DESIGNERBENCH: BENCHMARKING IMAGE GENERATION FOR REAL-WORLD DESIGN WORKFLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Image generative models are increasingly used in professional design, yet evaluation for design applications has not matured into a coherent practice attuned to real design workflows. Evaluation of generative systems in design is not a question of overall output preference, but of whether the operational requirements of a given workflow are satisfied. Existing benchmarks, however, remain largely confined to text-to-image synthesis and generic editing, leaving workflow-specific failures largely unexamined. To close this gap, we introduce **DESIGNERBENCH**, a workflow-centric benchmark that evaluates generative models as design collaborators, **spanning 2,000 curated instances across seven design task families**: asset creation, element extraction, multi-source composition, professional editing, rendering, multi-view dissection, and design language extension, which together reflect the canonical progression of real design projects. Rather than reducing evaluation to a single holistic score, **DESIGNERBENCH** scores each output against a sample-specific, multidimensional checklist of task-specific objectives and preservation constraints, using an **evidence-first vision-language judge with a four-level rubric**. Our evaluation of 14 state-of-the-art models reveals substantial variation in performance across design tasks, with rankings shifting according to the operation under evaluation. These results challenge the notion of a single best generative model, underscoring the need for workflow-aware evaluation.

1 INTRODUCTION

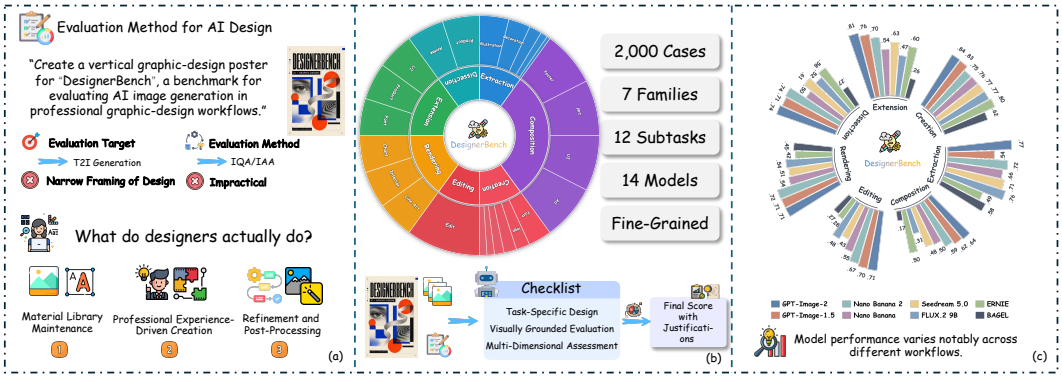


Figure 1: Overview of **DESIGNERBENCH**. (a) Motivation. (b) Benchmark composition and evaluation pipeline. (c) Model performance across seven workflows.

The rapid progress of generative AI over the past several years has begun to reshape professional design practice — not incrementally, but in ways that are already visible in the tools designers use, the workflows they follow, and the expectations they bring to their work. Designers report that AI has become a routine part of ideation, prototyping, and production, with adoption rates climbing to 91% weekly usage (Designer Fund et al., 2026; Adobe, 2025). Yet systematic evaluation of generative performance on tasks central to design practice has not kept pace with this adoption.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This lag reflects a deeper mismatch between existing evaluation paradigms and the realities of professional design, evident in three respects. ***First, prevailing benchmarks overlook the processual nature of design work***, in which designers extend and adapt existing artifacts across iterative stages; outputs are therefore intermediate states rather than terminal deliverables (Schön, 2017; Lawson, 2006). Even design-oriented benchmarks remain confined to isolated capabilities such as generation or controllability (Li et al., 2026; Liang et al., 2025), offering no systematic evaluation of design performance across workflow stages. ***Second, existing benchmarks construe design restrictively***, reducing it to a single task type rather than a broad professional domain. In practice, design spans a wide spectrum of media and application contexts, from brand identities and editorial systems to digital products and spatial layouts, each governed by its own conventions, constraints, and success criteria. Although recent benchmarks have moved beyond pure generation, they still cover only a narrow slice of this spectrum (Zhang et al., 2026; Xu et al., 2026; Yi et al., 2026), leaving much of professional design beyond their scope. ***Third, the benchmarks discussed above remain limited in scale***, with most containing no more than 1,000 evaluation instances. Such scale is insufficient for drawing reliable conclusions about model competence across varied design constraints, and their evaluation criteria often remain coarse. Evaluation therefore remains fragmented, and no large, systematically structured benchmark currently supports fine-grained analysis of design performance.

To address these gaps, we introduce **DESIGNERBENCH**, a workflow-centric benchmark that evaluates generative systems on the design operations that arise across real design workflows. **DESIGNERBENCH** is built around three principles. First, it is ***practice-oriented***: its tasks are defined by what designers actually do rather than by what models can already generate well. Second, it is ***multi-type***: it spans seven complementary workflow families, namely asset generation, element extraction, multi-source composition, professional editing, design language extension, multi-view dissection, and structural rendering, which together reflect the breadth of professional design work. Third, it is ***multi-dimensional***: rather than scoring each output holistically, it evaluates against sample-specific checklists that encode task-specific objectives and preservation constraints, scored by an evidence-first vision-language judge on a 0–3 rubric. As can be seen in Figure 1, **DESIGNERBENCH** comprises 2,000 curated instances across 12 scoring units, with 8 to 16 criteria per sample depending on task complexity. We evaluate 14 state-of-the-art models under this protocol and find that no single model dominates across all operations and that controllability and reference fidelity, not aesthetic quality, remain the primary bottlenecks, particularly for open-source models.

