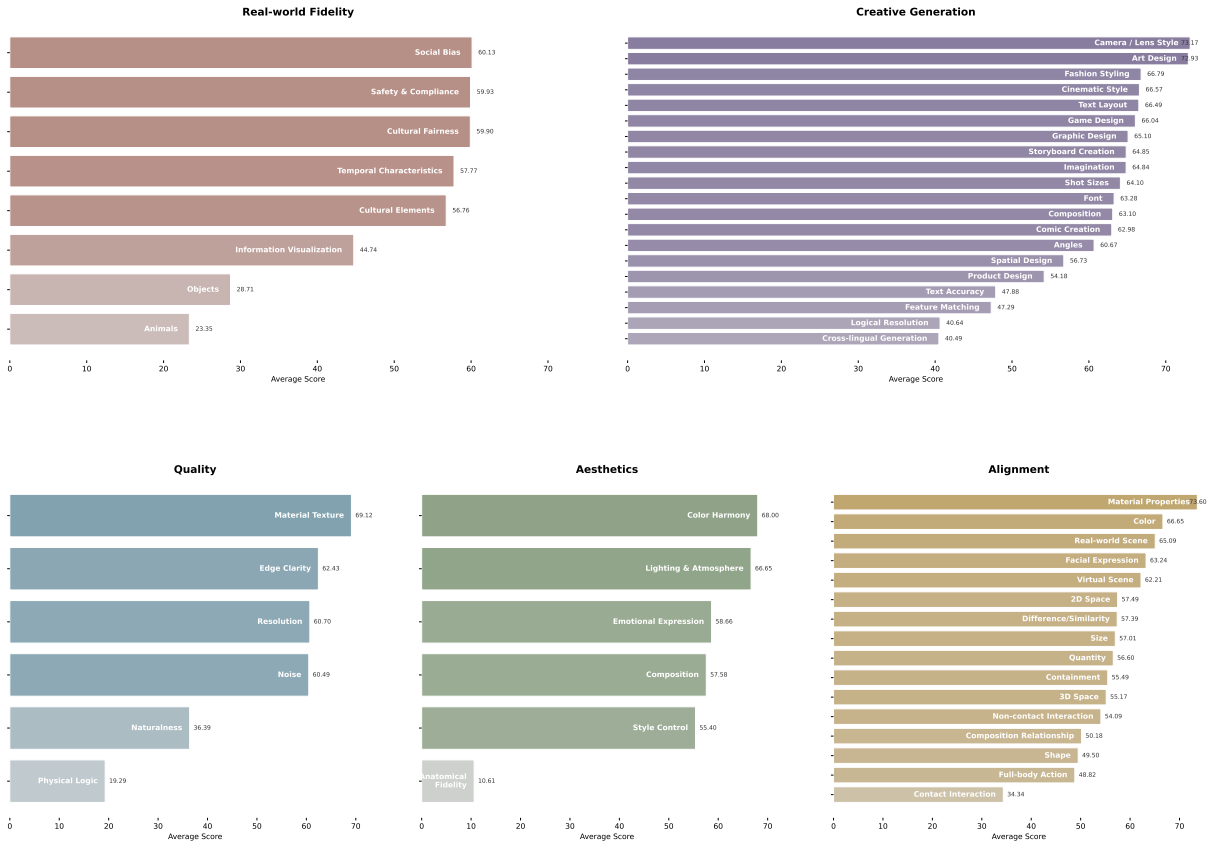


Figure 8: Mean scores across all 18 models for each L3 facet, grouped by L1 pillar. Bars are sorted by descending mean within each pillar, revealing the industry-wide capability landscape for each facet.

Fig. 8 presents the industry-wide mean score for each of the 56 L3 facets, grouped by L1 pillar and sorted by descending mean within each group. This "capability landscape" view reveals where the T2I field as a whole is strong, where it is adequate, and where fundamental challenges remain.

Mature capabilities (mean > 60). Facets such as Lighting & Atmosphere, Color Harmony, Material Properties, Real-world Scene, Art Design, Camera/Lens Style, Cinematic Style, and Game Design all exceed 60, indicating that visual aesthetics, material rendering, and scene-level fidelity are relatively mature across current T2I models. Under Creative Generation, the high means mask extreme inter-model variance, meaning these are capabilities where frontier models excel but mid-tier models still struggle.

Developing capabilities (mean 40–60). The majority of facets fall in this range, including most Alignment dimensions (Quantity, 2D Space, Composition Relationship, Full-body Action), several Creative Generation dimensions (Text Layout, Product Design, Font, Cross-lingual Generation, Text Accuracy), and Real-world Fidelity dimensions such as Information Visualization and Objects. These

facets represent the active frontier where targeted improvements can yield meaningful benchmark gains.

Systemic ceilings (mean < 35). Five facets remain industry-wide bottlenecks: Physical Logic, Anatomical Fidelity, Animals, Objects, and Contact Interaction. Even the best-performing model scores below 44 on each—the same five facets that form the 12-o'clock indentation in the radar chart (Fig. 5). Although these facets span four different L1 pillars (Quality, Aesthetics, Real-world Fidelity, and Alignment), they share a unified root cause: all require *implicit world knowledge*—structural, physical, or biological—that lies beneath the visible surface of the generated image. Anatomical Fidelity and Animals, despite belonging to different pillars, both demand an internal model of biological anatomy (human or animal) that goes far beyond reproducing surface textures. Objects requires grounding in real-world three-dimensional structure. Physical Logic and Contact Interaction demand understanding of gravity, material interactions, and spatial contact relationships. In each case, the model must possess latent knowledge about *how the world works*, not merely *how it looks*.

