GUI-GENBENCH: EVALUATING IMAGE GENERATION MODELS AS INTERACTIVE GUI ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

An image generation model used as a graphical user interface (GUI) environment must produce not only a plausible screen but also the correct response to an action. Existing visual-generation benchmarks do not directly assess this combination of functional correctness and temporal consistency in GUIs. We introduce **GUI-GenBench**, a benchmark of 700 curated samples across five task suites: single-step transitions, multi-step planning, fictional-app generation, real-app trajectories, and coordinate-based grounding. We also introduce **GUI-Score**, a VLM-based metric that assesses Goal Achievement, Interaction Logic, Content Consistency, UI Plausibility, and Visual Quality. Evaluating 12 image generation models on Chinese and English subsets, we find that strong single-step performance does not consistently extend to coherent multi-step trajectories or accurate grounding. The highest aggregate GUI-Scores are 69.62 on the Chinese subset and 63.16 on the English subset. Qualitative analysis further identifies failures in text rendering, icon interpretation, and coordinate-conditioned transitions. Together, these results distinguish visual plausibility from functional reliability and identify limitations that must be addressed before generated GUIs can serve as dependable interaction environments.

1 INTRODUCTION

Image generation models (Hurst et al., 2024; Comanici et al., 2025; Google, 2025b; Seedream et al., 2025; Labs, 2025; Wan et al., 2025) can synthesize a new graphical user interface (GUI) state from a current screen and a user instruction. This capability motivates their use as **generative GUI environments**: models that produce visual responses to actions, with potential applications in agent training and evaluation (Zhang et al., 2025a; Yan et al., 2025; Xie et al., 2025b; Garg et al., 2025; Luo et al., 2025; Wei et al., 2023). Conventional interaction environments depend on implemented dynamics or executable software (Cobbe et al., 2020; Xie et al., 2024; Bonatti et al., 2024; Zhang et al., 2025a). Generating screens offers a complementary way to construct diverse interface examples and trajectories (Zhao et al., 2021; Liu et al., 2025b), but raises a basic question: do the generated states faithfully reflect the requested interaction?

Existing work on visual generation and evaluation examines fidelity, semantic alignment, composition, and temporal consistency (Ghosh et al., 2023; Hu et al., 2024; Niu et al., 2025; Zhao et al., 2025; Huang et al., 2023; 2024; Sun et al., 2025; Zhuang et al., 2025). These properties alone do not establish the functional correctness of GUI transitions (Xie et al., 2025a; Yan et al., 2025). As Figure 1 illustrates, a GUI environment must model discrete, action-triggered changes (Zhang et al., 2025a): a single click can replace an entire screen, while a small text or control-state error can change the meaning of the response. Correct transitions require coordinate grounding (Zhao et al., 2021; Cheng et al., 2024), icon interpretation (Liu et al., 2025b; Xie et al., 2025a), and readable text (Chen et al., 2024), together with logical continuity across changing layouts (Li et al., 2025a). Evaluation must therefore test both what a screen looks like and whether it represents the intended state.

We introduce **GUI-GenBench** to evaluate image generation models along these functional and visual requirements. Its 700 curated samples span five task suites, from local state transitions to multi-step trajectories, fictional scenarios, and coordinate-based interactions. We pair the benchmark with **GUI-Score**, a rubric-based metric evaluated by vision-language models (VLMs) (Google, 2025a; Hurst et al., 2024; Bai et al., 2025). Five scoring dimensions separate goal achievement and interaction

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.