Our contributions are threefold: ***(i) a workflow-centric benchmark*** that captures the breadth of real design practice across 2,000 instances and 7 workflow families; ***(ii) an evidence-first evaluation protocol*** that enforces diagnostic, reproducible scoring through structured checklists and vision-language judging; and ***(iii) a comprehensive empirical study*** of 14 models across all operations, showing that no single model dominates, that rankings shift substantially across tasks, and that different models exhibit distinct weaknesses across operations.

2 RELATED WORK

2.1 IMAGE GENERATION AND EDITING MODELS

Image generation models first established prompt-driven synthesis as the dominant paradigm (Ramesh et al., 2022; Rombach et al., 2022), with evaluation centered on prompt alignment and perceptual quality (Huang et al., 2023; Ghosh et al., 2023). Subsequent work shifted toward editing under textual instructions or spatial controls (Brooks et al., 2023; Meng et al., 2021), with evaluation targeting edit fidelity and content preservation (Ku et al., 2023). More recent systems unify generation and editing within a single architecture and support reference-conditioned synthesis (Black Forest Labs, 2026; Zhao et al., 2026; OpenAI, 2026a; NanoBanana, 2026). These advances have motivated new benchmarks (Huang et al., 2025; Yang et al., 2026), but evaluations remain centered on isolated operations, such as prompt alignment, aesthetic quality, or single-step editing, and do not characterize how model behavior changes as a design artifact progresses through a workflow.

2.2 DESIGN-ORIENTED BENCHMARKS

Prior work on design-oriented evaluation has examined isolated aspects of graphic design, including content-aware layout and element placement (Hsu et al., 2023), text rendering accuracy (Zhang et al.,

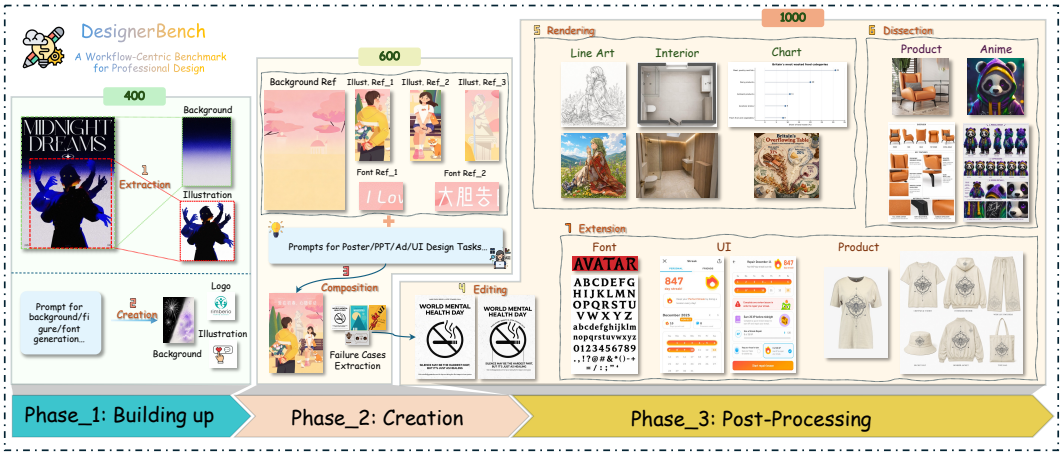


Figure 2: Task examples in **DESIGNERBENCH**. The benchmark comprises 2000 cases across three design phases: Building up, Creation, and Post-Processing.

2025). More recent benchmarks broaden task coverage but remain confined to isolated capabilities, such as generation (Liang et al., 2025; Li et al., 2026), controllability (Zhang et al., 2026), and editing (Xu et al., 2026; Yi et al., 2026). **DESIGNERBENCH** is complementary: it focuses on raster generation and editing across workflow transformations such as extraction, composition, rendering, dissection, and extension. Its central object is not only the final design but the correspondence among the instruction, each ordered reference, and the properties that may or may not change.

3 DESIGNERBENCH

Professional design is processual: designers build on existing assets and constraints, work across media and task types, and iterate through intent, composition, and revision before a result is usable. To systematically evaluate generative models on the workflow transformations that arise in real design practice, we construct **DESIGNERBENCH** based on three design principles: practice-oriented, multi-type, and validity. The benchmark comprises seven task families (§3.1), where success depends not on standalone image quality but on changing only what is authorized, preserving what is fixed, and treating each input as evidence for the property it constrains. To realize this, we build an asset-centric data construction pipeline with expert-validated prompts (§3.2) and a sample-specific checklist evaluation protocol with an evidence-first vision-language judge (§3.3).