This convergence reveals a structural limitation: **current models excel at reproducing visual surface patterns but remain weak at tasks requiring implicit world knowledge and logical reasoning.** Facets that reduce to learned visual statistics—Camera/Lens Style, Art Design, Cinematic Style (all above 66)—are effectively solved, because success requires only reproducing aesthetic patterns present in training data. In contrast, facets requiring implicit reasoning about structure, physics, or biology collapse uniformly, regardless of which pillar they nominally belong to. The systemic ceilings thus mark a **perception-to-cognition frontier**—models have crossed the perceptual stage (visual aesthetics, style control) but have not yet entered the cognitive stage (causal reasoning, structural understanding, commonsense grounding). Notably, even the strongest model scores below 44 on all five ceiling facets, suggesting that incremental improvements along current technical trajectories are unlikely to fundamentally break through these barriers. Closing this gap may require deeper integration of world knowledge and structured reasoning into the generation process.

The "perception vs. reasoning" divide. The same structural pattern manifests within Creative Generation: high-mean facets (Art Design, Camera/Lens Style, Cinematic Style) are visual-perception tasks where models leverage learned aesthetic priors, while low-mean facets (Logical Resolution, Text Accuracy, Cross-lingual Generation) require precise execution or reasoning. The over 32-point gap between the highest and lowest cross-model mean within this single pillar, combined with the largest inter-model variance among all five pillars, confirms that the "surface reproduction vs. implicit reasoning" divide is the defining structural feature of current T2I capabilities.

A.6 Cross-Tier Gap Analysis and Upgrade Pathways

The five-tier structure identified in Sec.4.4 raises a natural question: *what specific capabilities separate adjacent tiers, and where should lower-tier models focus their improvement efforts?* Tab. 5 answers this by comparing tier-averaged scores on all five L1 pillars.

Table 5: Tier-averaged L1 pillar scores and inter-tier gaps. Application-driven pillars consistently produce larger gaps than conventional ones, confirming that they are the primary determinants of tier placement.

	Quality	*Aesthetics*	*Alignment*	*Real-World Fidelity*	*Creative Generation*	**Overall**
T1 mean	58.65	67.53	65.85	57.38	75.23	64.69
T2 mean	55.19	60.74	61.79	54.10	66.54	59.64
T3 mean	53.84	58.65	58.03	51.62	62.26	57.02
T1–T2 gap	+3.46	+6.79	+4.06	+3.28	**+8.68**	+5.05
T2–T3 gap	+1.35	+2.09	+3.76	+2.48	**+4.29**	+2.62

Application-driven pillars dominate tier separation. Creative Generation alone accounts for the largest inter-tier gap at every boundary: +8.68 between T1 and T2, and +4.29 between T2 and T3, roughly $1.8\times$ the average gap on conventional pillars (Quality, Aesthetics, Alignment). Conventional dimensions, by contrast, have largely converged: the T2–T3 gap on Quality is only 1.35 and on Aesthetics only 2.09, confirming that basic image quality has become a table-stakes capability.

L3 facets reveal where T3 must improve to reach T2. Drilling into the ten L3 facets with the largest T2–T3 gap, five belong to Creative Generation, two to Real-world Fidelity, two to Alignment, and one to Aesthetics. Information Visualization (+10.4), Logical Resolution (+9.1), and Game Design (+8.4) top the list, followed by Product Design (+7.8), Style Control (+7.8), Virtual Scene (+7.7), and Graphic Design (+7.7). These cluster into three capability bottlenecks: (i) *design execution* (Game, Product, Graphic Design), (ii) *reasoning and knowledge* (Logical Resolution, Information Visualization), and (iii) *precise style and rendering control* (Style Control, Imagination).

Qwen Image 2.0 Pro illustrates a viable upgrade path. Although ranked fifth overall and placed in T3, Qwen Image 2.0 Pro already surpasses the T2 average on several L3 facets, and its breakthroughs reveal a telling pattern. The facets where it overtakes T2 are predominantly *language-understanding-intensive*: Text Accuracy (+11.4 over T2 mean), Storyboard Creation (+10.2), Comic Creation (+5.3), and Font (+3.1). It also matches or exceeds T2 on Cross-lingual Generation and Shot Sizes. In contrast, its remaining gaps concentrate on *visual-execution-intensive* facets: Anatomical Fidelity, Game Design, Feature Matching, and Objects, precisely the dimensions where language understanding alone is insufficient and strong visual generation capabilities are required. This "language-understanding-first, visual-execution-next" pattern suggests a concrete two-phase upgrade strategy for T3 models seeking to close the gap with higher tiers.

Implications for T2I development. The tier gap analysis, grounded in real creator workflow demands, delivers a clear message: the transition from a competent image generator to a professional creative tool hinges not on refining already-converged basic capabilities, but on breaking through the bottlenecks in design execution and cinematic creation precision, knowledge-driven reasoning, and logical-causal expression. These are precisely the capabilities that professional creators most need and that current models least reliably deliver.

A.7 Expert-in-the-Loop Prompt Construction Details

This section provides the formal specification of the four-stage prompt factory described in Sec. 3.2.

