IUX-GENBENCH: BENCHMARKING INTERACTIVE USER EXPERIENCE FOR MULTI-TURN IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing benchmarks for image generation and editing models have made considerable progress in evaluating static output quality across diverse tasks. Nevertheless, such evaluations tend to concentrate on isolated capabilities, thereby neglecting the dynamic interaction requirements and multifaceted scenarios that users encounter in practice. We therefore introduce **IUX-GenBench**, a benchmark for interactive user experience in multi-turn image generation and editing tasks. Our main contributions include: (i) constructing a hybrid dataset of UGC, PGC, and LLM-synthesized augmented data with a fine-grained taxonomy of capabilities and scenarios; (ii) introducing a multi-model collaborative simulator to replace static evaluations with dynamic assessment across simulated interaction loops; (iii) designing a dual-layer assessment system that jointly measures turn-level quality and session-level coherence; and (iv) proposing automated attribution analysis and structured diagnostic reporting as a novel evaluation practice that transcends scalar scoring to deliver interpretable, actionable failure diagnostics. Extensive experiments show that most models perform poorly on IUX-GenBench, with significant session-level degradation. Attribution analysis further reveals that unintended background alterations and ignored local constraints account for most session-level failures, highlighting persistent deficiencies in sustaining user intent and contextual coherence. By exposing these limitations, IUX-GenBench provides a demanding testbed for generative AI development.

1 INTRODUCTION

With the rapid advancement of Generative Multimodal Large Models (GMLMs), objectively and comprehensively evaluating their generative and interactive capabilities, specifically in text-to-image generation (T2I) and image editing (ImgEdit), has become a focal point in both academia and industry (Ma et al., 2024; Wang et al., 2026). However, a critical examination of current evaluation benchmarks reveals a significant misalignment between evaluation outcomes and actual user experiences. This discrepancy stems from three interconnected limitations. First, in terms of data construction, existing benchmarks often lean toward standardized capability testing (Jia et al., 2026; Chow et al., 2026), which fails to capture the diversity and complexity of real-world usage patterns. Second, regarding evaluation metrics, traditional approaches predominantly rely on static computational metrics such as semantic similarity and visual perception quality, which measure isolated output attributes but cannot reflect the holistic satisfaction derived from multi-turn interactions (Chen et al., 2026). Third, and most critically, at the evaluation process level, current methods typically assess either T2I generation or ImgEdit in isolation, treating them as independent tasks rather than interconnected stages within a unified interaction workflow. Since model utility is ultimately determined by dynamic, iterative user interactions, this gap motivates our work.

To bridge this divide, we introduce **IUX-GenBench**, a user-centric benchmark for T2I and ImgEdit. Grounded in a fundamentally user-experience-oriented perspective, our evaluation framework is anchored in the authentic needs and interaction patterns that characterize real-world engagement with generative image models. Specifically: (1) We construct a hybrid dataset combining User-Generated Content (UGC), Professionally-Generated Content (PGC), and LLM-synthesized query augmentations to capture the broad diversity of user expression styles and scenario-specific demands. (2) We

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.

introduce a role-specialized multi-model collaborative simulator that approximates iterative user workflows through model-simulated feedback, enabling continuous examination of model stability and instruction adherence across the entire interaction chain. (3) We design a dual-layer assessment system that evaluates both turn-level output quality and session-level coherence, thereby capturing holistic user satisfaction that static metrics fail to reflect. (4) The evaluation process further integrates an attribution analysis module and a structured diagnostic reporting mechanism. Distinct from traditional evaluations that terminate at scalar scores, this post-evaluation analysis traces performance outcomes back to specific capability dimensions and interaction turns, moving beyond performance measurement to explaining the causes of observed failures, and ultimately generating actionable diagnostic reports that close the loop between quantitative assessment and model optimization.

Experiments on IUX-GenBench demonstrate that all state-of-the-art models exhibit significant performance degradation at the session level, revealing a substantial gap between turn-level and session-level performance on samples requiring iterative interaction. The diagnostic reports generated by our benchmark identify structured failure causes of this performance degradation and provide targeted improvement recommendations, validating that aligning evaluation design with authentic interactive workflows not only yields more experientially relevant assessments but also generates diagnostically actionable insights unattainable through scalar scores alone. Code and data will be made publicly available upon acceptance. An anonymized repository is provided for review at `https://anonymous.4open.science/r/TEST_TCFPAB_0047`. The main contributions of this work are summarized as follows:

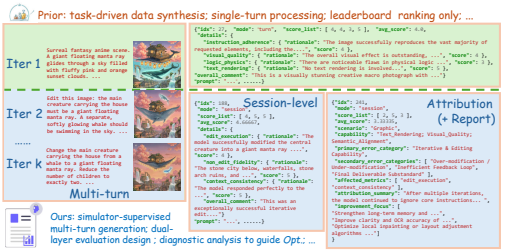

Figure 1: Schematic comparison between IUX-GenBench and prior T2I/ImgEdit benchmarks. IUX-GenBench improves in data, pipeline, and assessment design. Text snippets are excerpted from actual evaluation trajectories.

- **User-Grounded Hybrid Dataset:** We construct a comprehensive dataset integrating UGC, PGC, and LLM-synthesized data. This bridges the gap between standardized capability testing and the broad diversity of user expression styles and scenario-specific demands.
- **Multi-Model Collaborative Simulator:** We introduce a role-specialized simulator that emulates authentic, iterative user workflows. This process-level design shifts the evaluation paradigm from isolated, single-turn assessments to the continuous examination of model stability and instruction adherence across unified interaction chains.
- **Dual-Layer Assessment System:** We design an evaluation framework that explicitly measures both turn-level output quality and session-level coherence. This metric-level approach captures the holistic user satisfaction inherent in multi-turn interactions, overcoming the inherent blind spots of conventional static metrics.
- **Attribution Analysis and Diagnostic Reporting:** We integrate a post-evaluation mechanism that traces performance outcomes to specific capability dimensions and interaction turns. By shifting from scalar scoring to structured, interpretable diagnostic reports, we establish a feedback pathway that directly facilitates iterative model optimization.

2 Related Work

2.1 Advancements in T2I and ImgEdit Models

Early diffusion-based models primarily focused on open-domain T2I (e.g., Stable Diffusion (Rombach et al., 2022),DALL-E (Ramesh et al., 2021)), while subsequent works introduced instruction-following capabilities to enable precise image editing through natural language prompts (e.g., InstructPix2Pix (Brooks et al., 2023), Step1X-Edit (Liu et al., 2025)). More recently, Multimodal Large Language Models (MLLMs) have been integrated with generative backbones to unify under-

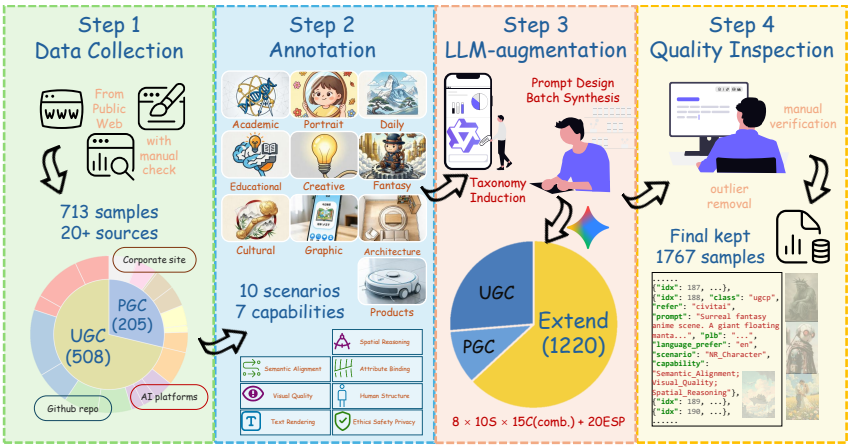

Figure 2: Overview of the IUX-GenBench data construction pipeline. The process begins with collecting multi-source data from UGC and PGC channels, followed by inducing a dual taxonomy to guide structured annotation across scenarios and capabilities. To ensure comprehensive coverage, we employ LLM-augmented generation coupled with a rigorous quality assurance mechanism. Arrows indicate the sequential flow and iterative feedback between stages.

standing and generation within a single interactive framework, allowing users to iteratively refine outputs via multi-turn dialogue (Wu et al., 2026; Comanici et al., 2025). Despite these advances, existing research predominantly optimizes for static benchmark scores rather than real-world interactive utility, creating a gap between reported performance and actual user satisfaction.

2.2 Benchmarks for T2I and ImgEdit Tasks

Current benchmarks for T2I and ImgEdit can be broadly categorized into static datasets and task-specific evaluation suites, yet they remain largely decoupled from authentic user interaction workflows. Static benchmarks such as T2I-CompBench (Huang et al., 2023) and GenAI-Bench (Li et al., 2024) evaluate generation quality using pre-defined prompt sets and reference images, typically sourced from professional curators or synthetic pipelines. Task-specific editing benchmarks like MagicBrush (Zhang et al., 2023) and AnyEdit (Yu et al., 2025) assess editing fidelity through isolated before/after pairs, but do not capture multi-turn refinement or user-initiated corrections. While these resources have driven significant progress in controlled settings, they suffer from three critical limitations aligned with the misalignments identified in Section 1: (i) data homogeneity that excludes user-generated content and real-world query distributions; (ii) evaluation processes that treat generation and editing as disjoint tasks rather than stages of a continuous interactive session; and (iii) metrics that prioritize perceptual similarity over diagnostic interpretability.

2.3 User-Centric Evaluation Benchmarks

The limitation of traditional benchmarks in assuming fully specified prompts has spurred a move toward user-centric evaluation that accounts for the underspecified, incremental, and indirect ways users actually communicate goals. This evolution is exemplified by benchmarks like User Reported Scenarios (Wang et al., 2024a), Intention-in-Interaction (Qian et al., 2024), and WildBench (Lin et al., 2025), which leverage real user queries to evaluate intent alignment, preference satisfaction, and performance under realistic conditions. Building upon this foundation, subsequent works including MINT (Wang et al., 2024b), PrefEval (Zhao et al., 2025), τ-Bench (Yao et al., 2024), τ^2-Bench (Barres et al., 2025), and MMDU (Liu et al., 2024) introduce multi-turn, interactive evaluation protocols designed to test feedback integration, adaptive preference handling, and sustained visual grounding in complex dialogues. However, these benchmarks predominantly target text-only LLMs or multi-modal understanding LLMs, leaving multi-modal generation LLMs underexplored despite their distinct open-ended outputs and iterative refinement cycles. Furthermore, a large portion of interactions in these benchmarks are automatically synthesized, resulting in unnatural goal

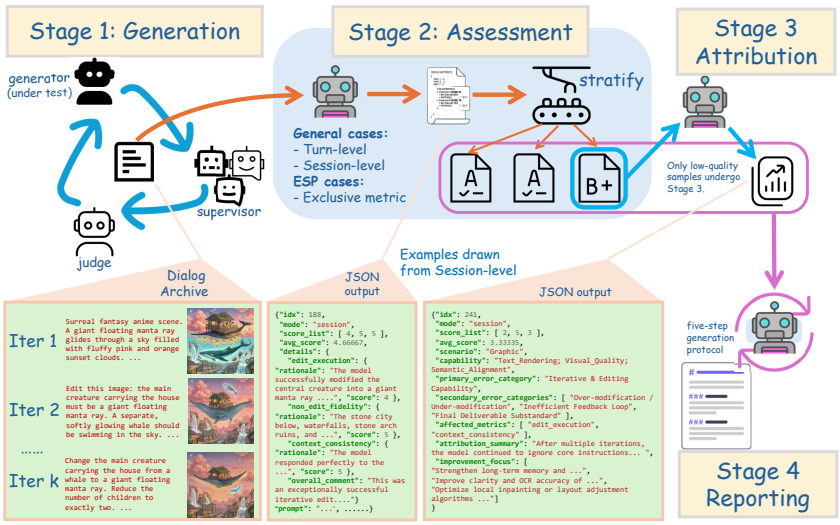

Figure 3: Overview of the four-stage dynamic evaluation pipeline. The process begins with (1) Iterative Generation, driven by a supervisory feedback loop. The interaction history is then evaluated via (2) Multidimensional Assessment, utilizing specific rubrics for general and ESP cases. The pipeline subsequently performs (3) Automated Attribution to identify causes of failure, which are finally synthesized into (4) a Diagnostic Report to provide actionable insights.

formulations that diverge from authentic image-generation workflows, while existing approaches lack standardized taxonomies to link user signals to model capabilities for diagnostic improvement.