3.1 TASK DESCRIPTION

To ground the task families in real design practice, we surveyed professional designers with relevant industry experience and distilled seven recurring workflow families. All tasks share a common interface: each is instantiated as a tuple $(I, \mathcal{V}, O)$, where I is the design instruction, $\mathcal{V}$ is an optional set of visual inputs, and O is the required output. Creation is prompt-driven with $\mathcal{V} = \emptyset$; Composition is multi-image to single-image, taking 2–8 ordered references; the remaining families are image-to-image with a single source S. We describe each family below (Figure 2).

Creation. Generate reusable design assets, including backgrounds, fonts, icons, logos, decorations, and illustrations, driven solely by the prompt.

Extraction. Recover a prompt-targeted element from a source design while preserving internal detail, thin parts, enclosed holes, and semi-transparent edges without neighbor contamination.

Composition. Orchestrate ordered references into a complete deliverable. The model must integrate fonts, layout, attributes, and visual elements so that each reference is bound to its intended role; a composition can look coherent while assigning the wrong reference to the wrong role.

Editing. Apply a design-oriented instruction to an existing design, making a composite change while preserving text semantics, protected content, and unmodified regions.

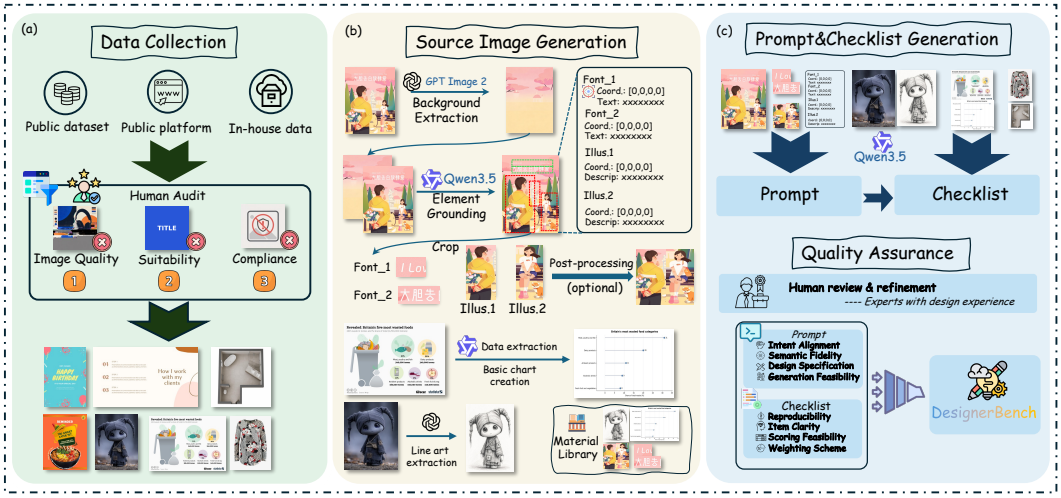


Figure 3: Data construction pipeline of **DESIGNERBENCH**. (a) Data collection and human audit. (b) Source image generation. (c) Prompt and checklist generation with quality assurance.

Rendering. Translate preliminary or structured inputs into presentable outputs, where line-art and floor-plan rendering preserve reference geometry while applying prompt-specified appearance, and chart rendering produces a creative presentation while preserving the data truth of the source chart.

Dissection. Generate multiple views, expressions, or detail panels of the same character or product, constrained by cross-panel identity, view coverage, and structural consistency. This family tests whether a model can decompose an artifact without inventing a different object.

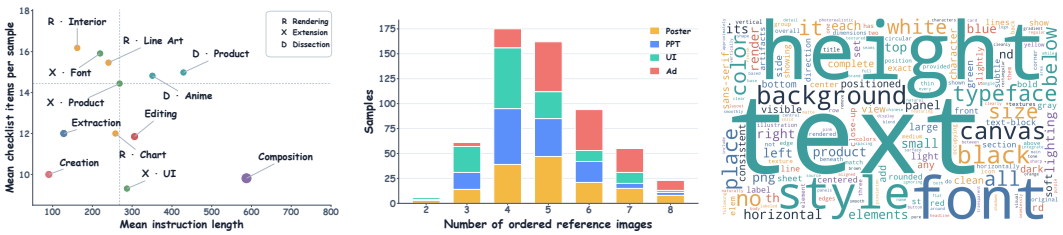
Extension. Create related variants while separating requested change from invariant identity, where product and font extension evaluate pattern/material preservation, typeface metrics, and family-level continuity, while UI extension generates an adjacent page that preserves the source product's visual language, text semantics, and component system.

3.2 DATA CONSTRUCTION

The data construction of **DESIGNERBENCH** follows an asset-centric strategy that grounds all tasks in a common design substrate.

Collection. Our data construction combines a shared layout-derived core with task-specific source collections. For the shared core, we curate professionally designed advertisements, posters, presentation pages, and mobile/web interfaces from open-source datasets, public platform, and in-house data. As illustrated in Figure 3, a dedicated pipeline first identifies the background and then decomposes each layout into provenance-linked font exemplars, text groups, and visual elements, forming a reusable material library, while recording the design plan with element positions, layout, color scheme, and layer relations. Beyond this core, we additionally collect product and anime images as well as creative data tables for the remaining families. Decomposed assets are further augmented to remove residual information that could affect generation, and together with the intact layouts they form a reusable design corpus that links Creation, Extraction, Composition, and Editing through shared artifacts rather than disconnected datasets.