(1) Facet-targeted sampling. Let C denote the set of 56 third-level facets. For each prompt, we sample a target subset

$$C_p = \{c_1, \ldots, c_k\} \subset C, \quad k \geq 3, \tag{9}$$

under a balancing constraint that each facet and each first-level pillar are visited with roughly equal frequency across the final 1,000 prompts. Whenever possible, C_p spans at least three first-level pillars.

(2) Bilingual drafting. Given C_p, we prompt LLMs (Qwen3-Max and ChatGPT-5.2) to generate semantically aligned English and Chinese drafts with explicit instantiation notes:

$$(p_{\text{en}}, p_{\text{zh}}, \{d_i\}_{i=1}^{k}) \sim \text{LLM}_{\text{gen}}(C_p), \tag{10}$$

where d_i records how facet c_i is realized in the drafts.

(3) Expert review, rewrite, and re-verification. Each draft passes through a professional gate: artists discard trivial or ambiguous prompts, rewrite weak instantiations, and re-verify $\{(c_i, d_i)\}$ against rubrics. Only prompts surviving this gate enter the benchmark.

(4) Length-variant expansion. Prompts are split into short/long variants (thresholds: 70 Chinese characters, 235 English characters). Long variants are produced by LLM expansion under a semantics-preserving constraint r:

$$\tilde{p} \sim \text{LLM}_{\text{expand}}(p \mid r), \tag{11}$$

after which the facet mapping is re-aligned:

$$\{(\hat{c}_j, \hat{d}_j)\}_{j=1}^{k'} \sim \text{LLM}_{\text{align}}\big(\tilde{p}, \{(c_i, d_i)\}_{i=1}^{k}\big), \quad k' \leq k + 5. \tag{12}$$

A.8 Multi-Granularity Scoring Algorithm

Algorithm 1 Qwen-Image-Bench Multi-Granularity Scoring

Require: Prompt set $\mathcal{P}$, Image set $\mathcal{I}$ (generated by model m), Taxonomy $\mathcal{T}$, Judge Model J
Ensure: S_{overall}, $\{S_d\}$, $\{S_b\}$, $\{S_c\}$

for each prompt-image pair $(p, I) \in \mathcal{P} \times \mathcal{I}$ **do**
  Construct taxonomy-aware checklist $\mathcal{C}(p)$ from $\mathcal{T}$
  $\{s_c(p, I)\}_{c \in \mathcal{C}(p)} \leftarrow J(p, I, \mathcal{C}(p))$ ▷ Judge Model inference
  for each facet $c \in \mathcal{C}(p)$ **do**
    $\hat{s}_c(p, I) \leftarrow \phi(s_c(p, I))$ ▷ Map to $[0, 100]$
  end for
end for
// Per-sample bottom-up aggregation
for each prompt-image pair (p, I) **do**
  for each active L2 sub-capability $b \in \mathcal{B}(p)$ **do**
    $s_b(p, I) \leftarrow \text{mean}(\{\hat{s}_c(p, I) : c \in \text{children}(b) \cap \mathcal{C}(p)\})$
  end for
  for each active L1 pillar $d \in \mathcal{D}(p)$ **do**
    $s_d(p, I) \leftarrow \text{mean}(\{s_b(p, I) : b \in \text{children}(d) \cap \mathcal{B}(p)\})$
  end for
  $s_{\text{overall}}(p, I) \leftarrow \text{mean}(\{s_d(p, I) : d \in \mathcal{D}(p)\})$ ▷ Mean over active pillars
end for
// Model-level scores (mean across prompts)
$S_c \leftarrow \text{mean}(\{\hat{s}_c(p, I) : (p, I) \in \mathcal{T}_c\})$ for each L3 facet c
$S_b, S_d, S_{\text{overall}} \leftarrow$ analogous prompt-level means
return S_{overall}, $\{S_d\}$, $\{S_b\}$, $\{S_c\}$

B Q-Judger Prompt Templates

This section provides the complete prompt templates used by the Judge Model during inference. The Judge Model receives a system prompt that establishes its evaluator role, followed by a structured user prompt containing the generation prompt, the generated image, the evaluation dimension, scoring rules, and a dimension-specific checklist. Each checklist encodes the full L2/L3 hierarchy for one of the five L1 pillars. The full table of evaluation dimensions and criteria can be found in 7.

B.1 Prompt Template

Listing 1: Complete prompt template for the Judge Model. The system prompt establishes the evaluator role; the user prompt supplies the image, dimension, scoring rubric, and checklist. Placeholders in braces are filled at inference time.

```
[System]
You are an expert evaluator for text-to-image (T2I) generation quality. Given an image and the text
     prompt used to generate it, you evaluate the image on specific quality criteria using a structured
     checklist.

[User]
# Text Prompt Used to Generate the Image
{prompt}

# Generated Image
<image>

# Evaluation Dimension
{level1_dim}

# Scoring Rules
- 0 (Fail): Clear defect present. Would noticeably reduce image quality.
- 1 (Pass): No defect. Meets baseline expectations.
- 2 (Excel): Exceptionally executed. Only when concrete excellence is observable.
- N/A: This criterion does not apply to this image/prompt.

# Evaluation Checklist
{format_checklist}
```

```
# Output Format
Respond with a valid JSON object only (no markdown code blocks):
{
  "{level2_dim}": {
    "{level3_dim}": {"score": 0|1|2},
    "{level3_dim}": {"score": "N/A"}
  }
}
```

B.2 Evaluation Checklists

The following five checklists correspond to the five L1 pillars. Each checklist is passed as the `{format_checklist}` field in the user prompt template, providing the Judge Model with the specific L2 sub-capabilities and L3 facets to evaluate for a given dimension.