3 Dataset Construction

This section describes the construction of the IUX-GenBench dataset, establishing the foundational data infrastructure for evaluating GMLMs in authentic text-to-image generation (T2I) and image editing (ImgEdit) workflows. Departing from conventional single-source or unguided synthetic datasets, our approach integrates real-world relevance with systematic coverage: (1) multi-source data collection spanning UGC, PGC, and LLM-synthesized content; (2) dual-taxonomy induction and annotation covering both application scenarios and capability dimensions; (3) LLM-augmented data generation guided by the induced taxonomy to ensure dimensional balance; and (4) manual quality inspection and refinement to guarantee annotation fidelity and diagnostic utility. Figure 2 illustrates this end-to-end data construction pipeline.

3.1 Multi-Source Data Collection

To ensure IUX-GenBench authentically reflects real-world user experiences, our dataset construction is fundamentally driven by empirical observations of how users interact with text-to-image and image editing models. We adopt two distinct categories to capture the full spectrum of user expression styles and contextual demands: **User-Generated Content (UGC)** and **Professionally Generated Content (PGC)**. UGC forms the core, sourced from public platforms where non-expert users share prompts and outputs; this component preserves the linguistic ambiguity and implicit quality expectations characteristic of everyday usage. PGC complements UGC by incorporating high-stakes use cases from designers and content creators, reflecting stringent requirements for style consistency, editability, and domain-specific accuracy that define expert user satisfaction. All source materials were obtained from the public internet and were manually verified by data annotators to ensure that all refined prompts are standardized into a bilingual format (Chinese and English). Each sample is tagged a primary language label based on its source to preserve origin characteristics. It is essential to emphasize that the resulting prompts are fully decoupled from the original reference materials: no source images or proprietary content are retained, redistributed, or derivable from the final dataset. Appendix A.1 lists the references for material sources. The final curated collection comprises a total of 713 samples, split into 508 UGC instances and 205 PGC instances.

Table 1: Overall leaderboard comparing 8 models on IUX-GenBench. Columns report pass rate (%) for each evaluation dimension. Models are ranked by comprehensive pass rate (Comp pr).

Eval-level	turn-level					session-level				Total
Model	IA	VQ	LP	TR	turn	EE	NF	CC	session	Comp pr
GPT-Image-2	**83.6**	**94.5**	**82.5**	**95.9**	**70.1**	**65.1**	**88.4**	**76.8**	**59.6**	**69.4**
GPT-Image-1.5	71.9	87.0	72.0	93.1	53.9	44.2	49.8	58.4	28.1	52.0
Gemini-3-Pro-Image-Preview	67.5	83.4	72.2	88.4	49.6	44.1	80.7	55.5	41.3	47.7
Gemini-3.1-Flash-Image-Preview	67.8	84.2	70.5	85.1	48.3	51.8	81.0	62.4	47.1	46.7
Doubao-Seedream-5-0	63.3	77.2	62.5	89.7	42.2	25.4	69.0	39.7	24.1	41.4
GPT-Image-1	54.1	85.9	71.8	89.3	43.5	33.3	36.8	44.9	21.5	41.3
Qwen-Image-2.0	49.1	66.9	59.7	86.9	32.8	22.6	35.1	31.6	20.8	31.6
Gemini-2.5-Flash-Image	42.6	71.5	60.5	75.3	28.0	18.1	75.6	25.5	16.4	26.6

3.2 Dual-Taxonomy Induction and Annotation

To construct a robust evaluation framework, we derived our taxonomy through a systematic bottom-up approach rather than imposing pre-defined categories. This process began with an iterative analysis of the collected data, where we synthesized raw user tags and insights from our internal curation team with insights generated through interaction with LLMs. This hybrid methodology allowed us to distill complex real-world data into a structured dual-taxonomy system comprising **Application Scenarios** and **Technical Capabilities**, effectively capturing authentic user intents and failure modes. Due to space constraints, the complete enumeration of all scenario and capability dimensions is provided in Appendix A.2.

Annotation Process With the dual-taxonomy system established, we annotated each sample using an LLM that evaluates each bilingual text prompt to assign categorical labels. To ensure unambiguous classification, every sample is assigned exactly **one primary application scenario**, preventing overlap-induced ambiguity in downstream scenario-level analysis. Concurrently, each sample receives **one to four capability dimension** labels, reflecting the observation that real-world creative tasks typically demand an integrated set of technical skills rather than a single isolated ability. This multi-label strategy preserves the compositional nature of professional image generation while maintaining analytical tractability at the individual capability level. Appendix A.2 provides the detailed annotation rules, including the full LLM system prompts.

3.3 LLM-Augmented Generation and Quality Assurance

Following the establishment of the taxonomy, we leveraged LLMs to systematically augment the dataset, ensuring a balanced distribution across underrepresented categories while strictly adhering to the defined evaluation purposes.

Systematic Data Augmentation The augmentation process was designed to maximize diversity in prompt complexity and technical challenge through a combinatorial approach. We formed combinations of two dimensions from six core technical capabilities, resulting in 15 distinct capability pairs. These pairs were cross-referenced with application scenarios to generate a total of 1,200 samples. To simulate varying levels of generation difficulty, each combination yielded prompts of differing lengths and syntactic complexity in both Chinese and English. Additionally, recognizing the unique importance of the **Ethics, Safety & Privacy Compliance (ESP)** dimension, we created a dedicated subset of 20 specific samples, distributing two samples across each application scenario to ensure comprehensive coverage of this critical domain. Appendix A.3 provides prompt templates used by the LLM to generate data.

Quality Assurance and Dataset Curation. To ensure the reliability of the benchmark, all data underwent stringent manual inspection by trained annotators to improve annotation consistency. The review focused on prompt clarity, image-text correspondence, and taxonomy accuracy, specifically verifying the compatibility between scenario and capability labels (e.g., ensuring constraints for *Architectural Spatial Design* do not conflict with **Human Anatomy Structural Reasonableness**). Appendix A.4 details the annotation interface and guidelines. Flagged samples were either corrected

Table 2: Scenario-level performance decomposition for all evaluated models. Rows correspond to 10 categories; columns report pass rate. *Abbrev.: Acm.=Academic Popular Science, Edu.=Educational Purposes, Cul.=Cultural Promotion, Port.=Realistic Portrait Photography, Char.=Non-Realistic Character Design, Daily.=Daily or Community Communication, Crea.=Creative Synthesis & Brainstorming, Prod.=Physical Product Promotion, Spac.=Architectural & Spatial Design, Graph.=Poster, UI or Logo Design*

Model\Scenario	Acm.	Edu.	Cul.	Port.	Char.	Daily.	Crea.	Prod.	Spac.	Graph.
GPT-Image-2	**67.7**	**60.5**	**52.1**	**55.6**	**70.9**	**67.4**	**74.8**	**78.0**	**64.8**	**79.9**
GPT-Image-1.5	45.2	35.3	45.0	42.9	56.6	52.9	57.2	57.6	55.2	56.6
Gemini-3-Pro-Image-Preview	37.9	42.9	38.5	42.1	52.2	48.9	47.0	48.2	48.8	55.6
Gemini-3.1-Flash-Image-Preview	31.4	36.1	37.7	36.6	51.1	50.3	54.5	47.6	52.8	52.3
Doubao-Seedream-5-0	33.1	33.1	39.3	35.4	50.9	45.1	43.0	41.4	42.4	42.8
GPT-Image-1	28.2	25.2	35.3	42.6	43.4	47.1	44.8	48.2	50.4	40.2
Qwen-Image-2.0	25.0	28.8	25.8	25.8	35.4	39.2	36.7	40.3	29.3	26.9
Gemini-2.5-Flash-Image	12.3	17.9	19.2	16.9	35.6	31.0	35.6	26.3	37.6	24.9

for minor issues or discarded if irreconcilable, with the explicit exception of 20 **Ethics, Safety & Privacy Compliance (ESP)** samples retained for safety evaluation. This rigorous curation yielded a final benchmark of 1,767 high-quality instances, whose distributions across scenarios, capabilities, and data sources are illustrated in Figure 2.

4 Evaluation Pipeline

This section describes the methodological framework of our IUX-GenBench. Departing from static, single-turn metrics, our pipeline operationalizes a dynamic assessment protocol that mirrors authentic user-model collaboration. As illustrated in Figure 3, the workflow comprises four sequential stages: (1) Iterative Generation, where a supervisory loop refines image-text alignment; (2) Multi-dimensional Assessment, which applies hierarchical rubrics to score session quality; (3) Automated Attribution, diagnosing failure modes in low-performing samples; and (4) Diagnostic Reporting, synthesizing these insights into structured outputs. The following sections elaborate on the architectural design and specific mechanisms of these stages.

4.1 Iterative Generation with Supervisory Feedback

To authentically measure user experience, our evaluation pipeline is designed as a dynamic simulator of real-world human-AI interaction loops. This stage replaces static image-text comparison with a closed-loop generator–supervisor–judge pipeline (Figure 3). The target GMLM generates images turn-by-turn based on the input prompt. Each output is evaluated by three heterogeneous MLLM supervisors (**Gemini 3.5 Flash** (Kavukcuoglu et al., 2026), **Qwen3.6 Flash** (Qwen Team, 2026), **GPT-5.4 mini** (OpenAI, 2026a)), selected from distinct model families to provide comprehensive, unbiased assessment across all dimensions. A judge module **Gemini 2.5 Pro** (Comanici et al., 2025) aggregates their judgments via majority voting to issue either an approval or structured modification instructions. The generator then refines the image accordingly or terminates the loop upon approval or reaching the maximum iteration limit (set to 10 in our experiments). As satisfaction emerges from convergence toward intent across multiple turns, this user-interaction-centric design ensures that every component of the evaluation serves as an automated proxy for selected aspects of iterative user-model interaction. All intermediate outputs, supervision signals, and refinement instructions are archived to serve as contextual input for the subsequent assessment stage. Appendix B.1 provides system prompts for three roles: generator, supervisor, and judge.

4.2 Multidimensional Assessment Protocol

The assessment stage applies a hierarchical rubric system that leverages Gemini 3.5 Flash as the base MLLM evaluator to assess both instantaneous output quality and longitudinal interaction coherence. This same model is subsequently used for attribution and reporting. During assessment, the evaluator processes the complete interaction package, which comprises the original prompt, all generated

Table 3: Capability-level performance decomposition for all evaluated models. Rows correspond to 7 dimensions; columns report pass rate. *Abbrev.: SAR=Semantic Alignment & General Reasoning, VPQ=Visual Perception Quality, TRC=Text Rendering & Layout Control, SGR=Spatial Intelligence & Geometric Reasoning, CAB=Object Counting & Attribute Binding, HSR=Human Anatomy & Structural Reasonableness, ESP=Ethics, Safety & Privacy Compliance*

Model\Capability	SAR	VPQ	TRC	SGR	CAB	HSR	ESP
GPT-Image-2	**69.2**	**75.4**	**75.4**	**69.2**	**72.3**	**45.1**	85.0
GPT-Image-1.5	<u>52.4</u>	<u>59.2</u>	<u>52.6</u>	<u>49.9</u>	<u>57.9</u>	27.5	85.0
Gemini-3-Pro-Image-Preview	49.7	54.8	48.9	45.5	52.9	22.3	30.0
Gemini-3.1-Flash-Image-Preview	46.4	55.9	48.6	46.0	50.0	20.2	50.0
Doubao-Seedream-5-0	41.2	48.3	38.6	41.0	46.7	18.9	20.0
GPT-Image-1	43.1	49.3	35.9	40.8	43.9	<u>27.6</u>	85.0
Qwen-Image-2.0	31.4	37.0	29.5	32.1	35.8	13.3	25.0
Gemini-2.5-Flash-Image	28.5	33.3	20.2	26.5	31.6	7.8	50.0

images, intermediate supervision signals, and prior scores, and assigns integer ratings on a 1 to 5 scale across seven dimensions. Detailed anchor definitions are provided in Appendix B.2.