Task Construction. The material library and design plans from the previous pipeline support Creation, Extraction, Composition, and Editing, while the remaining families use task-specific sources. As can be seen in Figure 3, in every family, a VLM drafts the generation instruction from these inputs. For Creation, the VLM describes an element decomposed by the pipeline (one of six asset types), and that description is used as the text-only prompt. Extraction conditions on a source design and a named target, asking the model to recover that element. Composition conditions on the design plan and ordered decomposed assets, asking the model to assemble a complete layout. Editing conditions on a flawed Composition result and a specified change, asking for localized repair or a



(a) Task landscape (b) Composition references (c) Prompt Word Clouds

Figure 4: Data analysis of **DESIGNERBENCH**.

common edit. Rendering conditions on a structural input together with appearance cues from the source—for example, line-art rendering feeds the original product or anime image and its reverse-traced sketch so the VLM can write a prompt to color the line art in the original style; floor plans start from generated sketches, and charts from extracted data with a refined visual scheme. Dissection conditions on a collected anime or product image and asks for multiple consistent views. Extension conditions on a source artifact to expand it in-family: Latin or Chinese glyphs from pipeline-extracted font images, design variants from product photos, or an adjacent UI page from collected screenshots or web pages, preserving the source visual language.

Data Analysis. The benchmark contains 2,000 task instances and approximately source images. Creation, Extraction, and Editing contribute 200 instances each. Composition is the largest family, with 600 multi-reference layouts. Rendering and Extension contribute 300 instances each, split evenly across three subunits (line art, interiors, and charts; fonts, products, and UI). Dissection contributes 200 instances (100 character images and 100 product images). Figure 1 shows this distribution. Creation is reference-free (200 instances). Every other family except Composition conditions on a single source image (1,200 instances). Only Composition uses multiple ordered references: 2–9 images per instance, with a mean of 5.12. Prompt lengths average 317.7 tokens (range 56–4,296).

Prompt quality validation. All prompts are validated, revised and refined by human experts with design experience along four dimensions: intent alignment, semantic fidelity, design specification, and generation feasibility. Intent alignment verifies that the prompt matches the intended design task; semantic fidelity ensures that subjects, attributes, and relations are accurate and unambiguous; design specification checks that sufficient design constraints such as style, color, material, and composition are provided; and generation feasibility assesses whether the prompt can be executed by models.

3.3 EVALUATION PROTOCOL

Holistic preference scores are insufficient for design evaluation: an output can be aesthetically strong while violating a non-negotiable constraint, and a single score provides no evidence for why it succeeds or fails. We therefore adopt checklist-based evaluation, scoring each instance against sample-specific criteria that target observable relationships, boundaries, mappings, or rendering behavior. These criteria treat each reference as evidence for a property rather than a pixel-level target, while the instruction remains authoritative about what may change.

Checklist construction. Each checklist is written against its instance: the prompt and valid reference images determine which visual obligations can be checked, so two samples in the same family can receive different tests. We decompose the requested workflow into operation points—elements, regions, structures, or representations to be changed or preserved—and translate each into a visually verifiable question. The prompt authorizes what may change, while reference images provide evidence for properties that must remain faithful. Harder or finer-grained operations have more independent failure points and therefore more criteria.

Checklists share a common schema of task dimensions and four-level rubrics. Each criterion records its dimension, observable test, authoritative reference, authorized-change or ignore scope, and criterion-specific 0–3 anchors, with provenance linked to the checklist version. The benchmark

Table 1: Family-level results on **DESIGNERBENCH**. Overall is the sample-level micro average over the 2,000-instance benchmark. and indicate the best and second-best reported scores. "–" denotes scoring units that are not supported because of the model's input format limitations.

Model	Generate	Extraction	Composition	Editing	Rendering	Extension	Dissection	Overall
Closed-source models								
GPT Image 2	**0.841**	**0.770**	**0.642**	**0.713**	0.715	**0.814**	**0.745**	**0.729**
GPT Image 1.5	0.831	0.537	0.625	0.703	0.712	0.763	0.715	0.688
NanoBanana 2	0.753	0.717	0.587	0.669	**0.720**	0.705	0.736	0.679
NanoBanana	0.758	0.661	0.499	0.550	0.535	0.544	0.497	0.559
Seedream 5.0	0.765	0.709	0.483	0.427	0.513	0.633	0.611	0.569
Seedream 4.5	0.728	0.536	0.460	0.314	0.426	0.520	0.442	0.482
Wan2.7-Image	0.755	0.717	0.501	0.638	0.585	0.572	0.652	0.601
Qwen-Image 2 pro	0.695	0.502	–	0.585	0.518	0.490	0.602	–
Open-source models								
FLUX.2 [klein] 9B	0.771	0.763	0.310	0.477	0.538	0.473	0.517	0.498
FLUX.2 [klein] 4B	0.712	0.505	0.273	0.414	0.484	0.391	0.336	0.410
ERNIE-Image	0.803	0.494	0.503	0.284	0.421	0.597	0.580	0.519
Qwen-Edit-2509	–	0.616	–	0.295	0.502	0.381	0.284	–
HiDream-I1	0.673	0.506	0.278	0.269	0.411	0.339	0.344	0.375
BAGEL	0.619	0.582	0.169	0.268	0.445	0.258	0.270	0.330

contains 25,438 scorable criteria (12.72 per instance). Budgets reflect independently observable constraints: Composition averages 9.81 criteria, while Extraction averages 11.35.