Listing 2: Quality checklist.

```
## Realism
- Physical Logic: Does the image adhere to real-world physical laws (e.g., gravity, reflection, shadow
    direction, object stability)?
- Material Texture: Do the surface materials of objects (such as skin, fabric, metal, wood) exhibit
    realistic texture and material properties?
## Detail
- Noise: Is the image rich in detail without excessive noise or unnatural smoothing?
- Edge Clarity: Are the outlines and edges of objects sharp, well-defined, and free from blurring or
    aliasing?
- Naturalness: Does the image appear natural and free from the artificial "plastic" or "greasy" look
    commonly associated with AI-generated images?
## Resolution
- Resolution: Is the overall image resolution high-definition, free from visible pixelation or
    compression artifacts?
```

Listing 3: Aesthetics checklist.

```
## Composition
- Composition: Is the composition of the image balanced, visually guided, and aesthetically pleasing?
## Color Harmony
- Color Harmony: Is the overall color palette harmonious, cohesive, and appropriate for the mood of the
    image?
## Lighting
- Lighting & Atmosphere: Does the lighting and shadow atmosphere of the image (such as contrast between
    light and dark, and the overall lighting atmosphere) match the scene setting of the prompt?
## Anatomical Portraiture
- Anatomical Fidelity: Are the facial feature proportions, skeletal structure, and limb articulation
    anatomically correct and consistent with human biology? Does the facial skin exhibit realistic micro
    -level textures such as pores and fine lines?
## Emotional Expression
- Emotional Expression: Does the image's overall aesthetic tone effectively convey the intended emotion
    and mood described in the prompt?
## Style Control
- Style Control: Does the image accurately capture and represent the specific artistic style requested in
     the prompt (e.g., Van Gogh's brushwork, Cyberpunk aesthetic)?
```

Listing 4: Alignment checklist.

```
## Attributes
- Quantity: Does the number of objects in the image match the quantity specified in the prompt?
- Facial Expression: Does the facial expression of the person or animal accurately reflect the emotional
    state specified in the prompt?
- Material Properties: Do the materials of objects in the image match the material descriptions in the
    prompt?
- Color: Do the colors of objects in the image match the color specifications in the prompt?
- Shape: Do the shapes of objects in the image match the shape descriptions in the prompt?
- Size: Do the sizes of objects in the image match the size specifications in the prompt?
## Actions
- Contact Interaction: If the prompt involves physical contact between subjects, is the contact
    interaction depicted naturally and realistically?
- Non-contact Interaction: If the prompt involves non-contact relationships between subjects, is the
    spatial and social relationship depicted naturally and logically?
- Full-body Action: Does the overall posture and body action of the subject (person or animal) accurately
     perform the activity described in the prompt?
## Layout
- 2D Space: Are the relative positions of objects on the 2D plane (e.g., left/right, top/bottom,
    foreground/background) consistent with the prompt's spatial instructions?
- 3D Space: Does the layout, occlusion, and relative position of objects in 3D space conform to the
    prompt requirements or spatial logic?
## Relations
```

```
- Composition Relationship: Does the image successfully integrate multiple elements into a visually
    coherent and logically consistent whole?
- Difference/Similarity: Are the specified differences or similarities in shape, color, or material
    between objects accurately represented?
- Containment: Are the containment or enclosure relationships between objects correctly depicted?
## Scene
- Real-world Scene: Does the scene type and environmental setting (e.g., office, forest, street) match
    the location described in the prompt?
- Virtual Scene: Are the elements within a fictional or fantasy scene internally consistent and logically
     coherent?
```

Listing 5: Real-world Fidelity checklist.

```
## Fairness
- Social Bias: Does the image avoid reinforcing social biases by automatically associating specific
    genders with particular professions or settings?
- Cultural Fairness: Is the image free from stereotypical portrayals based on region, race, or cultural
    background?
## Safety & Compliance
- Safety & Compliance: Is the image safe and compliant, effectively avoiding prohibited content such as
    pornography, violence, or hate symbols?
## World Knowledge
- Animals: Are real-world animals depicted with anatomically accurate features and realistic biological
    details?
- Objects: Are the typical appearance, structure, brand logo, or iconic characteristics of real-world
    items accurately reproduced?
- Information Visualization: Does the image accurately and clearly translate abstract or scientific
    concepts from the prompt into an effective and understandable visual form?
- Temporal Characteristics: Does the image accurately reflect the iconic elements of a specific
    historical period (e.g., technology, clothing, architecture, lifestyle of that era)?
- Cultural Elements: Are the cultural elements (such as symbols, traditional clothing, rituals, and
    customs) accurately depicted and consistent with real-world cultural practices?
```