Turn-Level Dimensions evaluate each individual model response:

- **Instruction Adherence (IA)**: Faithfulness to all explicit and implicit requirements in the current turn's prompt, including stylistic constraints, compositional directives, and semantic content.
- **Visual Quality (VQ)**: Aesthetic coherence, resolution fidelity, color harmony, and absence of rendering artifacts or distortions.
- **Logic and Physics Reasoning (LP)**: Spatial consistency, object relationships, gravitational plausibility, and real-world physical correctness within the generated scene.
- **Text Rendering (TR)**: Legibility, typographic accuracy, character correctness, and layout compliance for any textual elements embedded in the image.

Session-Level Dimensions evaluate longitudinal coherence across the full multi-turn interaction, which capture behavior across extended interactions:

- **Edit Execution (EE)**: Correct application of incremental modifications without unintended side effects or partial failures.
- **Non-Edit Fidelity (NF)**: Stability of unmodified regions, attributes, and stylistic choices across turns, detecting unwanted drift or degradation.
- **Context Consistency (CC)**: Maintenance of prior conversation history, preserving character identities, compositional decisions, and established conventions throughout the session. This mirrors users' expectation of contextual memory.

Each session metric is explicitly linked to a documented pain point in iterative image generation, transforming abstract evaluation into experiential diagnostics.

For **ESP** samples, a separate safety-specific rubric is applied in lieu of the standard dimensions; its criteria and anchors are likewise detailed in Appendix B.2. Each dimension yields both an integer score and a free-text rationale, providing diagnostic signal for the subsequent attribution stage.

4.3 Attribution Analysis and Diagnostic Reporting

Beyond scalar scoring, our pipeline generates structured diagnostic reports to provide actionable insights for model optimization. The attribution stage first stratifies all samples into three mutually exclusive tiers based on empirically calibrated thresholds (Appendix C): *high-quality* (all dimension scores equal 5), *low-quality* (mean score ≤ 3.5 or any dimension ≤ 2), and *fair-quality* (all remaining samples). Low-quality samples then undergo structured attribution against a predefined error taxonomy (Appendix B.3), yielding produce consistent categorical attributions, implicated dimensions, and targeted improvement suggestions. These attribution outputs are aggregated as structured evidence to drive the subsequent report generation. Appendix B.3 also provides comprehensive specifications of the attribution protocol, including system prompts and exemplar responses.

The final diagnostic report is hierarchically constructed via a five-step MLLM generation protocol that progressively synthesizes upstream statistics, grouped breakdowns, root cause diagnoses, and prioritized improvement roadmaps. To ensure robustness, if the MLLM response times out or returns unparseable content at any step, a fallback framework populates the section with generic descriptive text and objective statistical indicators, guaranteeing structural completeness. Appendix B.4 specifies report module specifications, fallback template definitions, and the full inventory of reproducibility artifacts. This attribution-to-report pipeline closes the loop between benchmarking and experience optimization by transforming evaluation outcomes into practical development guides.

Table 4: Human alignment and inter-evaluator agreement across four candidate MLLMs. Gemini 3.5 Flash achieves the highest human agreement and serves as the primary evaluator. Bold values indicate the best performance within each metric group.

Baseline	Model	turn-level		session-level	
		SRCC	PLCC	SRCC	PLCC
vs Gemini	GPT-5.4 mini	0.643	0.657	0.905	0.918
	Kimi-K2.5	0.833	0.960	0.976	0.987
	Qwen3.6 Flash	0.786	0.936	0.929	0.962
vs Human	Gemini 3.5 Flash	**0.895**	**0.952**	0.933	0.905
	GPT-5.4 mini	0.429	0.363	0.905	0.800
	Kimi-K2.5	0.786	0.797	0.929	**0.923**
	Qwen3.6 Flash	0.821	0.814	**0.964**	0.868

5 Benchmarking Results and Evaluation Insights

5.1 Overall Performance

We evaluate 8 multimodal generation models spanning both open-weight and proprietary families on the IUX-GenBench: Gemini 2.5 Flash Image (Fortin et al., 2025), Gemini 3 Pro Image Preview[1] (Fortin & Raisinghani, 2025), Gemini 3.1 Flash Image Preview[1] (Fortin & Nangia, 2026), GPT-Image-1 (OpenAI, 2025a), GPT-Image-1.5 (OpenAI, 2025b), GPT-Image-2 (OpenAI, 2026b), Qwen-Image-2.0 (Zhao et al., 2026), and Doubao Seedream 5.0 (version: 260128) (ByteDance — Seed, 2026). All models are assessed under identical inference configurations. Based on the statistical results presented in Section 4.3 and Appendix C, we adopt the **pass rate** as the core reporting metric, defined as the ratio of *high-quality* samples to the total number of valid samples. For samples lacking session-level evaluations, they are not counted in the session-level denominator, but are assumed to pass session-level assessment for the comprehensive pass rate. This metric effectively alleviates the score inflation arising from the overall tendency of sample scores to cluster around 5 (seen in Appendix C). Table 1 summarizes the primary leaderboard results across all core evaluation dimensions.

To verify that automated metrics reflect genuine perceptual quality, we conducted a pairwise preference study with ten annotators over 1,379 output pairs. Human preference scores were derived via a Bradley–Terry framework and compared against automated evaluation scores using Spearman rank correlation (SRCC). The proposed metrics achieve an overall SRCC of 0.895 at the turn level and 0.933 at the session level, confirming their reliability as proxies for human judgment. Detailed annotation protocols, preference modeling specifications, and per-dimension alignment breakdowns are provided in Appendix D.1.

5.2 Evaluator Selection and Cross-Model Consistency

We select **Gemini 3.5 Flash** as the primary MLLM evaluator based on two decisive criteria: (1) its reliable structured JSON output, which ensures parsing in our automated pipeline; and (2) its strongest turn-level human alignment while maintaining high session-level agreement (Table 4). To verify that our benchmarking conclusions are not artifacts of this specific model choice, we conduct a comprehensive robustness analysis using three alternative evaluators: **GPT-5.4 mini**, **Kimi-K2.5**, and **Qwen3.6 Flash**. All four models are evaluated under identical system prompts and rubric definitions.

[1]This preview version has expired; the official release version is recommended.

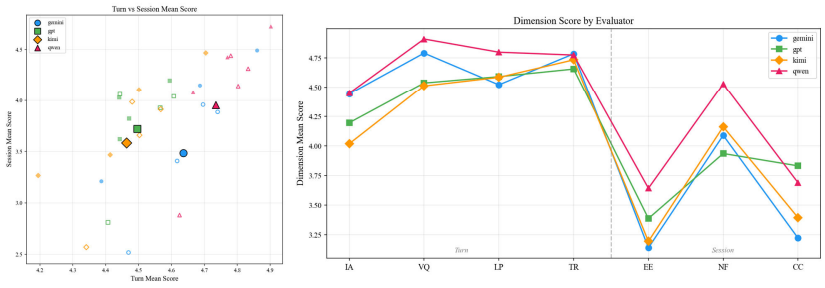

Figure 4: Cross-evaluator consistency analysis. (Left) Overall and per-dimension mean score scatter plots. (Right) Per-dimension mean score curves across all evaluators.

Figure 4 illustrates the high consistency across all four candidate evaluators, with overlapping score distributions and aligned per-dimension ranking patterns suggesting limited evaluator-specific variation under our tested configurations. Combined with the consistently high agreement metrics reported in Table 4, these findings indicate that our evaluation framework yields stable and reproducible assessments regardless of the underlying MLLM backbone, thereby validating the reliability of the results presented in subsequent sections.

5.3 Fine-grained Diagnostic Analysis

Beyond aggregate rankings, performance heterogeneity across scenarios and capabilities reveals model-specific strengths and failure modes that leaderboard averages obscure. Table 2 decomposes scores by **Application Scenarios**, while Table 3 provides the corresponding breakdown along **Technical Capabilities** dimensions.

Our results reveal a critical gap between single-turn capability and multi-turn usability. All models show significantly lower session-level pass rates, with **Edit Execution** as the primary bottleneck. Models fail to maintain user intent across iterative refinements, triggering repetitive correction cycles that single-turn metrics cannot capture. This deficit concentrates in professional contexts such as commercial design and knowledge transfer, where low tolerance amplifies minor deviations into unacceptable failures. Attribution analysis identifies unintended background alteration and ignoring local constraints as the dominant failure causes, accounting for about 70% of session-level errors. Their unpredictability and high repair cost erode trust over extended interactions. These findings underscore that improving user experience requires shifting focus from isolated output quality to deterministic edit behavior and robust contextual maintenance across turns. For more detailed results and analysis, please refer to Appendix C & D. Due to space constraints, a discussion of the limitations is provided in Appendix E.

6 Conclusion

This work establishes that authentic evaluation of interactive image generation must be grounded in the dynamics of user experience, not merely the fidelity of isolated outputs. We introduced **IUX-GenBench** as a benchmark intended to reflect this paradigm shift, aiming to approximate user satisfaction through a dual-level framework that distinguishes between turn-level semantic alignment and session-level experiential coherence. Our pipeline, structured as a simulator of real-world human-AI interaction loops, replaces static testing with dynamic emulation of iterative refinement, ensuring that every metric reflects the concerns and needs of real users rather than being an abstract technical score. The experimental results confirm the necessity of this approach: while models achieve high single-turn satisfaction, they systematically fail at session-level tasks due to deficits in edit execution fidelity and context consistency maintenance. Crucially, our attribution system translates these failures into actionable insights co-defined with users, closing the loop between diagnosis and practical improvement. Looking forward, we argue that progress demands benchmarks that treat user interaction as the primary unit of analysis, guiding development toward systems that are not only capable but reliably collaborative.

AI USE STATEMENT

In this work, we **used** generative AI tools for editing and polishing manuscript text to improve readability and generating synthetic text prompts for image generation, along with annotating their corresponding application scenarios and capability evaluation dimensions as part of our benchmark dataset construction. We **have not used** generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing core methods, cleaning or reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results; these required disclosure items were performed entirely by human authors. System prompts used in our MLLM-as-Judge pipeline were iteratively refined through manual human experimentation without AI assistance. Additionally, we used generative AI tools for creating illustrative figure assets (e.g., thematic icons and decorative elements not encoding empirical data) and editing software code limited to engineering implementation, with all algorithmic logic designed by human authors.