Scoring and aggregation. The 0–3 rubric captures partial compliance with criterion-specific anchors. For model m, sample n, dimension d, and criterion i, let $x_{mni} = s_{mni}/3 \in [0, 1]$ be the normalized criterion score and I_{nd} the criteria assigned to dimension d. We combine the mean criterion performance with the weakest criterion in that dimension:

$$\bar{D}_{mnd} = \frac{1}{|I_{nd}|} \sum_{i \in I_{nd}} x_{mni}, \quad D^{\min}_{mnd} = \min_{i \in I_{nd}} x_{mni}, \quad D_{mnd} = \alpha \bar{D}_{mnd} + (1 - \alpha) D^{\min}_{mnd}, \tag{1}$$

where $\alpha \in [0, 1]$ is a trade-off parameter that balances average performance against worst-case failure. We then aggregate dimensions using their frozen task-specific weights:

$$S_{mn} = \frac{\sum_d w_{nd} D_{mnd}}{\sum_d w_{nd}}. \tag{2}$$

The second term penalizes a single non-negotiable failure that would be hidden by a dimension mean, while the first term preserves graded credit across all criteria. Creation has no named dimension schema and therefore retains its original direct mean over prompt-specific criteria. Weights w_{nd} are determined by design experts to reflect the relative importance of each dimension for a given task (Table C.4). Composition uses equal dimension weights, while other tasks assign higher weights to their defining fidelity dimensions.

Judge protocol. For each instance, the judge receives the output, instruction, and only the task-authorized inputs: the prompt for Creation, a single source image for single-reference tasks, indexed references for Composition, and the plain chart for Chart Rendering. The judge evaluates the checklist for the instance and returns a score and evidence for each criterion. For each criterion, it records visible observations, a concrete measurement or reference comparison, and any visible deviations before assigning a score in $\{0, 1, 2, 3\}$; a score of 3 requires complete evidence. Scores are aggregated using the frozen Mean–Minimum rule and task-specific dimension weights.

Checklist quality validation. Because the benchmark contains 25,438 criteria, checklist validation uses a hybrid protocol combining automated screening with expert adjudication. An VLM first screens each criterion along four dimensions: reproducibility (whether the criterion is stated objectively enough to be judged independently), clarity (unambiguous wording), observability (judgeable from observable visual evidence), and dimension alignment (consistent with the dimension it is assigned to). Criteria flagged as problematic are adjudicated by human experts with design experience.

Table 2: Detailed scoring-unit results of **DESIGNERBENCH**. and mark the best and second-best reported scores; "–" denotes scoring units that are not supported because of the model's input format limitations.

Model	Composition				Rendering			Extension			Dissection	
	Poster	Ad	PPT	UI	Line Art	Floor	Chart	Product	Font	UI	Anime	Product
Closed-source models												
GPT Image 2	**0.656**	**0.615**	0.625	**0.672**	0.689	0.789	0.669	**0.751**	**0.816**	**0.875**	0.754	0.736
GPT Image 1.5	0.628	0.560	**0.643**	0.666	0.669	0.771	**0.696**	0.751	0.716	0.822	**0.756**	0.674
NanoBanana 2	0.628	0.476	0.614	0.623	**0.705**	**0.797**	0.659	0.737	0.620	0.757	0.692	**0.780**
NanoBanana	0.493	0.482	0.520	0.501	0.649	0.608	0.349	0.559	0.480	0.595	0.537	0.457
Seedream 5.0	0.480	0.442	0.498	0.508	0.574	0.660	0.306	0.689	0.551	0.659	0.656	0.567
Seedream 4.5	0.453	0.390	0.499	0.490	0.496	0.538	0.243	0.529	0.406	0.625	0.504	0.380
Wan2.7-Image	0.465	0.458	0.517	0.567	0.668	0.712	0.376	0.634	0.576	0.508	0.702	0.602
Qwen-Image 2 pro	–	–	–	–	0.689	0.561	0.303	0.566	0.410	0.494	0.662	0.542
Open-source models												
FLUX.2 [klein] 9B	0.337	0.348	0.328	0.226	0.680	0.727	0.208	0.543	0.467	0.408	0.570	0.464
FLUX.2 [klein] 4B	0.291	0.285	0.316	0.201	0.682	0.564	0.205	0.400	0.414	0.359	0.389	0.283
ERNIE-Image	0.473	0.422	0.547	0.569	0.527	0.457	0.252	0.636	0.536	0.612	0.683	0.474
Qwen-Edit-2509	–	–	–	–	0.629	0.669	0.209	0.438	0.479	0.226	0.325	0.242
HiDream-I1	0.281	0.313	0.305	0.214	0.567	0.505	0.162	0.420	0.456	0.140	0.374	0.315
BAGEL	0.174	0.144	0.207	0.153	0.566	0.595	0.173	0.278	0.285	0.211	0.319	0.221