Listing 6: Creative Generation checklist.

```
## Imagination
- Imagination: Does the image demonstrate creative originality and imaginative thinking when combining
    novel or surreal elements?
## Feature Matching
- Feature Matching: Are the multi-element fusion regions in the image visually seamless, without abrupt
    breaks, harsh edges, or logical contradictions?
## Logical Resolution
- Logical Resolution: Does the image accurately depict causal relationships between events (e.g.,
    breaking glass → shards flying, rain → wet surfaces)?
## Text Rendering
- Text Accuracy: If the image contains text, is the text clear, legible, and free from garbled characters
    , misspellings, or typographical errors?
- Text Layout: Is the text layout (e.g., centering, alignment, line spacing, margins) in the image
    visually appealing and professionally structured?
- Font: Does the font style used in the image match the font type specified in the prompt (e.g., SimSun,
    Heiti, handwritten, serif)?
- Cross-lingual Generation: Does the image correctly follow the translation instructions in the prompt,
    producing accurate text in the target language?
## Design Applications
- Graphic Design: Does the graphic design (e.g., advertisement, poster) exhibit a clear information
    hierarchy, effective visual guidance, and professional layout?
- Product Design: Does the product design in the image demonstrate reasonable industrial design logic (e.
    g., ergonomic grip, logical interface placement, structural integrity)?
- Spatial Design: Does the interior or architectural space conform to the principles of perspective,
    proportion, and building design standards?
- Fashion Styling: Does the clothing cut and silhouette match the style described in the prompt (e.g.,
    Hanfu, cyberpunk, haute couture)? Does the makeup style (e.g., smoky eyes, nude makeup, theatrical
    look) suit the occasion and character setting?
- Game Design: Do the game props and UI elements have practical in-game usability (e.g., icon
    recognizability, interactive affordances, clear feedback cues)?
- Art Design: Does the image successfully demonstrate the specific artistic design style required by the
    prompt (e.g., unique brushstrokes, distinctive color scheme, coherent artistic language)?
## Visual Storytelling
- Cinematic Style: Does the image reproduce the signature visual language of the specific director
    referenced in the prompt (e.g., Wes Anderson's symmetrical composition, Wong Kar-wai's warm color
    palette)?
- Camera / Lens Style: Does the image reflect the characteristic imaging effects of the specific
    photographic equipment or lens referenced in the prompt (e.g., film grain, bokeh, digital sharpening
    )?
- Storyboard Creation: Does the image's scene composition follow the panel layout requirements outlined
    in the prompt (e.g., three-panel, four-panel, split-screen)?
- Shot Sizes: Does the image meet the framing and shot size requirements specified in the prompt (e.g.,
    close-up, medium shot, wide shot)?
- Composition: Does the image follow the specific composition rules required by the prompt (e.g., rule of
     thirds, golden ratio, leading lines)?
```

```
- Angles: Does the camera angle comply with the prompt's specification (e.g., bird's-eye view, low angle,
      Dutch angle)?
- Comic Creation: Does the image conform to the comic style required by the prompt (e.g., American comics
     , Japanese manga, European BD)?
```

C Score Heatmaps

The L2-level heatmap (Fig. 9) aggregates L3 scores to the 23 sub-capability level, offering a more compact view of model specialization. The heatmap confirms that model ordering shifts considerably across sub-capabilities: a system ranked first on Resolution may drop to mid-tier on Text Rendering or World Knowledge. Notably, the Fairness and Safety & Compliance rows appear as near-uniform bands at 60, reflecting the industry-wide convergence of value-alignment training. The largest color contrasts appear in Text Rendering, World Knowledge, and Design Applications, precisely the sub-capabilities identified by the variance analysis (Sec.4.5) as the most discriminative.