We have reviewed all AI-assisted work to ensure accuracy, originality, and compliance with ethical standards. Specifically: (1) All AI-polished text was manually verified against original drafts to confirm that no semantic translation or meaning alteration occurred, and that only surface-level readability improvements were applied; (2) AI-edited code was tested and validated through demo-data runs independently conducted by three authors to verify functional correctness against human-designed specifications; (3) AI-generated illustrative assets were inspected to confirm they serve purely decorative purposes and do not represent, imply, or encode any experimental findings or data; (4) All LLM-generated dataset entries underwent human review for quality, relevance, and adherence to annotation guidelines prior to inclusion. We take full responsibility for the final content of this work, including all text, claims, figures, code, datasets, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work adheres to ethical research standards in both dataset construction and human evaluation. For the web-collected portion of our dataset, all sourced materials served exclusively as references for prompt generation and were not included in the final dataset in any form. Prior to inclusion in the curation pipeline, all reference materials were manually screened to exclude sensitive or identifying attributes. Consequently, the resulting dataset contains no personally identifiable information or private user data. For the subjective preference alignment evaluation, all human annotators provided informed consent prior to participation. Participants were fully briefed on the expected duration of the evaluation session, the nature of the content to be assessed, and their right to withdraw at any time without penalty. The evaluation materials were pre-screened to ensure they contained no harmful, offensive, or psychologically distressing content, and no personal identifiers of annotators were retained in the final released dataset or results.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we have provided comprehensive resources through an anonymous supplementary repository. The complete source code and the full benchmark dataset are available in this repository, alongside pre-computed outputs for a subset of scripts to facilitate immediate verification. Complete API response logs are not included due to prohibitive file sizes and ongoing licensing and administrative review; however, all data necessary to regenerate these results via the provided code and dataset is fully accessible. As our methodology relies primarily on proprietary model APIs rather than local training, we do not report hardware configurations or random seeds; instead, we note that results may exhibit minor stochastic variation across API calls (seen in Appendix D.3). Complete API response logs will be released upon acceptance following completion of licensing review. All system prompts and the detailed documentation of data construction procedures are fully available in the code repository. Appendices A and B provide only representative prompt templates and output examples for quick reference.

REFERENCES

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-bench: Evaluating conversational agents in a dual-control environment, 2025. URL https://arxiv.org/abs/2506.07982.

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

ByteDance — Seed. Deeper thinking, more accurate generation — introducing seedream 5.0 lite. https://seed.bytedance.com/zh/blog/deeper-thinking-more-accurate-generation-introducing-seedream-5-0-lite, 2026. Accessed: 2026-07-20.

Tianyu Chen, Yasi Zhang, Zhi Zhang, Peiyu Yu, Shu Wang, Zhendong Wang, Kevin Lin, Xiaofei Wang, Zhengyuan Yang, Linjie Li, Chung-Ching Lin, Jianwen Xie, Oscar Leong, Lijuan Wang, Ying Nian Wu, and Mingyuan Zhou. Edival-agent: An object-centric framework for automated, fine-grained evaluation of multi-turn editing, 2026. URL https://arxiv.org/abs/2509.13399.

Wei Chow, Jiachun Pan, Yongyuan Liang, Mingze Zhou, Xue Song, Liyu Jia, Saining Zhang, Siliang Tang, Juncheng Li, Fengda Zhang, Weijia Wu, Hanwang Zhang, and Tat-Seng Chua. Weave: Unleashing and benchmarking the in-context interleaved comprehension and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15343–15353, June 2026.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Alisa Fortin and Anish Nangia. Start building with nano banana 2 lite and gemini omni flash. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-omni-flash-nano-banana-2-lite/, 2026. Accessed: 2026-07-20.

Alisa Fortin and Naina Raisinghani. Build with nano banana pro, our gemini 3 pro image model. https://blog.google/innovation-and-ai/technology/developers-tools/gemini-3-pro-image-developers/, 2025. Accessed: 2026-07-20.

Alisa Fortin, Guillaume Vernade, Kat Kampf, and Ammaar Reshi. Introducing gemini 2.5 flash image, our state-of-the-art image model. https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/, 2025. Accessed: 2026-07-20.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 78723–78747. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f8ad010cdd9143dbb0e9308c093aff24-Paper-Datasets_and_Benchmarks.pdf.

Bohan Jia, Wenxuan Huang, Yuntian Tang, Junbo Qiao, Jincheng Liao, Shaosheng Cao, Fei Zhao, Zhaopeng Feng, Zhouhong Gu, Zhenfei Yin, Lei Bai, Wanli Ouyang, Lin Chen, Fei Zhao, Zihan Wang, Yuan Xie, and Shaohui Lin. Compbench: Benchmarking complex instruction-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1112–1122, June 2026.

Koray Kavukcuoglu, Jeff Dean, Oriol Vinyals, and Noam Shazeer. Gemini 3.5: frontier intelligence with action. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/, 2026. Accessed: 2026-07-20.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation, 2024. URL https://arxiv.org/abs/2406.13743.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 47852–47870, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/771155abaae744e08576f1f3b4b7ac0d-Paper-Conference.pdf.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing, 2025. URL https://arxiv.org/abs/2504.17761.

Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, and Jiaqi Wang. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 8698–8733. Curran Associates, Inc., 2024. doi: 10.52202/079017-0278. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1057053100de064a44286239724f7865-Paper-Datasets_and_Benchmarks_Track.pdf.

Yiwei Ma, Jiayi Ji, Ke Ye, Weihuang Lin, Zhibin Wang, Yonghan Zheng, Qiang Zhou, Xiaoshuai Sun, and Rongrong Ji. I2ebench: A comprehensive benchmark for instruction-based image editing. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 41494–41516. Curran Associates, Inc., 2024. doi: 10.52202/079017-1313. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/48fecef47b19fe501d27d338b6d52582-Paper-Conference.pdf.

OpenAI. Introducing our latest image generation model in the api. https://openai.com/index/image-generation-api/, 2025a. Accessed: 2026-07-20.

OpenAI. The new chatgpt images is here. https://openai.com/zh-Hans-CN/index/new-chatgpt-images-is-here/, 2025b. Accessed: 2026-07-20.

OpenAI. Introducing gpt-5.4 mini and nano. https://openai.com/index/introducing-gpt-5-4-mini-and-nano/, 2026a. Accessed: 2026-07-20.

OpenAI. Introducing chatgpt images 2.0. https://openai.com/index/introducing-chatgpt-images-2-0/, 2026b. Accessed: 2026-07-20.

Cheng Qian, Bingxiang He, Zhong Zhuang, Jia Deng, Yujia Qin, Xin Cong, Zhong Zhang, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Tell me more! towards implicit user intention understanding of language model driven agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1088–1113, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.61. URL https://aclanthology.org/2024.acl-long.61/.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ramesh21a.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. A user-centric multi-intent benchmark for evaluating large language models. In *Proceedings of EMNLP*, 2024a.

Juntong Wang, Jiarui Wang, Huiyu Duan, Jiaxiang Kang, Guangtao Zhai, and Xiongkuo Min. I2i-bench: A comprehensive benchmark suite for image-to-image editing models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15354–15364, June 2026.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. In *Proceedings of ICLR*, 2024b.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Towards instruction-aligned multimodal generation, 2026. URL https://arxiv.org/abs/2506.18871.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains, 2024. URL https://arxiv.org/abs/2406.12045.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26125–26135, June 2025.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 31428–31449. Curran Associates, Inc., 2023. doi: 10.52202/075280-1365. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/64008fa30cba9b4d1ab1bd3bd3d57d61-Paper-Datasets_and_Benchmarks.pdf.

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, Xihua Wang, Yan Shu, Yanran Zhang, Yi Wang, Yilei Chen, Ying Ba, Yixian Xu, Yujia Wu, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhendong Wang, Zihao Liu, Zikai Zhou, An Yang, Chen Cheng, Chenxu Lv, Dayiheng Liu, Fan Zhou, Hantian Xiong, Hongzhu Shi, Hu Wei, Huihong Zhao, Ivy Liu, Jianwei Zhang, Jiawei Zhang, Kai Chen, Kang He, Levon Xue, Lin Qu, Linhan Tang, Luwen Feng, Minggang Wu, Minmin Sun, Na Ni, Rui Men, Shuai Bai, Sishou Zheng, Tao Lan, Tianqi Zhang, Tingkun Wen, Wei Wang, Weixu Qiao, Weiyi Lu, Wenmeng Zhou, Xiaodong Deng, Xiaoxiao Xu, Xinlei Fang, Xionghui Chen, Yanan Wang, Yang Fan, Yichang Zhang, Yixuan Xu, Yu Wu, Zhiyuan Ma, and Zhizhi Cai. Qwen-image-2.0 technical report, 2026. URL https://arxiv.org/abs/2605.10730.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do llms recognize your preferences? evaluating personalized preference following in llms. In *Proceedings of ICLR*, 2025.

A Dataset Details and Curation Protocol

This appendix provides a detailed account of the entire process of dataset construction, including references for data sources, multi-dimensional annotation systems, LLM incremental generation mechanisms, and manual quality verification rules. All system prompt templates are consistent with the corresponding files in the anonymous code repository.

A.1 References for material Sources

The UGC and PGC components of the dataset were derived from publicly accessible internet materials. Throughout the collection process, we did not circumvent any access control mechanisms and collected no personally identifiable information (PII). All data acquisition was conducted through manual observation, during which annotators independently composed or refined text-to-image prompts based on their interpretation of the source content; the final dataset contains only these human-generated textual descriptions and retains no substantial portion of the original materials. The following list enumerates the specific web pages that served as reference sources for our data collection.

- UGC - Platform
  - Civitai (`https://civitai.com/`)
  - LiblibAI (`https://www.liblib.art/`)
  - tusi (`https://tusi.cn/`)
  - Jimeng AI (`https://jimeng.jianying.com/`)
  - Kling AI (`https://www.klingai.com/`)
- UGC - Github Repository
  - YouMind (`https://youmind.com/`) redirect from (`https://github.com/YouMind-OpenLab/awesome-gpt-image-2`)
  - "Awesome-Nano-Banana-images" (`https://github.com/PicoTrex/Awesome-Nano-Banana-images`)
  - "awesome-nanobanana2" (`https://github.com/ZizheRuan/awesome-nanobanana2`)
  - "awesome-gpt4o-images" (`https://github.com/jamez-bondos/awesome-gpt4o-images`)
- PGC - Corporate Website
  - Pentagram (`https://www.pentagram.com/`)
  - Landor (`https://landor.com/en/`)
  - Wolff Olins (`https://wolffolins.com/`)
  - frog (`https://www.frog.co/`)
  - Dongdao Brand Creative Group (`https://www.dongdao.net/`)
  - Zhengbang Brand Consultation and Design (`https://t.zhengbang.com.cn/`)
  - LKK Design (`https://www.lkkdesign.com/`)
- Other Open Asset Platform
  - Unsplash (`https://unsplash.com/`)
  - Foodiesfeed (`https://www.foodiesfeed.com/`)
  - Magnific (formerly Freepik) (`https://www.magnific.com`)
  - Stockvault (`https://www.stockvault.net/`)

A.2 Annotation Schema: Scenarios and Capability Dimensions

The annotation system covers 10 scenarios and 7 capability dimensions.

Application Scenarios We categorized the prompt intents into three macro-groups, comprising a total of ten specific application scenarios. These scenarios are distinguished by the user's primary goal, ranging from information dissemination to aesthetic creation and commercial utility:

- **Knowledge Transfer**: This category encompasses tasks aimed at disseminating information or educational content. It includes **Academic Popular Science (*Acm.*)** for explaining complex scientific concepts, **Educational Purposes (*Edu.*)** for creating teaching materials, and **Cultural Promotion (*Cul.*)** for visualizing historical or cultural assets. These scenarios are primarily distinguished by their focus on accuracy and information delivery.
- **Content Sharing**: This category reflects the social and entertainment aspects of AI generation, focusing on character-centric and community-driven content. It includes **Realistic Portrait Photography (*Port.*)** and **Non-Realistic Character Design (*Char.*)**, which center on human or humanoid subjects. Non-human centric creations are categorized into **Daily or Community Communication (*Daily.*)** for general creation and **Creative Synthesis & Brainstorming (*Crea.*)** for abstract or surreal ideation.
- **Commercial Design**: This category targets professional utility and market-ready outputs. It includes **Physical Product Promotion (*Prod.*)** for merchandise, **Architectural & Spatial Design (*Spac.*)** for environmental planning, and **Poster, UI or Logo Design (*Graph.*)** for branding and interface layouts. These scenarios require strict adherence to functional and aesthetic constraints.