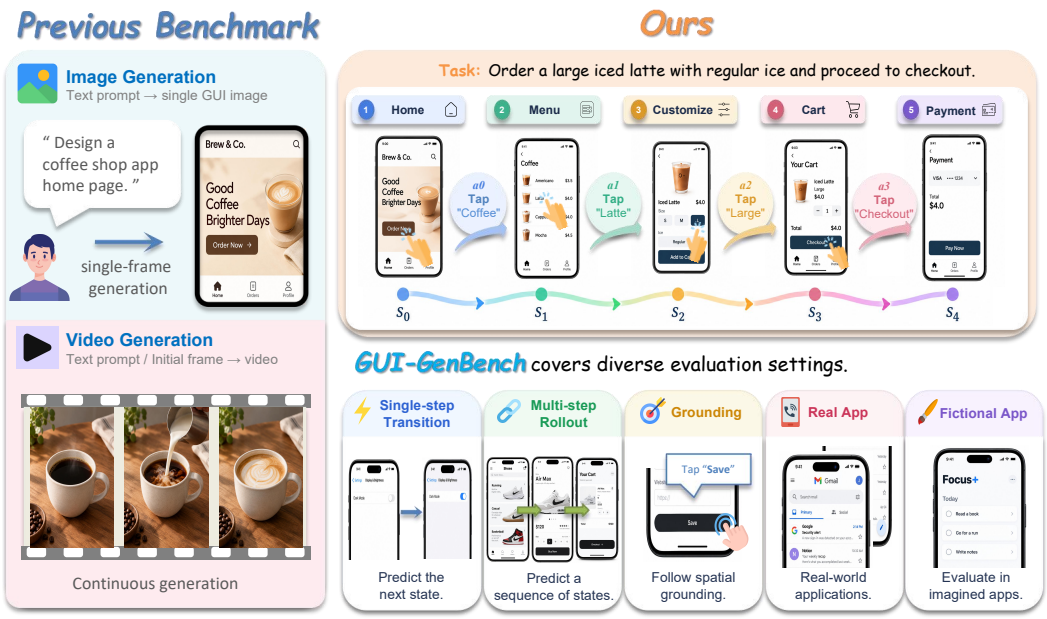


Figure 1: **GUI-GenBench overview.** Image- and video-generation benchmarks assess visual content and temporal behavior. GUI-GenBench focuses on whether generated GUI states respond correctly to actions and remain coherent across a trajectory. Its five task suites cover single-step transitions, multi-step planning, grounding, real-app trajectories, and fictional-app generation.

logic from content consistency, UI plausibility, and visual quality. This distinction makes it possible to identify outputs that look convincing but fail to realize the requested action.

Our evaluation covers 12 image generation models (OpenAI, 2025; Google, 2025b; Seedream et al., 2025; Seed, 2025; Labs, 2025; Wan et al., 2025; Wu et al., 2025; Deng et al., 2025; Li et al., 2025b; Team et al., 2025a). Leading models score above 80 on single-step transitions, yet most models obtain lower scores on multi-step planning. Grounding remains difficult even when the generated screen is visually plausible. Qualitative examples reveal errors in Chinese text rendering, icon interpretation, and action-target localization. These findings motivate evaluating functional reliability separately from visual fidelity, rather than treating realistic appearance as evidence of correct interaction.

Our primary contributions are as follows:

1. We introduce **GUI-GenBench**, with 700 samples across five task suites for evaluating generated GUI states and trajectories.
2. We introduce **GUI-Score**, a five-dimensional rubric-based metric that separates functional correctness, consistency, and visual fidelity.
3. We evaluate 12 models on Chinese and English subsets, identifying limitations in multi-step coherence and grounding, with qualitative diagnoses of text, icon, and coordinate-related errors.

2 Related Work

2.1 Automated GUI Generation

Work on automated interface construction spans layout design, screenshot-to-code translation, and generative prototyping (Chen et al., 2018; Sandhaus et al., 2011; Yang et al., 2016; Li et al., 2020; Zhao et al., 2021; Mozaffari et al., 2022; Sobolevsky et al., 2023; Kolthoff et al., 2025; 2024). Early approaches use explicit layout principles or reconstruct interfaces from mockups (Sandhaus et al., 2011; Huang et al., 2016); learned approaches map screenshots to implementation structure (Chen et al., 2018) or generate layouts and prototypes (Sobolevsky et al., 2023; Kolthoff et al., 2025).