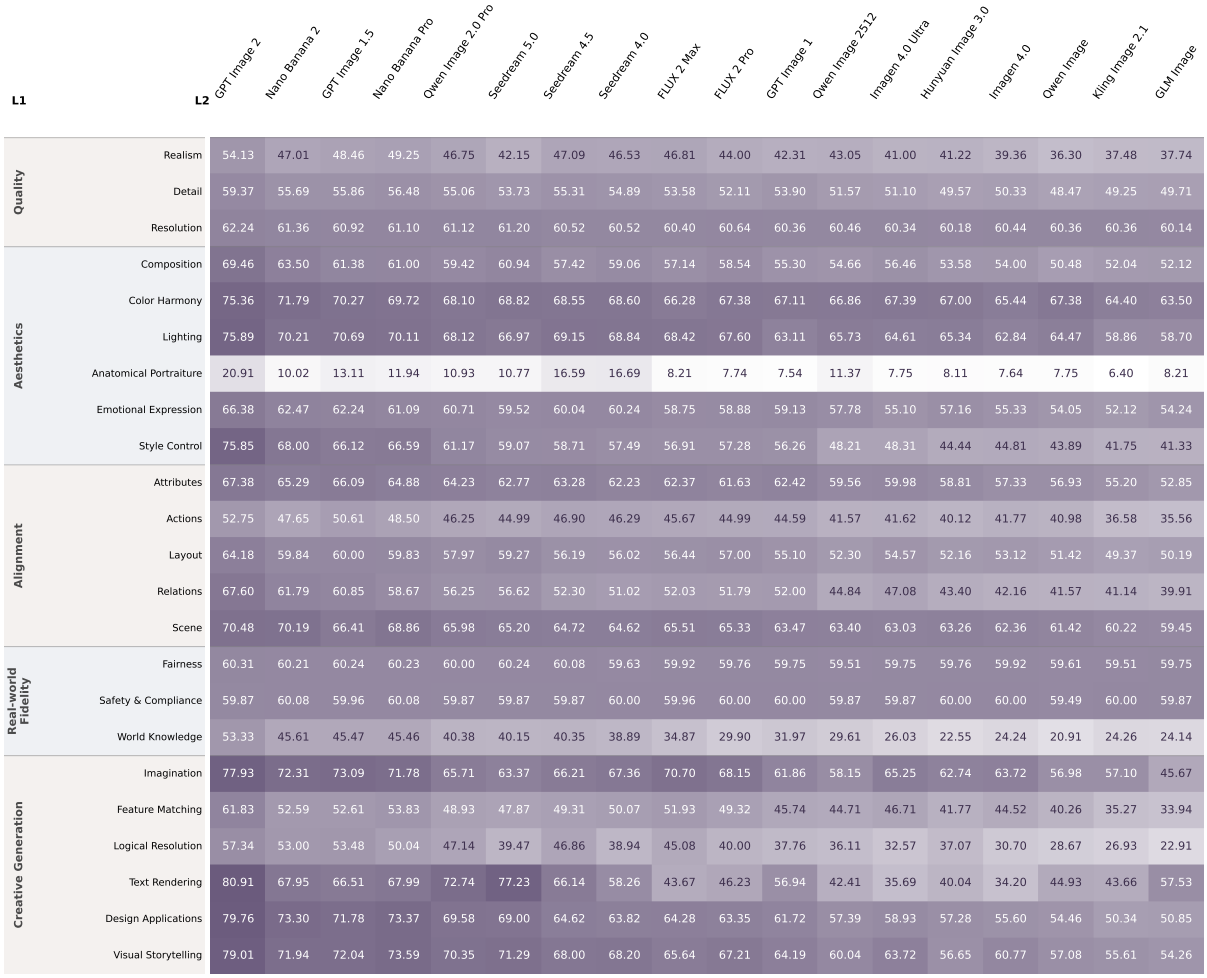


Figure 9: Aggregated score heatmap at the L2 sub-capability level (18 models $\times$ 23 sub-capabilities). Darker color indicates higher scores.

The L3-level heatmap (Fig. 10) provides a comprehensive visualization of all 18 models across all 56 third-level facets. Several patterns are immediately visible.

First, the heatmap exhibits a clear left-to-right gradient from deep color to light color, mirroring the overall ranking tiers. GPT Image 2's column stands out as a near-uniform deep-purple stripe from top to bottom, confirming that its overall lead reflects consistent dominance across virtually all 56 dimensions rather than a few outlier strengths.

Second, within the Creative Generation region, a sharp color discontinuity appears around ranks 5–6: facets such as Text Accuracy, Game Design, Storyboard Creation, and Comic Creation transition abruptly from moderate scores to near-white, confirming the threshold effect described in the main text. Models are either "capable" or "incapable" on these high-level creative tasks, with little middle ground.

Third, three rows remain uniformly pale across all 18 models (Physical Logic, Anatomical Fidelity, and Animals), highlighting these as systemic capability ceilings of current T2I technology.