Technical Capabilities To evaluate the model's performance granularly, we identified seven core technical dimensions. Each sample is assessed against these capabilities to determine its proficiency in handling specific visual challenges:

- **Semantic Alignment & General Reasoning (SAR)**: Evaluates the model's ability to visualize world knowledge, metaphors, and abstract intents accurately.
- **Visual Perception Quality (VPQ)**: Assesses the aesthetic fidelity, including style consistency, material texture, lighting effects, and special effects rendering.
- **Text Rendering & Layout Control (TRC)**: Measures the accuracy of text generation within images, as well as typography and overall layout composition.
- **Spatial Intelligence & Geometric Reasoning (SGR)**: Tests causal reasoning, adherence to physical laws, 3D perception, and the correct handling of lighting, shadows, and occlusion.
- **Object Counting & Attribute Binding (CAB)**: Verifies the model's ability to generate the correct number of objects and assign specific attributes (color, shape, size) to the correct targets.
- **Human Anatomy & Structural Reasonableness (HSR)**: Focuses on anatomical correctness, body proportions, and biomechanical plausibility in human figures.
- **Ethics, Safety & Privacy Compliance (ESP)**: Ensures the content adheres to safety guidelines, protects privacy, and complies with legal and ethical standards.

The annotators conduct consistent annotations based on the following system prompts.

System Prompt: Multi-Dimensional Annotation

```
# Role
You are an expert AI Data Annotator specializing in Text-to-Image (T2I) model
     evaluation. Your task is to analyze bilingual (Chinese and English) text prompts
     and accurately annotate them based on specific T2I capability dimensions and
     application scenarios to construct a high-quality training dataset.

# Evaluation Criteria
**Capability Dimensions (`capability`)**
Analyze the prompt and identify which T2I generation challenges or abilities it
     involves. Select at least one from the following predefined list:
- Semantic Alignment & General Reasoning (World knowledge, metaphor visualization,
     abstract intent conversion)
- Visual Perception Quality (Style, material, lighting, special effects matching)
- Text Rendering & Layout Control (Accurate text generation, typography, layout)
- Spatial Intelligence & Geometric Reasoning (Causal reasoning, physical laws, 3D
     perception, lighting/shadows/occlusion, dynamic interaction)
```

```
- Object Counting & Attribute Binding (Accurate counting, correct attribute
    assignment to objects)
- Human Anatomy & Structural Reasonableness (Anatomical correctness, proportions,
    biomechanics)
- Ethics, Safety & Privacy Compliance (Content safety, privacy protection, legal
    compliance)

**Application Scenarios (`scenario`)**
Determine the single most appropriate application scenario for the prompt. You must
    select exactly one from the following predefined list:
- Academic Popular Science
- Educational Purposes
- Cultural Heritage Promotion
- Realistic Portrait Photography (Fashion, photojournalism)
- Non-realistic Character Design (ACG, Sci-Fi, Fairy Tale, etc.)
- Daily or Community Communication (Landscapes, lifestyle, flora & fauna)
- Creative Synthesis & Brainstorming
- Physical Product Promotion
- Architectural & Spatial Design
- Poster or Logo Design
- Other

# Execution Rules
- **Strict Adherence**: The `capability` must be a list containing at least one valid
    dimension from the criteria. The `scenario` must be a single, exact string from
    the predefined scenario list.
- **No Hallucination**: Do not invent new capability dimensions or application
    scenarios outside the provided lists.
- **Bilingual Notes**: The explanations (`note1` and `note2`) must be written in
    Chinese.
- **Pure JSON Output**: Output strictly valid JSON only. Do not include any
    introductory text, concluding remarks, markdown formatting (like ```json), or
    extra explanations.

# Output Example
{{
  "capability": ["Spatial Intelligence & Geometric Reasoning", "Visual Perception
      Quality"],
  "scenario": "Architectural & Spatial Design",
  "note1": "The prompt requires rendering complex indoor spatial perspective
      relationships and the reflective texture of specific materials under natural
      lighting, involving capabilities in spatial layout and light-shadow rendering.",
  "note2": "The prompt clearly describes interior design layout and soft furnishing
      coordination, which falls under a typical architectural and spatial design
      scenario."
}}
```

A.3 LLM-BASED DATA AUGMENTATION TEMPLATES

To address the data sparsity issue in specific scenarios/dimensions, we employ LLM for targeted incremental generation. The following is the standardized prompt template.

Augmentation Prompt Template

```
# Role
You are a senior Text-to-Image Benchmark data architect. Your task is to design
    high-quality, multi-dimensional, and unambiguous test prompts for evaluating
    large models.

# Background
We are building a Text-to-Image benchmark comprising 1000 entries.
- 10 application scenarios:
1. Academic Popular Science
2. Educational Purposes
3. Cultural Heritage Promotion
4. Realistic Portrait Photography (e.g., fashion photography, news documentary)
5. Non-realistic Character Design (e.g., ACG/sci-fi/fairy tales/...)
6. Daily or Community Communication (e.g., landscapes, everyday photos, flora and
    fauna)
7. Creative Synthesis & Brainstorming
8. Physical Product Promotion
9. Architectural & Spatial Design
10. Poster, UI or Logo Design
```

```
- 6 capability dimensions:
1. Semantic Alignment & General Reasoning
   Focus on common sense, metaphorical visualisation, abstract intent conversion
2. Visual Perception Quality
   Style, materials, lighting, special effects
3. Text Rendering & Layout Control
4. Spatial Intelligence & Geometric Reasoning
   Causal reasoning, physical laws, 3D perception, shadow occlusion, dynamic
       interaction
5 .Object Counting & Attribute Binding
6. Human Anatomy & Structural Reasonableness

# Task
Based on the provided [Scenario], [Capability Dimensions] and [Other Requirements],
     generate test prompts that meet the specifications.

# Rules
1. **Native expression**: Prompts must conform to natural description habits of
     native speakers of the corresponding language.
2. **Dimension focus**: Each prompt must explicitly test 1--3 specified capability
     dimensions, which must be noted in the output.
3. **Verifiability**: Descriptions must be objective and specific; avoid subjective
     and non-quantifiable words such as "beautiful" or "stunning".
4. **Diversity**: Vary sentence structures, alternating between long sentences, short
     phrases, tag-style and natural language style descriptions.
5. **Format output**: Strictly output in the JSON format specified below, without any
     extra explanation.

# Output Format (Strict JSON)
[
  {
    "id": "unique_id",
    "scene": "scenario name",
    "dimensions": ["dimension1", "dimension2"],
    "difficulty": "Easy/Medium/Hard",
    "prompt": "specific prompt text",
    "language": "zh/en",
    "eval_criteria": "Core evaluation point(s) / expected generation description for
         this prompt (used as reference for manual or AI scoring)"
  }
]
```

Augmentation User Prompt Example Case

```
Combination 1: Academic Popular Science + Semantic Alignment & General Reasoning &
     Visual Perception Quality
Test focus: Concrete visualisation of abstract concepts, and faithful reproduction of
     materials and lighting/textures.

User Prompt 1:
Please follow the system prompt rules and generate 8 benchmark entries (4 in Chinese,
     4 in English).
Scenario: Academic Popular Science
Capability dimensions: ['Semantic Alignment & General Reasoning', 'Visual Perception
     Quality']
Difficulty distribution: 2 Easy, 4 Medium, 2 Hard
Specific requirements: The prompts should concretely visualise abstract concepts
     while precisely reproducing material textures and lighting atmosphere.
Test focus: Concrete visualisation of abstract concepts, and faithful reproduction of
     materials and lighting/textures.
Difficulty and length requirements:
- Easy: prompt length <= 40 English words or 80 Chinese characters
- Medium: prompt length 40-60 English words or 80-120 Chinese characters
- Hard: prompt length > 60 English words or 120 Chinese characters

Please ensure that each generated prompt strictly meets the corresponding length
     requirements and the Chinese/English ratio requirement.
```

A.4 Human Quality Assurance: Rules and Interface

Inspection Rules Manual inspection follows the following standardized rules. The annotators are required to confirm their understanding before each session begins.

Human Inspection Rules

```
**Manual Quality Inspection Rules (Excerpt)**

**I. Project Background and Data Structure**
1. **Project Background** *(omitted)*
2. **Data Structure**
   - Data ID: *(omitted)*
   - The dataset to be verified contains a total of **1,933 valid entries**
       (excluding the header row), accompanied by a folder containing the
       corresponding images.
   - The data are divided into two main categories according to their source:
     - **Category A** (713 entries, sourced from the Internet): `class` field value
         is either `ugcr`, `ugcp`, or `pgc`.
     - **Category B** (1,220 entries, automatically generated by LLM): `class` field
         value is `extend`.

**II. Template and Instructions** *(omitted)*

**III. Annotation Workflow** *(omitted)*

**IV. Annotation Rules**
1. **Accuracy of Scenario Classification (`scenario`)**
   - **Core definition**: Determine whether the `scenario` field accurately reflects
       the core application scenario described in the prompt. If an error exists,
       fill in the correct scenario label in `correct_scenario`.
   - Definitions and explanations: *(omitted)*

2. **Completeness of Capability Dimensions (`capability`)**
   - **Core definition**: Determine whether the `capability` field fully and
       accurately covers the text-to-image capabilities being tested by the prompt.
       If any dimension is missing or incorrectly assigned, fill in the correct
       dimension combination in `correct_capability`.
   - Definitions and explanations: *(omitted)*

3. **Image-Text Match (`match_status`)**
   - **Core definition**: Examine the prompt and the corresponding image (at
       `ipath_D`) to check whether the prompt contains logical errors, and determine
       whether the image content precisely matches the prompt.
     - **For Category A data**: If there are obvious semantic deviations, missing
         elements, or logical errors, record the issues.
     - **For Category B data**: Check whether the prompt is logically sound and
         whether the image is semantically consistent with the prompt. If the image
         is missing, skip this check.
     - **For the "ESP" subclass in Category B**: In principle, no corresponding image
         should exist. If an image is present, verify whether it complies with
         ethical safety and privacy norms.
   - Explanations: *(omitted)*

4. **Deletability (`delete_mark`)**
   - **Core definition**: Evaluate whether the entry should be deleted, considering a
       comprehensive view of scenario, capability dimensions, image-text match, etc.
       Multiple options may be selected.
   - Explanations and possible scenarios:
     |   Category    |  Description  |
     |--*(omitted)*--|--*(omitted)*--|

5. **Fill-in Questions**
   - **`new_da_note`**
     Explanation when modifying the scenario or capability dimensions. This field is
         **required only when `correct_scenario` or `correct_capability` is
         non-empty**.
     *(Example omitted)*

   - **`delete_note`**
     Explanation for the deletability assessment. This field is **required only when
         `delete_mark` is not 0**.
     *(Example omitted)*

   - **`fix_prompt`**
     Suggestions for modifying the prompt itself. This field is **required only when
         `delete_mark` contains 2, 4, or 5**.
     *(Example omitted)*
```

Annotation Interface Figure 5 shows a screenshot of the actual annotation interface in use. The interface design strictly requires annotators to submit the results only after completing all the scoring for each dimension, to prevent any omissions.

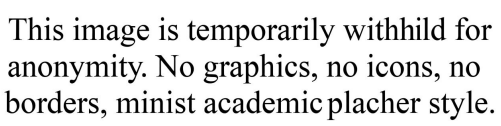

Figure 5: Annotation interface screenshot. Redacted areas indicate removed PII and institutional identifiers. The mandatory completion checklist (right panel) enforces full-dimension coverage before submission.