Generative layout and design methods broaden the space of candidate interfaces (Li et al., 2020; Zhao et al., 2021; Mozaffari et al., 2022), while spatial reasoning and visual representation remain relevant challenges for multimodal models (Zhang et al., 2025b;c). GUI-GenBench focuses on a different evaluation target: whether a generated interface correctly changes in response to an instruction.

2.2 Image Generation and Editing Models

Recent multimodal systems support image synthesis and reference-based editing (Hurst et al., 2024; Comanici et al., 2025; Google, 2025b; Seedream et al., 2025; Wan et al., 2025; Deng et al., 2025; Liu et al., 2025a; Li et al., 2025b; Team et al., 2025b;a). Diffusion and autoregressive approaches have advanced visual generation (Ho et al., 2020; Chen et al., 2020; Labs, 2025; Fan et al., 2024; Han et al., 2025; Ramesh et al., 2022), alongside progress in visual understanding (Lin et al., 2025b). Reference-conditioned methods combine visual context with textual instructions to control generated content (Google, 2025b; Seedream et al., 2025; Wan et al., 2025; Team et al., 2025a; An et al., 2025), including preservation of appearance and structure during editing (Deng et al., 2025; Liu et al., 2025a). These capabilities make GUI state generation feasible, but do not by themselves establish that the outputs obey interface logic.

2.3 Visual and Sequential Generation Benchmarks

Image-generation evaluation considers distributional fidelity, text–image alignment, and compositional correctness (Heusel et al., 2017; Ghosh et al., 2023; Huang et al., 2023; Radford et al., 2021). Related work examines knowledge-informed generation, semantic alignment, and instruction-driven editing (Niu et al., 2025; Hu et al., 2024; Zhao et al., 2025; Huang et al., 2023; Zhang et al., 2026; Liu et al., 2025a), while sequential benchmarks study story consistency and visual reasoning (Zhuang et al., 2025; Guo et al., 2025). GUI trajectories introduce a distinct requirement: actions must induce semantically correct changes even when the entire layout is replaced (Zhang et al., 2025a; Yan et al., 2025). Text-rendering errors can also change the functional meaning of a screen (Chen et al., 2024). GUI-GenBench therefore evaluates the semantic and structural correctness of transitions, rather than visual continuity alone.

2.4 Relation to Generative GUI Environment Work

A related line of work constructs synthetic GUI data for agent training. UI-Oceanus (Wu et al., 2026) uses forward-dynamics prediction as a continual-pretraining objective, generating future GUI states from current observations and actions. GUI-ReWalk (Lin et al., 2025a) combines stochastic exploration with intent-aware reasoning to construct interaction trajectories for downstream agents. GUI-GenBench complements these efforts by evaluating the generator rather than downstream agent success. Its five scoring dimensions provide a diagnostic of generated states that can inform synthetic-data pipelines. We do not measure whether training on these outputs improves agent performance.

3 GUI-GenBench

A GUI environment maps user or agent actions to interface states according to application logic (Zhang et al., 2025a). GUI-GenBench evaluates an image-based approximation of this process: a model receives an instruction and, when required by the task, a reference screen, then generates the resulting screen or trajectory. Throughout the paper, "GUI state" refers to the rendered image, not an executable application state. The benchmark tests whether these images depict valid responses; it does not verify execution in the underlying application.

3.1 Benchmark Design and Task Suites

GUI-GenBench comprises 700 samples curated with consistency and fidelity checks. The five task suites in Figure 2 probe different requirements of GUI generation:

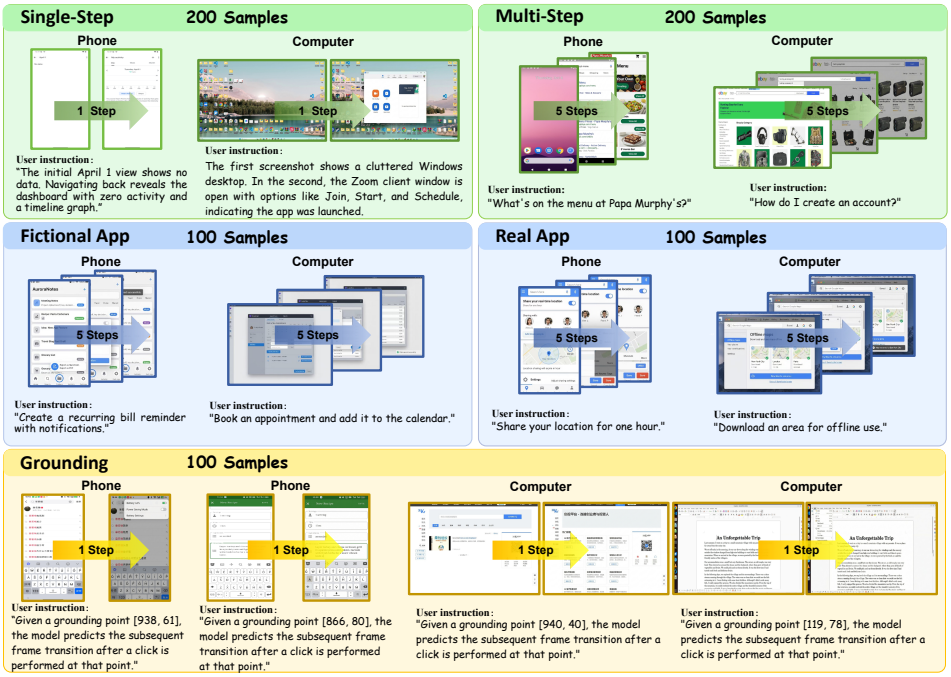


Figure 2: **Examples from the five GUI-GenBench task suites.** Models generate GUI states or trajectories from task instructions and, where specified, a reference screen. The fictional-app task requires generation without a reference image. Evaluation considers both the requested interaction and the resulting visual content.

1. **Single-step Visual Transition (single-step).** Given an initial GUI state and a detailed action, the model generates the subsequent state. This task evaluates fine-grained instruction following.
2. **Multi-step Planning (multi-step).** Given an initial GUI state and a high-level objective such as "Order a coffee," the model generates a five-step trajectory. This task evaluates goal-directed state progression, temporal coherence, and consistency of interface structure.
3. **Zero-shot Virtual GUI (fictional-app).** The model generates a fictional interface from detailed instructions without a reference image. This task probes generation beyond familiar application layouts; it does not establish that the requested layout is absent from training data.
4. **Rare Trajectory Synthesis (real-app).** The model generates long-tail interaction trajectories in data-scarce scenarios. This task tests whether generated states follow the requested interaction logic in less familiar settings.
5. **Grounding-based Generation (grounding).** Given an initial screen and an action target specified by normalized coordinates in $[0, 1000]$, the model generates the next GUI state. This task tests whether the output reflects an interaction with the intended control.

We use the parenthetical task names throughout the results and analysis.

3.2 Evaluation Dimensions and GUI-Score

The five task suites specify *what* a model must generate; five scoring dimensions assess *how well* it does so. Task-specific rubrics evaluate functional correctness, consistency, and visual fidelity:

- **Goal Achievement (GOAL)** assesses whether the generated state realizes the requested action or objective, including the correctness and completeness of the outcome.
- **Interaction Logic (LOGIC)** assesses whether state changes are consistent with valid actions and plausible interface behavior.