4 EXPERIMENTS

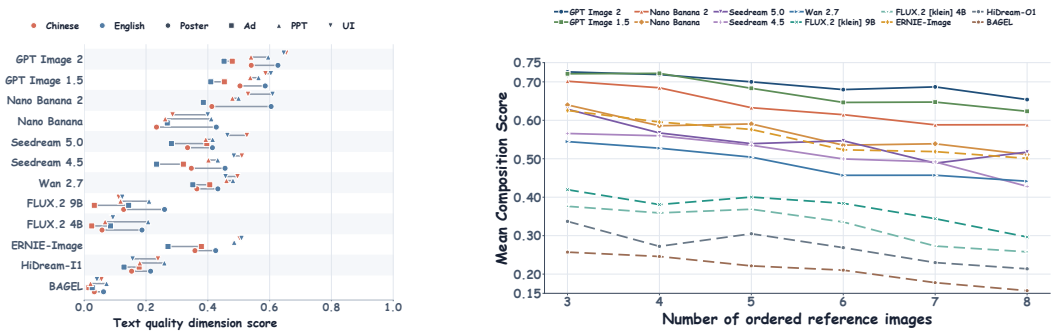
4.1 EXPERIMENTAL SETUP

We evaluate 14 image generation systems, including eight proprietary API models: GPT Image 1.5/2 (OpenAI, 2026a), NanoBanana 1/2 (NanoBanana, 2026), Wan2.7-image (Wan Team & Alibaba Group, 2026), Qwen-Image 2.0 pro (Qwen Team & Alibaba Group, 2026), and Seedream 4.5/5.0 (ByteDance, 2026), and six open or self-hosted models: FLUX.2 [klein] 4B/9B (Black Forest Labs, 2026), ERNIE-Image (Liu et al., 2026), Qwen-Edit-2509 (Wu et al., 2025), HiDream-I1 (Cai et al., 2025), and BAGEL (Deng et al., 2025). For each instance, every model receives the identical task instruction and ordered visual inputs as specified in §3.1. Instance scores follow the mean–minimum aggregation in §3.3. GPT5.4 (OpenAI, 2026b) is used as VLM judger in our work. We set $\alpha = 0.5$, so that a dimension score equally weights the mean criterion and the weakest criterion and therefore more strongly penalizes a single failed. Family scores are means over instances.

4.2 MAIN RESULTS

Different operations demand different capabilities. At the family level, GPT Image 2 is the most consistent operator, leading Creation, Extraction, Composition, Editing, Extension, and Dissection, but this lead does not extend to every scoring unit. ***Generation strength does not transfer to extraction***: GPT Image 2, GPT Image 1.5, and ERNIE-Image all exceed 0.80 on Creation, yet Extraction preserves a different order, and FLUX.2 [klein] 9B becomes the runner-up (0.763) despite falling to 0.310 on Composition. ***Editing tests localization rather than synthesis***: GPT Image 2 (0.713) and GPT Image 1.5 (0.703) lead, while ERNIE-Image shows a clear Creation–Editing gap and specialized editors (Qwen-Image 2 pro, Qwen-Edit-2509) do not match the strongest generalists. ***Results diverge further at the scoring-unit level***: NanoBanana 2 leads Rendering (0.720) and product dissection (0.780), GPT Image 1.5 leads chart rendering (0.696), and font extension favors models that preserve typeface metrics (GPT Image 2 0.816) rather than those with strong generation alone. A high score on one family is therefore not evidence of a reliable design operator across the workflow.

Composition and structured fidelity remain the main bottlenecks. Composition is the steepest drop for almost every system: BAGEL falls from 0.619 on Creation to 0.169 on Composition, both FLUX.2 [klein] variants remain at or below 0.310, and even the leading system reaches only 0.642. Multi-source composition requires assigning each reference a role and keeping text, layout, and assets aligned to those roles; attributes and composition quality stay high, while text and layout remain weak, so a composition can look fused while still corrupting copy or hierarchy. Rendering shows a related pattern: GPT Image 1.5 leads chart rendering (0.696) because the criterion is data accuracy rather than pictorial style, and open models that look plausible on product or anime rendering collapse on charts



(a) Text quality by type and language.

(b) Composition score versus reference count.

Figure 5: Text rendering quality analysis by type and language **(b)** and Composition score analysis by reference count **(a)**.