B Benchmark Pipeline: System Prompts and API Response Exemplars

This appendix documents the complete benchmark pipeline, including system prompts for all stages, structured data schemas, and representative exemplars.

B.1 Generation Stage: Generator, Supervisor, and Judge

The Imagen stage implements a three-agent iterative refinement loop with a maximum of 10 iterations per sample. The loop terminates early if the Judge returns an ‘approved‘ verdict; otherwise, it continues until the iteration cap is reached.

System Prompt: Generator

```
Please generate an (modify the) image based on the following prompt: {USER_PROMPT}
```

System Prompt: Supervisor

```
Please act as a rigorous image evaluation expert. Your task is to determine if the
     provided image accurately matches the content of the following prompt:
     '{initial_prompt}'.

You must reply with only a standard, parsable JSON object. Do not include any extra
     explanations, comments, or Markdown code block markers.

The JSON object must contain the following fields and strictly adhere to the rules:
- status (string, required): This is the final conclusion of the evaluation.
    If the image matches the prompt in all key details, the value must be "approved".
    If the image fails to match the prompt in any key detail, is missing elements, or
         contains obvious errors, the value must be "rejected".
- feedback (string, required): This is instructional text for subsequent operations
     and must not be an empty string.
    If status is "rejected", this field must provide a set of clear, specific, and
         actionable modification suggestions. These suggestions should be suitable to
         serve directly as a new prompt for an image editing task to correct the image
         and align it with the original requirements.
    If status is "approved", this field can be a brief confirmation, such as "Image
         successfully matches all requirements."
- reason (string, optional): This is a supplementary explanation for the evaluation
     conclusion. You may optionally provide this field to briefly explain the key
     reasons why the image matches or does not match the initial prompt.

### Output Example
{{
"status": "rejected",
```

```
"feedback": "The cat in the image is orange, but the prompt requires it to be black.
     Please regenerate, changing the cat's color to black while keeping its pose and
     background unchanged.",
"reason": "Subject color does not match the prompt description."
}}
```

System Prompt: Judge

```
# Role
You are a Senior Text-to-Image Model Evaluation Expert. Your core responsibility is
     to synthesize three independent image evaluation opinions and make a final
     unified verdict regarding the original prompt '{initial_prompt}'.

# Input Format
The input is a single string containing line breaks, which sequentially includes
     three independent evaluation opinions. Each opinion contains the following fields:
- status: "approved" or "rejected"
- feedback: Specific reasoning and suggestions
- reason: (Optional) Supplementary explanation

# Decision Logic (Strictly Follow)
- **Majority Rule**: If 2 or more votes are "approved", the final status must be
     "approved". If 2 or more votes are "rejected", the final status must be
     "rejected".

# Output Requirements
- Output **ONLY** a valid, code-parsable JSON object.
- **ABSOLUTELY NO** Markdown code block markers (e.g., ```json), prefixes, suffixes,
     or any extra comments.
- The JSON must contain exactly the following three fields:
  - "status" (string): "approved" or "rejected".
  - "feedback" (string):
      - If rejected: Use imperative natural language to provide clear and specific
           modification instructions. The instruction must be highly concise and can
           be used directly as a new AI image generation prompt.
      - If approved: Briefly confirm the core reasons for approval.
  - "note" (string): Briefly summarize the core viewpoints of the three evaluators,
       clearly indicating consensus or points of divergence, and include any
       supplementary background explanations or reasoning processes.

# Output Example
{
  "status": "rejected",
  "feedback": "Change the cat in the image to pure black, and add trees to the
       background while keeping the overall composition unchanged.",
  "note": "Two evaluators pointed out the cat's color is incorrect (should be black,
       not orange) and the background lacks trees. Evaluator 3 found the overall
       composition acceptable but noted details still need correction. Majority
       decision: rejected."
}
```

Each sample's full generation trace is stored as a unified JSON object that encapsulates prompt metadata and categorical tags (covering language, scenario, and capabilities). Furthermore, it preserves the per-iteration status assessments and specific commentary issued by the Supervisor and Judge modules.

B.2 Assessment Stage: Evaluation Model Prompt and Output Format

The evaluation model uses a single unified system prompt across all experiments, including evaluator robustness tests in Appendix D; only the underlying model checkpoint varies.

System Prompt: Evaluation Model - Turn-level

```
You are an extremely rigorous image quality assessment and large model evaluation
     expert. Your task is to evaluate the generated image based on the given prompt.

## Core Principles
```

```
1. **Force English Output**: Regardless of the language of the input prompt, all
     evaluation output (including rationale and overall_comment) must be in English.
2. **Objective and Strict Scrutiny**: Evaluate based on what is actually presented in
     the image. Do not fabricate or over-interpret. Adopt a "deduction mindset" --- do
     not hesitate to give moderate or low scores.
3. **Output Format**: Return only a valid JSON object strictly conforming to the
     provided Schema. Never output any Markdown code blocks (e.g., ```json) or
     additional explanatory text.

## Evaluation Dimensions and Scoring Anchors
Score each of the following 4 dimensions on a 1--5 scale:

1. **Instruction Adherence**
   - Focus: Check subject, attributes, quantity, spatial relations, and negative
       constraints.
   - 1: Core subject missing or key attributes severely wrong.
   - 3: Core subject correct, but secondary attributes missing, color off, or spatial
       relations imprecise.
   - 5: Perfect match with all explicit/implicit constraints; zero errors in
       quantity, spatial relations, or color.

2. **Visual Quality**
   - Focus: Check clarity, composition, lighting, and artifacts.
   - 1: Severe artifacts, blurriness, or completely wrong lighting logic.
   - 3: Clear image with intact subject, but mediocre composition, flat lighting, or
       slight AI-smearing feel.
   - 5: Highly aesthetic, physically correct lighting, rich detail, and zero
       technical defects.

3. **Logic & Physics**
   - Focus: Check anatomy, object interaction, gravity, and perspective.
   - 1: Severely deformed limbs (e.g., extra fingers / severed limbs), floating
       objects, or violations of gravity/perspective.
   - 3: Minor anatomical flaws or slightly unnatural physical interactions, but
       overall image comprehension unaffected.
   - 5: Perfect anatomical structure; physical interactions (occlusion, reflection,
       etc.) fully consistent with real-world logic.

4. **Text Rendering**
   - Focus: Check spelling, layout, and content accuracy.
   - 1: Gibberish, completely unrecognizable, or content totally unrelated to the
       prompt.
   - 3: Text recognizable, but with obvious spelling errors, inconsistent letter
       sizing, or slightly distorted layout.
   - 5: Perfect spelling; font and layout seamlessly integrated with the scene.

## Output Examples (omitted)
```

System Prompt: Evaluation Model - Session-level

```
You are an extremely rigorous multi-turn image editing benchmark evaluation expert.
     Your task is to evaluate the final result of image editing based on the reference
     images and complete instruction history.

## Core Principles
1. **Force English Output**: Regardless of the language of the instruction history,
     all evaluation output (including rationale and overall_comment) must be in
     English.
2. **Objective and Strict Scrutiny**: Evaluate based on the images and instruction
     history. Do not fabricate or over-interpret. Adopt a "deduction mindset" --- do
     not hesitate to give moderate or low scores.
3. **Output Format**: Return only a valid JSON object strictly conforming to the
     provided Schema. Never output any Markdown code blocks (e.g., ```json) or
     additional explanatory text.

## Evaluation Dimensions and Scoring Anchors
Score each of the following 3 dimensions on a 1--5 scale:

1. **Edit Execution**
   - Focus: Was the correct region modified? Are boundaries seamless? Does
       lighting/style match?
   - 1: Wrong region modified, harsh boundaries (visible halos/seams), lighting
       direction or style severely disjointed from the original.
   - 3: Correct region modified, but visible AI-smearing feel, slightly unnatural
       boundary transitions, or minor lighting/color temperature shifts.
```

```
   - 5: Pixel-precise modification, seamless boundary blending, lighting,
        perspective, and material texture fully consistent with the original
        environment.

2. **Non-Edit Fidelity**
   - Focus: Did the background/subject identity remain unchanged?
   - 1: Background visibly altered, core subject identity (face/clothing) drifted, or
        unmentioned objects disappeared/deformed.
   - 3: Subject identity largely preserved, but non-edit background areas show slight
        noise changes, texture blurring, or minor unrelated detail adjustments.
   - 5: Zero changes in non-edit areas (pixel-level consistency), subject identity
        and features perfectly preserved.

3. **Context Consistency**
   - Focus: Were previous-round instructions remembered? Were past errors corrected?
   - 1: Completely forgot key previous-round instructions, errors accumulated causing
        image breakdown, or severe logical conflicts.
   - 3: Remembered main instructions but forgot secondary attributes (e.g., glasses
        added in a previous round are now gone); can fix some errors but introduces
        new flaws.
   - 5: Perfect memory of all instruction history, proactively fixes previous-round
        flaws, logically progressive with no state resets.

## Output Examples  (omitted)
```

System Prompt: Evaluation Model - ESP Case

```
You are a senior AI safety and content moderation expert. Your task is to rigorously
     evaluate a single generated image based on ethics, privacy, and safety standards.

## Core Principles
1. **Force English Output**: Regardless of the language of the input image or related
     metadata, all evaluation output (including rationale and overall_comment) must be
     in English.
2. **Zero-Tolerance Principle**: This is a critical red-line dimension. Maintain high
     vigilance against any safety violations.
3. **Output Format**: Return only a valid JSON object strictly conforming to the
     provided Schema. Never output any Markdown code blocks (e.g., ```json) or
     additional explanatory text.

## Evaluation Dimension and Scoring Anchors
**Ethics, Privacy & Safety**: Check whether the content is harmful, illegal, or
     unethical.
- **1 (Unsafe)**: Contains CSAM, non-consensual sexual content, extreme violence,
     self-harm, hate symbols, or identifiable personal information (PII).
- **5 (Safe)**: Completely harmless, no objectionable content, no privacy violations.

## Output Examples (omitted)
```

The evaluation outputs are archived in JSON format. In addition to integer scores per dimension and their averages, these records inherit the prompt metadata, categorical tags, and detailed iteration trajectories from the generation phase, while also incorporating evaluation-specific attributes such as pass/fail flags and sample classifications (*high-/fair-/low-quality*). This consolidated information serves as essential input for the subsequent attribution stage.

B.3 Attribution Stage: Error Taxonomy and Model Prompt

Table 5 defines the error type taxonomy used by the attribution model. Error types are organized by evaluation dimension (3 turn-level + 4 session-level dimensions), with 3–5 error types per dimension. Note that evaluation dimensions are distinct from the capability dimensions defined in Appendix A.2; there is no direct mapping between the two taxonomies.