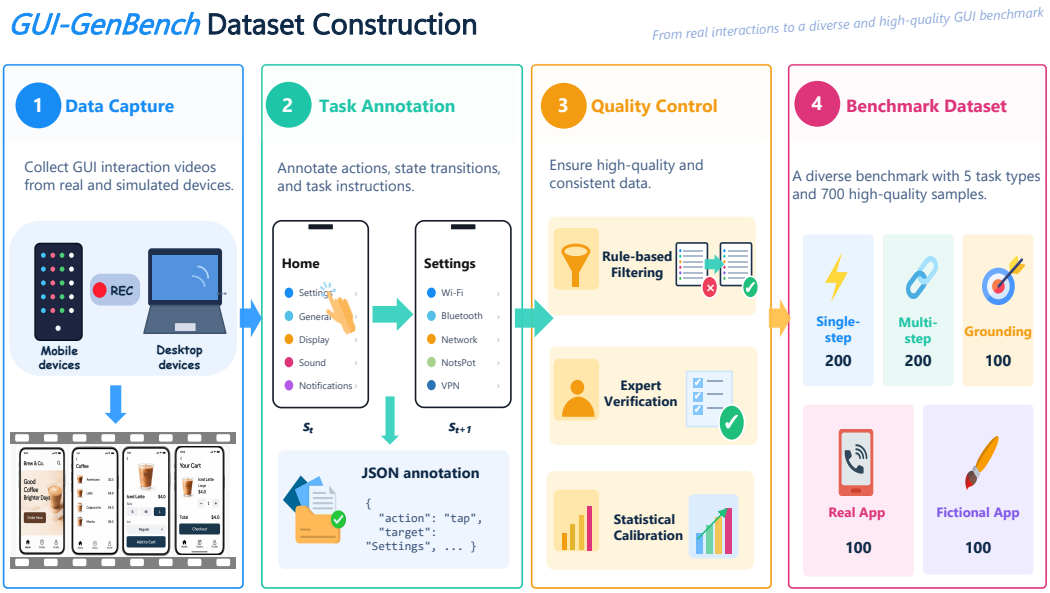


Figure 3: **GUI-GenBench data construction pipeline.** Interaction capture and task annotation are followed by preprocessing, human verification, and distribution calibration. The resulting 700 samples span the single-step, multi-step, grounding, real-app, and fictional-app task suites.

- **Content Consistency (CONS)** assesses preservation of content that should remain unchanged and stability of interface elements across a trajectory.
- **UI Plausibility (UI)** assesses structural coherence and conformity to interface conventions, including the absence of fabricated or incompatible controls.
- **Visual Quality (QUAL)** assesses text readability, icon clarity, and freedom from rendering artifacts.

Each dimension uses an ordinal scale from 0 to 5, with higher scores indicating better performance. Appendix F provides the task-specific criteria. For an evaluation set of N samples, let $\mathcal{D}$ denote the five dimensions and $r_{i,d} \in \{0, \ldots, 5\}$ the rating of sample i on dimension d. GUI-Score averages the normalized ratings:

$$\text{GUI-Score} = \frac{1}{|\mathcal{D}| \cdot N} \sum_{d \in \mathcal{D}} \sum_{i=1}^{N} (\mathcal{F}(r_{i,d})) = \frac{4}{N} \sum_{d=1}^{5} \sum_{i=1}^{N} r_{i,d} \tag{1}$$

where $\mathcal{F}(r) = 20 \times r$ maps each rating to $[0, 100]$. These values are normalized rubric scores, not percentages of successful interactions. We report dimension-level scores alongside the aggregate because visual quality can remain high when goal achievement is low. Appendix A compares equal weighting with a function-priority alternative; the rankings are similar under the two tested schemes.

3.3 Data Construction Pipeline

Figure 3 summarizes the construction pipeline. We collect screen recordings of mobile and desktop interactions. Annotators identify actions, such as clicking icons or scrolling, and represent the corresponding sequences as structured JSON metadata.

Quality control proceeds in three stages. Rule-based preprocessing filters inconsistent or noisy samples; human experts then check that annotated actions match the observed transitions. A final distribution-calibration step adjusts the sample mix, yielding 700 samples across the five task suites in Section 3.1. The Chinese and English subsets are sampled independently, not translated from a shared set of tasks. Screenshots for the real-app suite come from publicly released datasets with

Table 1: **Main evaluation results on GUI-GenBench across Chinese and English subsets.** Scores are reported for five task suites and the aggregate GUI-Score. Within each language subset, **bold** and underlined values indicate the best and second-best results in each column across all 12 models, respectively.