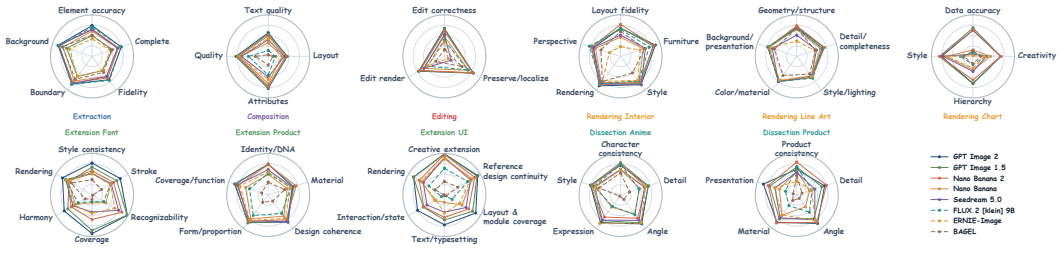


Figure 6: Dimension-level scores for eight representative models. Each panel is one scoring unit; axes are checklist dimensions on a 0–1 scale.

(FLUX.2 [klein] 9B 0.208, BAGEL 0.173). Across both families, the difficulty lies in preserving the structural and informational properties that make it usable as a design. Detailed dimension-level score can be seen in Appendix D.4.

4.3 Analysis and Discussion

Model generations substantially improve design capability. Across the evaluated model families, successive generations generally advance design capability (Table 1). Newer models generally outperform their predecessors on extraction, extension, rendering, and dissection, showing that ***architectural and training improvements translate into stronger design operations***. A few operations, such as certain generation or editing tasks, may exhibit marginal gains or remain comparable, and reference-grounded composition can stay stable under scale increases. These exceptions are limited and do not overturn the overall trend: model iterations systematically enhance design capability.

Text rendering remains a bottleneck in composition. Composition exposes a persistent text-quality gap, and the gap is not uniform across languages (Figure 5a). Models separate first by overall level: GPT Image 2 and GPT Image 1.5 occupy a high band on both Chinese and English (means 0.55/0.58 and 0.52/0.54), whereas FLUX.2 [klein] 9B, HiDream-I1, and BAGEL remain near the floor in both languages. A second split is bilingual symmetry. The GPT models show only a modest English edge, largest on posters. NanoBanana 2 and NanoBanana are English-leaning: their poster English scores exceed Chinese by about 0.19, and both FLUX.2 variants, though weak overall, are likewise higher on English than on Chinese. Seedream 5.0, Wan2.7-Image, and ERNIE-Image sit in a middle band with little net language bias. Advertisements are the hardest document type for almost every system, while UI copy is comparatively easier for the leading models. Thus a strong Composition score does not imply bilingual text fidelity: several otherwise capable systems remain English-leaning on the most copy-heavy layouts.

Composition degrades with reference count. Composition scores decline as the number of ordered references increases (Figure 5). With more references, the model must assign each input a distinct

design role and keep text, layout, and assets aligned to those roles, so role assignment and layout constraints come into conflict more often. A dimension-level breakdown in the Figure D.5 (Appendix) shows that this drop is not uniform: text and layout degrade faster than attribute compliance or fusion quality, and a composition can look fused while still corrupting copy or hierarchy. The overall trend is that reference-grounded composition becomes substantially harder as the reference set grows.

Each family is defined by a shared core criterion. Figure 6 groups subunits by a shared axis, and that axis—not pictorial finish—is what separates models. On Rendering, rendering quality is high for almost every system, but structural truth is not. GPT Image 2, GPT Image 1.5, and NanoBanana 2 keep layout, shape, and chart data accuracy together; FLUX.2 [klein] 9B remains strong on interior geometry and product form, yet its data accuracy collapses (0.141). On Extension, GPT Image 2 and GPT Image 1.5 hold reference identity across fonts, products, and UI (font recognizability and coverage near 0.96–0.98). NanoBanana 2 keeps product and UI continuity but is weaker on glyph coverage; FLUX.2 [klein] 9B and BAGEL drop to letter-like shapes and lose UI text and interaction. On Dissection, GPT Image 2 and GPT Image 1.5 stay high on identity, detail, and angle coverage for both characters and products. NanoBanana 2 leads product consistency (0.852) but is weaker on character angles; ERNIE-Image covers anime views well (0.883) and then loses product identity (0.442). A lead on one subunit's appearance therefore does not imply a lead on the family's core criterion, unless that lead holds across the shared axes.

Qualitative comparison. Beyond family scores, we compare model outputs on the same task to show how the core criteria in Figure 6 appear in actual deliverables. The Section D.6 in Appendix collects these side-by-side examples: each row holds one task instance and each column a model. Representative successes and remaining failure modes are both visible in that comparison.

4.4 Human–VLM Scoring Consistency

To validate the automatic evaluator, we conducted a blind human evaluation spanning both closed-source and open-source models. For each task unit, we selected a subset of generations from eight models, yielding 864 judging tasks, each with 10–16 criteria. Nine experts with design experience were recruited as annotators to score these anonymous outputs using the same task-specific 0–3 checklist as the GPT evaluator, without access to model identities, VLM scores, or VLM rationales. Human item scores were aggregated using the same criterion and dimension weights as those used for the VLM evaluator's scores, yielding human and VLM sample scores both in [0,1]. For each evaluation unit, we computed Spearman's rank correlation (SRCC) between the human rankings and each VLM judge. Figure 7 reports the model-level rank agreement under this protocol. GPT-5.4 achieved an overall SRCC of 0.976, demonstrating the effectiveness of VLM-based evaluators. More details are provided in the Appendix Section D.7.