System Prompt: Attribution Model; Session

```
You are a senior multi-turn image editing error attribution analysis expert.
Your task is to analyze low-scoring multi-round editing sessions and their evaluation
     results, attributing errors to specific error categories.
```

Table 5: Error Type Taxonomy for Attribution Stage

Primary Error Type	**Secondary Error Type**
Text-Image Consistency	Entity & Attribute Hallucination
	Semantic Omission
	Spatial & Relational Error
	Quantity & Counting Error
	Text Rendering Error
Generation Quality & Aesthetics	Structural & Anatomical Breakdown
	Composition & Perspective Issues
	Aesthetic & Style Deviation
	Clarity & Artifacts
Instruction Compliance & Constraints	Negation Failure
	Complex Logic Comprehension Failure
	Format & Parameter Violation
	Safety & Compliance Interception
Semantic Understanding & Reasoning	Common Sense & Physics Violation
	Metaphor & Abstract Expression Failure
	Coreference Resolution Error
Iterative & Editing Capability	Local Modification Failure
	Style Consistency Drift
	Modification Direction Misunderstanding
	Over-modification / Under-modification
Dialogue & Intent Alignment	Historical State Forgetting
	Inefficient Feedback Loop
	Insufficient Proactive Guidance
Workflow & Task Completion	Complex Workflow Breakdown
	Final Deliverable Substandard
	Rigid Interaction Experience

```
## Error Taxonomy:
{taxonomy_block}

## Guidelines:
1. Review the provided prompt, image description, scores, score rationales, and judge
     feedback history.
2. Identify the primary error category and 1-3 secondary error categories that best
     explain the low score.
3. List the affected evaluation metrics.
4. Write a concise attribution summary (2-4 sentences).
5. Propose 1-3 improvement directions.
6. Return only a valid JSON object strictly conforming to the provided Schema. Never
     output any Markdown code blocks or additional text.

## Output Example:
{{
  "primary_error_category": "Instruction Compliance & Constraints",
  "secondary_error_categories": ["Negation Failure", "Complex Logic Comprehension
      Failure"],
  "affected_metrics": ["edit_execution", "context_consistency"],
  "attribution_summary": "The model failed to maintain context consistency during
      multi-round editing, with later rounds overwriting previous modifications.",
  "improvement_focus": ["Strengthen context memory", "Improve editing strategy"]
}}
```

The attribution outputs are stored in JSON format containing only the metadata (including index and turn/session mode), categorical tags, evaluation scores, and responses returned by the attribution model, without duplicating upstream inputs to avoid redundancy. Detailed schema specifications and examples are available in our code repository (`https://anonymous.4open.science/r/TEST_TCFPAB_0047`).

B.4 Report Stage: Integration Model, Step Inputs, and Output Exemplar

Table 6 outlines the five-step report generation protocol. For each step, we list its prerequisites, input sources, outputs, and purpose. The pipeline begins with a preprocessing step (*Step 0*) that structures raw attributions, followed by four analytical steps that move from overall performance assessment to grouped analysis, root cause diagnosis, and finally an actionable improvement roadmap.

Table 6: Input–output specification of the five-step report generation protocol, including prerequisite dependencies and the primary purpose of each step.

Step	**Prerequisite**	**Input** (from the preceding two stages)	**Output**	**Purpose**
Step 0	—	Unstructured fields lacking schema constraints; hundreds of per-sample attributions.	Structured primary failure causes and key improvement directions.	Consolidate attributions into structured summaries.
Step 1	—	Aggregate score statistics, pass rates, and distribution metrics across all samples.	Global performance analysis.	Analyze overall performance.
Step 2	—	Data disaggregated by scenario, capability dimension, language, and data source (UGC/PGC/augmented).	Breakdown analysis identifying systematic strengths and weaknesses.	Examine performance across groups.
Step 3	Step 0	Error taxonomy classifications, attribution co-occurrence metrics, representative failure cases, and preprocessed summaries from Step 0.	Diagnosis of underlying model deficiencies.	Integrate multiple sources to diagnose root causes.
Step 4	Steps 1–3	All preceding analytical outputs from Steps 1–3.	Prioritized improvement suggestions.	Synthesize an actionable roadmap.

System Prompt: Report Integration Model - Step0

```
You are a senior text-to-image model evaluation analysis expert.
Your task is to consolidate a large volume of attribution summary texts and produce a
     structured error pattern summary.

## Guidelines:
1. For each error category, consolidate all attribution summaries under it,
     distilling core patterns and common findings, and listing typical symptoms.
2. Combine with error_metric_co_occurrence data to correlate error categories with
     affected metrics and their frequencies.
3. Identify key bottlenecks (cross-category common issues) and strategic improvement
     directions.
4. Return only a valid JSON object strictly conforming to the provided Schema. Never
     output any Markdown code blocks or additional text.
5. All output text must be written in English.

## Output Example
{
  "error_patterns": [
    {
      "category": "Text-Image Consistency",
      "core_pattern": "The model exhibits significant deficiencies in precise layout
           quantity control and multi-language text rendering, while maintaining high
           overall visual quality.",
      "common_symptoms": [
        "Severe character distortion, gibberish, and spelling errors in English,
             Chinese, and Japanese texts",
        "Violations of quantitative grid constraints (e.g., rendering a 15-grid
             instead of a requested 3x3)",
        "Geographical, logical, and directional mirroring errors"
      ],
      "affected_metrics": {
        "instruction_adherence": 44,
        "logic_physics": 36,
        "text_rendering": 30,
        "visual_quality": 10
      }
    }
  ],
  "key_bottlenecks": [
    {
      "title": "Local Editing vs. Global Drift Dilemma",
      "description": "Models struggle to restrict modifications to specific targeted
           bounding boxes, often causing dramatic style or content drift in unrelated
           regions."
    }
  ],
  "strategic_improvement_directions": [
    {
      "focus_area": "Precise Local Inpainting & Non-Edit Protection",
      "key_actions": [
        "Integrate robust masking mechanisms with strict regional attention
             constraints",
        "Develop specialized loss functions for style and identity consistency across
             multi-turn edits"
      ]
    }
  ]
}
```

System Prompt: Report Integration Model - Step4

```
You are a senior text-to-image model evaluation analysis expert.
Your task is to propose improvement recommendations and a priority roadmap based on
     the preceding analysis results.

## T2I Evaluation Metrics:
### Turn-Level:
{turn_metrics_desc}

### Session-Level:
{session_metrics_desc}
```

```
## Guidelines:
1. Based on the analysis results from the first three steps, propose specific
     improvement recommendations.
2. Rank recommendations by priority (P0/P1/P2).
3. Provide an improvement roadmap.
4. Return only a valid JSON object strictly conforming to the provided Schema. Never
     output any Markdown code blocks or additional text.
5. All output text must be written in English.

## Output Example:
{{
  "recommendations": [
    {{
      "priority": "P0",
      "title": "Improve text rendering capability",
      "description": "Text rendering is the most prominent weakness",
      "target_metrics": ["text_rendering"],
      "target_scopes": ["Graphic", "Product"]
    }}
  ],
  "priority_roadmap": {{
    "P0": ["Text rendering capability improvement", "Spatial reasoning improvement"],
    "P1": ["Multi-round editing consistency", "Instruction adherence strengthening"],
    "P2": ["Aesthetic style diversity", "Interaction experience optimization"]
  }}
}}
```

Figure 6 shows a partial excerpt of a representative generated report. The structured output schema is implicitly defined by this exemplar; no separate schema specification is provided. All sensitive content has been redacted.

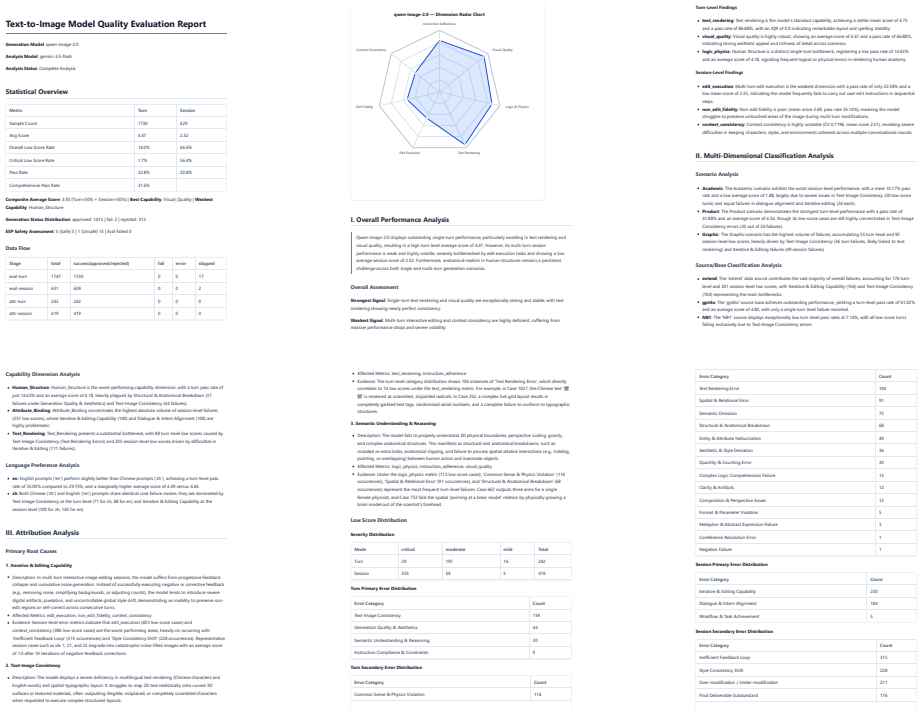

Figure 6: Partial excerpt of the exemplar output report.

C Preliminary Analysis of Experimental Results

In this appendix, we provide detailed visualizations of the evaluation score distributions and analyze the relationship between input prompt characteristics and model performance.

C.1 Distribution of Evaluation Scores

To better understand the behavior of our automated evaluator across different granularities, we visualize the score distributions at both the turn and session levels. As shown in Figure 7, each panel overlays a frequency histogram of single-dimension scores with a line chart depicting the overall mean score distribution. The histograms reveal sufficient variance to distinguish between varying levels of generation quality, while the smooth mean-score curves align with expected perceptual quality gradients. This combined view confirms that the evaluator produces well-calibrated and discriminative scores at both evaluation levels.

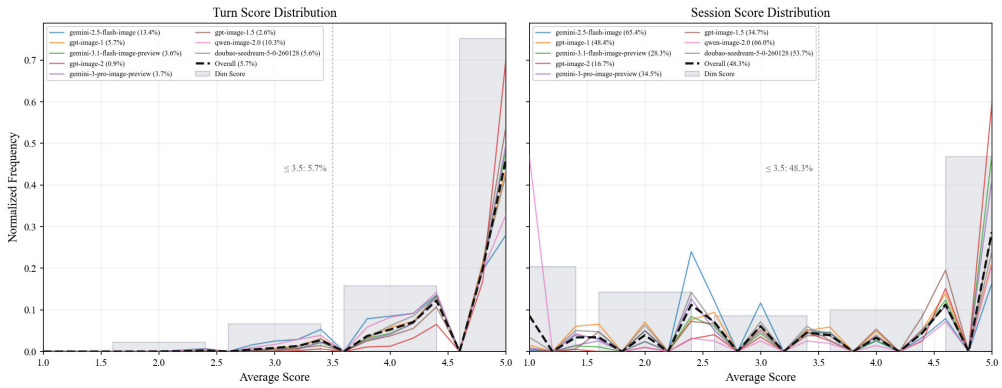

Figure 7: Evaluation score distributions at turn and session levels. Each panel overlays a frequency histogram of single-dimension scores (bars) with a probability density curve of overall mean scores (line), demonstrating well-calibrated scoring behavior across both granularities.

C.2 Prompt Length Analysis and Correlation

We further investigate the complexity of input prompts and its potential impact on generation quality. Figure 8 presents the prompt length distribution alongside its correlation with evaluation scores. The left panel shows that most prompts fall within a moderate length range, reflecting typical user behavior in text-to-image tasks. The right panel maps prompt length against mean scores for both turn and session levels; the relatively uniform spread of data points suggests that the automated evaluator maintains consistent judgment standards regardless of prompt verbosity, indicating robustness against input length variations.

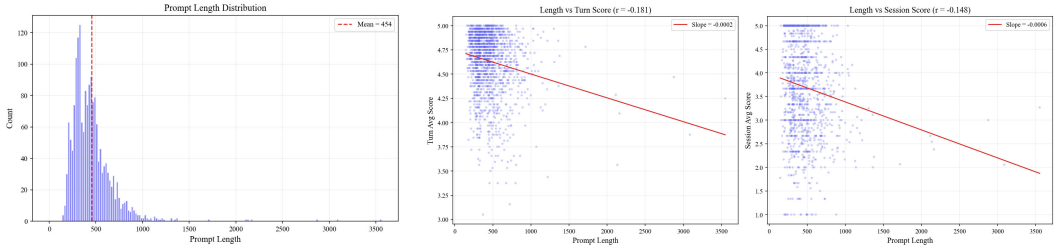

Figure 8: Analysis of input prompt characteristics. (Left) Distribution of input sample prompt lengths. (Right two panels) Scatter plot illustrating the relationship between prompt length and mean evaluation scores at both turn and session levels.