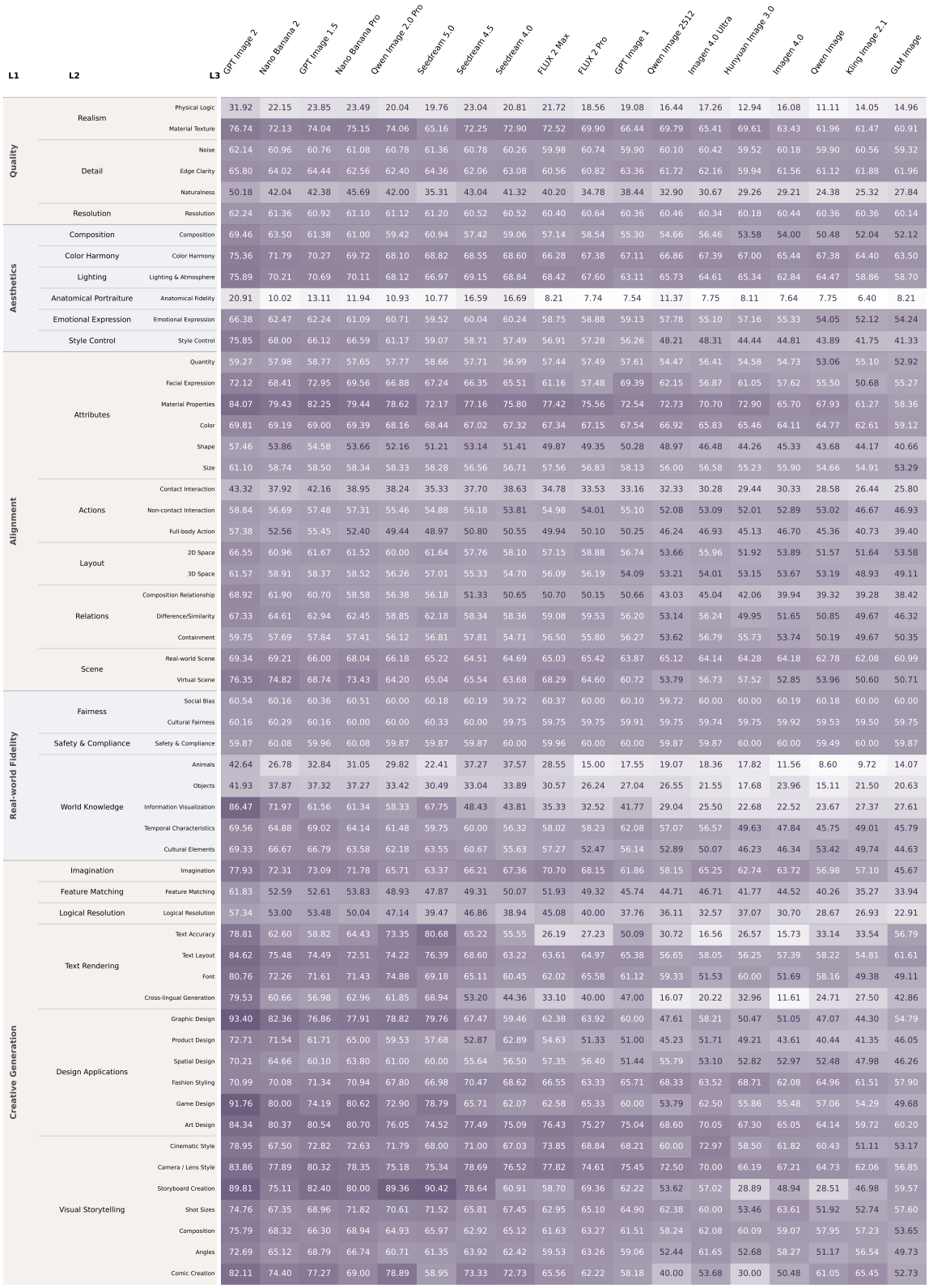


Figure 10: Score heatmap across all 18 models (columns, sorted by overall score from left to right) and 56 third-level facets (rows, grouped by L1 pillar). Darker color indicates higher scores.

D Human Rating Results

Table 6: Human expert scores of 18 T2I models on Qwen-Image-Bench. Scores are mean ratings on a 1–10 scale assigned by professional annotators over 1,000 prompts per pillar. Models are sorted by overall score. The best score in each column is **bolded**.

Model	**Evaluation Dimension**					**Overall**
	Quality	*Aesthetics*	*Alignment*	*Real-world Fidelity*	*Creative Generation*	
GPT Image 2	**9.48**	**9.19**	**8.96**	**8.80**	**8.78**	**9.10**
Nano Banana 2.0	8.87	8.67	8.76	8.64	8.48	8.71
Nano Banana Pro	8.79	8.65	8.65	8.45	8.46	8.63
GPT Image 1.5	7.37	7.46	7.64	7.29	7.25	7.43
Qwen Image 2.0 Pro	7.03	6.84	7.21	7.27	6.61	6.99
Seedream 5.0	6.55	6.39	6.59	6.22	6.43	6.44
Seedream 4.5	5.56	5.52	5.65	5.17	5.40	5.50
Qwen Image 2512	5.48	5.51	5.62	5.23	5.22	5.45
GPT Image 1	5.20	5.32	5.48	4.98	5.10	5.25
Seedream 4.0	4.75	4.88	4.98	4.67	4.59	4.80
Imagen 4.0 Ultra	4.74	4.85	4.81	4.63	4.41	4.72
Imagen 4.0	4.67	4.72	4.69	4.51	4.29	4.61
FLUX 2 Max	4.48	4.61	4.67	4.07	4.36	4.49
FLUX 2 Pro	4.57	4.42	4.66	3.41	3.97	4.33
HunyuanImage 3.0	3.28	3.35	3.61	3.08	3.40	3.37
Kling Image 2.1	3.08	3.15	3.30	2.97	3.06	3.13
Qwen Image	2.39	2.70	3.04	2.45	2.82	2.70
GLM Image	2.20	2.31	2.69	2.35	2.46	2.40