Figure 7: Model-level Human–VLM (GPT5.4) agreement.

5 Conclusion

We introduced **DesignerBench**, a workflow-centric and reference-aware benchmark for evaluating image generation systems as professional design collaborators. Its current audited core contains 2,000 instances across creation, extraction, composition, editing, rendering, dissection, and extension. The benchmark pairs explicit reference routing with sample-specific four-level checklists and version-locked, evidence-first scoring. This formulation turns design evaluation from a single aesthetic judgment into a diagnosis of controllability, fidelity, information correctness, and technical usability. Across 14 scored systems, the best model depends on workflow: the leaders in generation and extraction, editing, and rendering are different. Moreover, unequal task support materially changes aggregate conclusions, as demonstrated by ranking reversals between conditional and coverage-adjusted scores. These results motivate reporting workflow-level performance and explicit coverage instead of a single appearance-oriented score.

AI USE STATEMENT

In this work, generative AI tools were used solely for language polishing of the manuscript. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final content.

REFERENCES

Adobe. Inaugural Adobe Creators' Toolkit Report. Industry report, Adobe, October 2025. URL https://news.adobe.com/news/2025/10/adobe-max-2025-creators-survey. Global survey of 16,000+ creators across 8 countries.

Black Forest Labs. Flux.2. https://bfl.ai/blog/flux-2, 2026.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402. IEEE, 2023.

ByteDance. Seedream. https://www.bytedance.com, 2026. Multimodal image generation and editing system.

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Designer Fund, Foundation Capital, et al. AI in Design 2026. Industry report, Designer Fund, May 2026. URL https://designerfund.com/blog/ai-in-design-2026. Survey of 900+ designers across 60+ countries.

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

Hsiao Yuan Hsu, Xiangteng He, Yuxin Peng, Hao Kong, and Qing Zhang. Posterlayout: A new benchmark and approach for content-aware visual-textual presentation layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6018–6026, 2023.

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

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 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, 2025.

Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. ImagenHub: Standardizing the evaluation of conditional image generation models. *arXiv preprint arXiv:2310.01596*, 2023.

Bryan Lawson. *How designers think: The design process demystified*. Routledge, 2006.

Yan Li, Zezi Zeng, Ziwei Zhou, Xin Gao, Muzhao Tian, Yifan Yang, Mingxi Cheng, Qi Dai, Yuqing Yang, Lili Qiu, et al. Bizgeneval: A systematic benchmark for commercial visual content generation. *arXiv preprint arXiv:2603.25732*, 2026.

Chen Liang, Lianghua Huang, Jingwu Fang, Huanzhang Dou, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Junge Zhang, Xin Zhao, and Yu Liu. Idea-bench: How far are generative models from professional designing? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18541–18551. IEEE, 2025.

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

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

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

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

OpenAI. Gpt-5.4 thinking system card. Technical report, OpenAI, March 2026b. URL https://deploymentsafety.openai.com/gpt-5-4-thinking/gpt-5-4-thinking.pdf.

Qwen Team and Alibaba Group. Qwen-image-2.0-pro. https://www.alibabacloud.com/help/zh/model-studio/qwen-image-2-0-pro, 2026. Multimodal image generation and editing system.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Donald A Schön. *The reflective practitioner: How professionals think in action*. Routledge, 2017.

Wan Team and Alibaba Group. Wan-image: Pushing the boundaries of generative visual intelligence. *arXiv preprint arXiv:2604.19858*, 2026. URL https://arxiv.org/abs/2604.19858.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, et al. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Kai Xu, Ellis Brown, Shrikar Madhu, Rob Fergus, He He, and Saining Xie. Paintbench: Deterministic evaluation of precise visual editing. *arXiv preprint arXiv:2606.00188*, 2026.

Yujia Yang, Yuanxiang Wang, Zhenyu Guan, Tiankun Yang, Chenxi Bao, Haopeng Jin, Jinwen Luo, Xinyu Zuo, Lisheng Duan, Haijin Liang, et al. Omni IIE Bench: Benchmarking the practical capabilities of image editing models. *arXiv preprint arXiv:2603.16944*, 2026.

Hongzhu Yi, Yujia Yang, Yuanxiang Wang, Tong Li, Zhenyu Guan, Tianyu Zong, Jiahuan Chen, Chenxi Bao, Tiankun Yang, Haopeng Jin, et al. Vde bench: Evaluating the capability of image editing models to modify visual documents. *arXiv preprint arXiv:2602.00122*, 2026.

Hui Zhang, Dexiang Hong, Maoke Yang, Yutao Cheng, Zhao Zhang, Weidong Chen, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. Creatidesign: A unified multi-conditional diffusion transformer for creative graphic design. In *International Conference on Learning Representations*, volume 2026, pp. 111201–111214, 2026.

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

Bing Zhao, Chenfei Wu, et al. Qwen-image-2.0 technical report, 2026. URL https://arxiv.org/abs/2605.10730.