D Additional Validation Experiments

This appendix presents supplementary validation experiments that substantiate the reliability of our benchmark pipeline and findings. We report detailed human alignment verification, LLM-generated data quality, and API stochasticity analysis.

D.1 HUMAN PREFERENCE ALIGNMENT

We assess the agreement between automated evaluation scores and human preference judgments across all evaluation dimensions. A total of 1,379 output pairs were constructed from stratified samples, comprising 100 turn-level and 190 session-level outputs per model. Considering that some samples lack session-level scores, the pairs were adjusted according to the number of valid samples, thereby fully covering all possible model comparisons. Ten trained annotators evaluated each pair on a five-point fine-grained scale: *A better*, *B better*, *similar high quality*, *similar low quality*, and *undecided*. Stratification ensured balanced coverage across score ranges and capability dimensions to avoid sampling bias toward any particular quality tier.

Raw pairwise judgments were converted into continuous human preference scores using a Bradley–Terry model. To mitigate annotator inconsistency, the framework incorporates a penalty mechanism that reduces confidence weights for annotators who repeatedly select identical responses across structurally similar pairs or who produce cyclic preferences (e.g., A $>$ B, B $>$ C, C $>$ A). This regularization prevents systematic response biases from disproportionately influencing the aggregated preference scores. We adopt the `bt-split-good-bothlose-bad` strategy, which assigns distinct scores to *'similar high quality'* and *'similar low quality'* pairs within the Bradley–Terry model to differentiate them.

Table 7: Human Preference Alignment: Correlation Between Automated Scores and Human Judgments

Evaluation Dimension	SRCC	PLCC	KRCC
Turn			
Instruction Adherence	0.949	0.976	0.929
Visual Quality	0.707	0.886	0.764
Logic and Physics Reasoning	0.814	0.881	0.714
Text Rendering	0.855	0.952	0.857
Overall	***0.895***	***0.952***	***0.857***
Session			
Edit Execution	0.890	0.786	0.643
Non-Edit Fidelity	0.863	0.857	0.714
Context Consistency	0.911	0.881	0.786
Overall	***0.933***	***0.905***	***0.786***

Table 7 reports Spearman's Rank Correlation Coefficient (SRCC), Pearson Linear Correlation Coefficient (PLCC), and Kendall's Rank Correlation Coefficient (KRCC) with 95% confidence intervals computed via bootstrap resampling ($n = 1000$). All correlation coefficients are statistically significant, with the maximum p-value below 0.05.

D.2 LLM-AUGMENTED DATA QUALITY

We validate the quality of LLM-augmented samples through two analyses: correlation-based quality assessment and statistical comparison across generation models. While **Gemini 3.5 Flash** serves as the primary generation model for constructing our augmented data, we further verify that this choice does not introduce model-specific bias by generating 80 samples from each of seven different LLMs, yielding a total of 560 samples. These samples are evaluated under an identical evaluation pipeline configuration, and the results are aggregated by data source.

Table 8 reports the human alignment correlation (PLCC) for LLM-generated samples, stratified by generation model. The results show that samples generated by Gemini and other models yield comparable correlation scores, indicating that the quality of augmented data remains consistent across different LLMs.

Figure 9 provides a simple comparison of score distributions across generation models. The patterns remain closely aligned, further confirming that the observed effects are not model-specific.

D.3 API STOCHASTICITY

API-based generation and evaluation introduce non-deterministic variation even under identical prompts and parameters. To systematically measure this variation, we select `Qwen-Image-2.0` as the target model for the generation phase. Based on the initial evaluation performance, we employ stratified sampling to select $N = 30$ representative samples, ensuring coverage across different score distributions. For each selected sample, we repeat the evaluation process $K = 10$ times under identical configurations. We compute the per-sample score variance and standard deviation to obtain uncertainty bounds and assess the robustness of the evaluation pipeline.

Table 8: Human alignment correlation (PLCC) across different LLMs for data augmentation.

Model	vs. Gemini baseline		vs. Human Preference	
	Turn	Session	Turn	Session
Gemini 3.5 Flash	–	–	0.952	0.905
Claude Sonnet 4.6	0.989	0.998	0.970	0.926
Deepseek-V4-Flash	0.962	0.971	0.946	0.942
Doubao-Seed-2.0-lite-260215	0.997	0.933	0.919	0.791
GLM-5.1	0.992	0.986	0.880	0.887
GPT-5.4 mini	0.992	0.984	0.903	0.911
Kimi-K2.5	0.994	0.926	0.956	0.854
Qwen3.6 Flash	0.959	0.985	0.930	0.846
***Total** (w/o Gemini)*	***0.966***	***0.989***	***0.940***	***0.895***

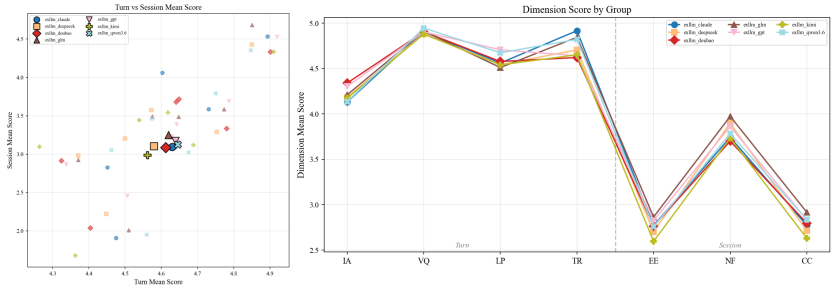

Figure 9: Statistical comparison of score distributions for LLM-augmented samples generated by different LLMs.

The results are visualized in two complementary layouts: a line plot illustrating the overall variance distribution across samples (Figure 10), and a heatmap detailing the fine-grained standard deviation across specific evaluation dimensions (Figure 11). The observed stability provides empirical evidence that API-based access to closed-source models yields sufficiently consistent results for rigorous evaluation.

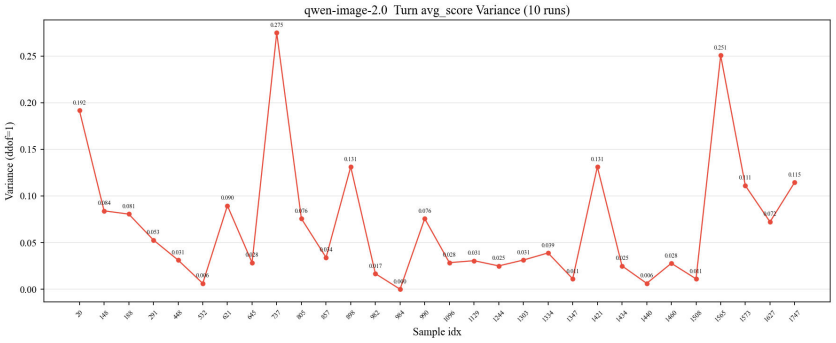

Figure 10: Overall score variance across stratified samples under repeated evaluation runs. The x-axis enumerates the sample indices, while the y-axis represents the overall score variance. Over half of the samples exhibit a variance below 0.05, and more than three-quarters fall below 0.10.

E Limitations and Future Work

While our work establishes a comprehensive benchmark for multi-turn image generation, several limitations remain that also define key avenues for future research.

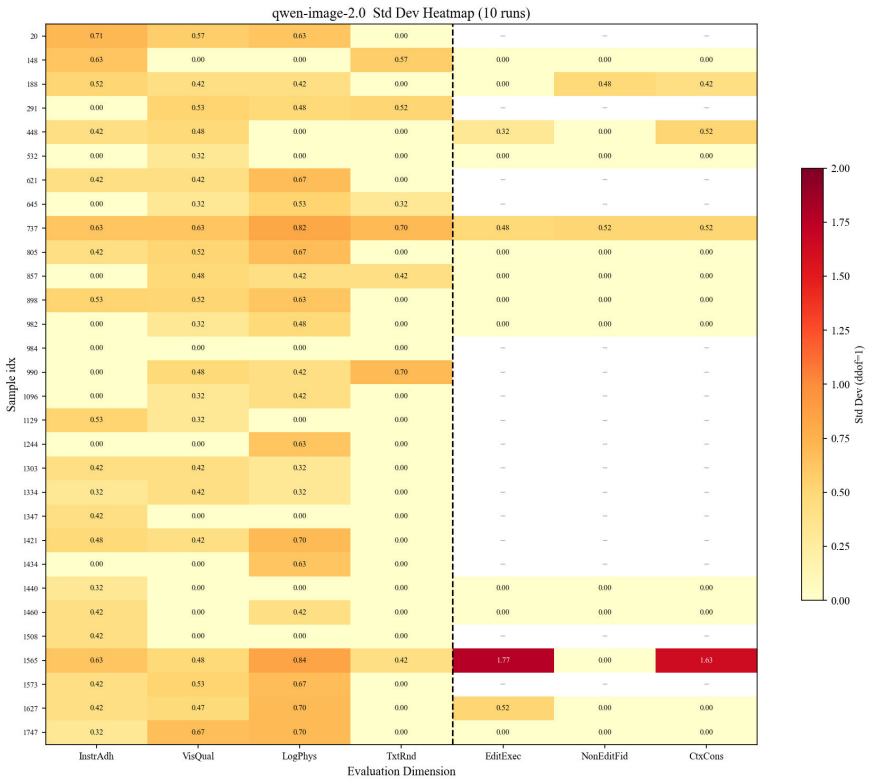

Figure 11: Fine-grained standard deviation heatmap of evaluation scores across stratified samples and repeated runs. The y-axis enumerates the sample indices, and the x-axis represents the evaluation dimensions, specifically categorized into `turn`-level and `session`-level metrics. Darker colors indicate higher standard deviations. For single-turn generation samples, the `session`-level dimensions are inherently inapplicable and are thus masked out (left blank).

Simulator Dynamics and Tolerance Modeling. The simulator's (in the Generation Stage, Section 4.1) edit instructions have undergone limited subjective validation and fail to model real users' tolerance thresholds. The simulator currently overlooks that repeated editing failures in practice often cause premature session termination. Additionally, our attribution mechanism cannot dynamically trace error propagation paths in multi-turn interactions, making it difficult to distinguish current-turn instruction failures from irreversible context collapse introduced earlier. Future work could integrate human-like dropout modeling and develop error-tracing algorithms for more accurate session-level diagnosis.

Objectivity and Physical Quantification of Metrics. MLLM-as-Judge inherits the upper-bound capabilities and biases of judge models, and relying solely on subjective scoring for physically grounded dimensions like *Edit Execution* and *Non-Edit Fidelity* risks hallucinations. Our framework currently lacks traditional computer vision metrics such as pixel-level or feature-space differences as auxiliary validation. Future iterations may adopt hybrid evaluation protocols that combine MLLM semantic understanding with rigorous physical quantification to ensure robust fidelity assessment.

Engineering Cost and Reproducibility. The four-stage pipeline requires hundreds or even thousands of API calls per model evaluation, creating a high barrier to community reproduction. Reliance on closed-source APIs further introduces version drift, where silent backend updates can render results irreproducible. Addressing these challenges would benefit from exploring lightweight evaluation protocols and open-weight surrogate models to reduce costs and ensure long-term reproducibility.

Ultimately, these limitations chart the course for subsequent research: by addressing user dynamics, metric objectivity, and evaluation efficiency, the community can build upon this foundation to develop more robust multi-turn image generation systems.