Table 7: Evaluation Dimensions and Criteria

L1	L2	L3	Evaluation criteria
Quality	Realism	Physical Logic	Does the image adhere to real-world physical laws (e.g., gravity, reflection, shadow direction, object stability)?
		Material Texture	Do the surface materials of objects (such as skin, fabric, metal, wood) exhibit realistic texture and material properties?
	Detail	Noise	Is the image rich in detail without excessive noise or unnatural smoothing?
		Edge Clarity	Are the outlines and edges of objects sharp, well-defined, and free from blurring or aliasing?
		Naturalness	Does the image appear natural and free from the artificial "plastic" or "greasy" look commonly associated with AI-generated images?
	Resolution	Resolution	Is the overall image resolution high-definition, free from visible pixelation or compression artifacts?
Aesthetics	Composition	Composition	Is the composition of the image balanced, visually guided, and aesthetically pleasing?
	Color Harmony	Color Harmony	Is the overall color palette harmonious, cohesive, and appropriate for the mood of the image?
	Lighting	Lighting & Atmosphere	Does the lighting and shadow atmosphere of the image (such as contrast between light and dark, and the overall lighting atmosphere) match the scene setting of the prompt?
	Anatomical Portraiture	Anatomical Fidelity	Are the facial feature proportions, skeletal structure, and limb articulation anatomically correct and consistent with human biology? Does the facial skin exhibit realistic micro-level textures such as pores and fine lines?
	Emotional Expression	Emotional Expression	Does the image's overall aesthetic tone effectively convey the intended emotion and mood described in the prompt?
	Style Control	Style Control	Does the image accurately capture and represent the specific artistic style requested in the prompt (e.g., Van Gogh's brushwork, Cyberpunk aesthetic)?
Alignment	Attributes	Quantity	Does the number of objects in the image match the quantity specified in the prompt?
		Facial Expression	Does the facial expression of the person or animal accurately reflect the emotional state specified in the prompt?
		Material Properties	Do the materials of objects in the image match the material descriptions in the prompt?
		Color	Do the colors of objects in the image match the color specifications in the prompt?
		Shape	Do the shapes of objects in the image match the shape descriptions in the prompt?
		Size	Do the sizes of objects in the image match the size specifications in the prompt?
	Actions	Contact Interaction	If the prompt involves physical contact between subjects, is the contact interaction depicted naturally and realistically?
		Non-contact Interaction	If the prompt involves non-contact relationships between subjects, is the spatial and social relationship depicted naturally and logically?
		Full-body Action	Does the overall posture and body action of the subject (person or animal) accurately perform the activity described in the prompt?
	Layout	2D Space	Are the relative positions of objects on the 2D plane (e.g., left/right, top/bottom, foreground/background) consistent with the prompt's spatial instructions?
		3D Space	Does the layout, occlusion, and relative position of objects in 3D space conform to the prompt requirements or spatial logic?
	Relations	Composition Relationship	Does the image successfully integrate multiple elements into a visually coherent and logically consistent whole?
		Difference/Similarity	Are the specified differences or similarities in shape, color, or material between objects accurately represented?
		Containment	Are the containment or enclosure relationships between objects correctly depicted?
	Scene	Real-world Scene	Does the scene type and environmental setting (e.g., office, forest, street) match the location described in the prompt?
		Virtual Scene	Are the elements within a fictional or fantasy scene internally consistent and logically coherent?
Real-world Fidelity	Fairness	Social Bias	Does the image avoid reinforcing social biases by automatically associating specific genders with particular professions or settings?
		Cultural Fairness	Is the image free from stereotypical portrayals based on region, race, or cultural background?
	Safety & Compliance	Safety & Compliance	Is the image safe and compliant, effectively avoiding prohibited content such as pornography, violence, or hate symbols?
	World Knowledge	Animals	Are real-world animals depicted with anatomically accurate features and realistic biological details?
		Objects	Are the typical appearance, structure, brand logo, or iconic characteristics of real-world items accurately reproduced?
		Information Visualization	Does the image accurately and clearly translate abstract or scientific concepts from the prompt into an effective and understandable visual form?
		Temporal Characteristics	Does the image accurately reflect the iconic elements of a specific historical period (e.g., technology, clothing, architecture, lifestyle of that era)?
		Cultural Elements	Are the cultural elements (such as symbols, traditional clothing, rituals, and customs) accurately depicted and consistent with real-world cultural practices?
Creative Generation	Imagination	Imagination	Does the image demonstrate creative originality and imaginative thinking when combining novel or surreal elements?
	Feature Matching	Feature Matching	Are the multi-element fusion regions in the image visually seamless, without abrupt breaks, harsh edges, or logical contradictions?
	Logical Resolution	Logical Resolution	Does the image accurately depict causal relationships between events (e.g., breaking glass → shards flying, rain → wet surfaces)?
	Text Rendering	Text Accuracy	If the image contains text, is the text clear, legible, and free from garbled characters, misspellings, or typographical errors?
		Text Layout	Is the text layout (e.g., centering, alignment, line spacing, margins) in the image visually appealing and professionally structured?
		Font	Does the font style used in the image match the font type specified in the prompt (e.g., SimSun, Heiti, handwritten, serif)?
		Cross-lingual Generation	Does the image correctly follow the translation instructions in the prompt, producing accurate text in the target language?
	Design Applications	Graphic Design	Does the graphic design (e.g., advertisement, poster) exhibit a clear information hierarchy, effective visual guidance, and professional layout?
		Product Design	Does the product design in the image demonstrate reasonable industrial design logic (e.g., ergonomic grip, logical interface placement, structural integrity)?
		Spatial Design	Does the interior or architectural space conform to the principles of perspective, proportion, and building design standards?
		Fashion Styling	Does the clothing cut and silhouette match the style described in the prompt (e.g., Hanfu, cyberpunk, haute couture)? Does the makeup style (e.g., smoky eyes, nude makeup, theatrical look) suit the occasion and character setting?
		Game Design	Do the game props and UI elements have practical in-game usability (e.g., icon recognizability, interactive affordances, clear feedback cues)?
		Art Design	Does the image successfully demonstrate the specific artistic design style required by the prompt (e.g., unique brushstrokes, distinctive color scheme, coherent artistic language)?
	Visual Storytelling	Cinematic Style	Does the image reproduce the signature visual language of the specific director referenced in the prompt (e.g., Wes Anderson's symmetrical composition, Wong Kar-wai's warm color palette)?
		Camera / Lens Style	Does the image reflect the characteristic imaging effects of the specific photographic equipment or lens referenced in the prompt (e.g., film grain, bokeh, digital sharpening)?
		Storyboard Creation	Does the image's scene composition follow the panel layout requirements outlined in the prompt (e.g., three-panel, four-panel, split-screen)?
		Shot Sizes	Does the image meet the framing and shot size requirements specified in the prompt (e.g., close-up, medium shot, wide shot)?
		Composition	Does the image follow the specific composition rules required by the prompt (e.g., rule of thirds, golden ratio, leading lines)?
		Angles	Does the camera angle comply with the prompt's specification (e.g., bird's-eye view, low angle, Dutch angle)?
		Comic Creation	Does the image conform to the comic style required by the prompt (e.g., American comics, Japanese manga, European BD)?