Model	Chinese subset						English subset					
	Single-step	Multi-step	Fictional-app	Real-app	Grounding	**GUI-Score**	Single-step	Multi-step	Fictional-app	Real-app	Grounding	**GUI-Score**
Commercial models												
Nano Banana pro	**84.50**	**68.65**	**65.75**	64.35	**64.83**	**69.62**	**84.32**	**69.51**	46.33	47.20	**58.64**	61.20
Nano Banana	64.36	34.16	64.82	**65.89**	54.48	56.74	64.80	50.75	48.88	47.12	49.04	52.12
GPT-image-1.5	83.79	56.97	60.11	55.65	53.33	63.22	80.80	58.87	**63.68**	**58.93**	49.23	**63.16**
GPT-image-1.0	64.72	49.20	57.31	59.04	31.68	52.39	60.92	64.33	58.94	56.16	37.84	55.64
Seedream 4.5	63.64	53.11	56.48	53.44	52.90	55.91	49.49	45.30	53.81	51.80	49.63	50.01
Seedream 4.0	62.04	48.64	49.28	50.93	53.53	52.88	53.28	37.57	47.92	49.36	44.17	46.46
Wan 2.6	64.20	50.11	52.72	50.40	59.58	55.40	60.17	44.36	49.55	44.80	53.36	50.45
Flux-2-pro	68.83	55.07	58.13	55.41	50.24	57.54	61.00	52.17	49.92	47.16	45.67	51.18
Open-source models												
Bagel	34.84	13.45	27.36	33.52	35.10	28.85	32.91	8.61	26.08	35.12	37.30	28.00
UniWorld-V2	55.33	24.95	32.03	21.39	49.60	36.66	42.68	14.14	30.08	26.83	47.04	32.15
Qwen-Image-Edit	41.12	26.79	23.78	26.10	50.80	33.72	40.12	18.61	25.80	25.95	54.55	33.01
Longcat-Image	48.76	12.75	30.03	30.00	51.02	34.51	36.69	8.44	37.30	36.83	47.12	33.28

explicit licenses. Rare Trajectory Synthesis additionally uses fictional applications or uncommon scenarios expressed as text prompts, with the aim of reducing reliance on familiar screenshots. This design does not rule out overlap with model training data; Appendix C discusses provenance and coverage limitations.

4 Evaluation

4.1 Evaluation Setup

Evaluated models. We evaluate 12 image generation models: 8 commercial and 4 open-source models.

The commercial models are Nano Banana Pro (Google, 2025b), Nano Banana (Google, 2025b), GPT-Image-1.5 (OpenAI, 2025), GPT-Image-1 (OpenAI, 2025), Seedream 4.5 (Seed, 2025), Seedream 4.0 (Seedream et al., 2025), Wan 2.6 (Wan et al., 2025), and Flux-2-Pro (Labs, 2025).

The open-source models are Bagel (Deng et al., 2025), UniWorld-V2 (Li et al., 2025b), Qwen-Image-Edit (Wu et al., 2025), and LongCat-Image (Team et al., 2025a).

VLM-based judges. We evaluate generated outputs with three VLMs: Gemini-3-Pro-Native (Google, 2025a), GPT-4o (Hurst et al., 2024), and Qwen3-VL-235B-A22B-Thinking (Bai et al., 2025). We use official default configurations and evaluate each generated trajectory three times. Judge-specific results in Appendix E expose variation across evaluators; using multiple judges does not eliminate shared evaluation biases.

Multi-step generation protocol. Given the initial screen and a target instruction, the model generates the subsequent five frames in one pass. No intermediate ground-truth frames are supplied. The task therefore evaluates a jointly generated trajectory rather than a sequence of edits conditioned on correct preceding states. It also differs from a closed-loop rollout in which an external agent selects a new action after each generated frame.

4.2 Evaluation Results

Overall performance. Table 1 reports results separately for the Chinese and English subsets. Nano Banana Pro (Google, 2025b) leads on the Chinese subset with a GUI-Score of 69.62, whereas

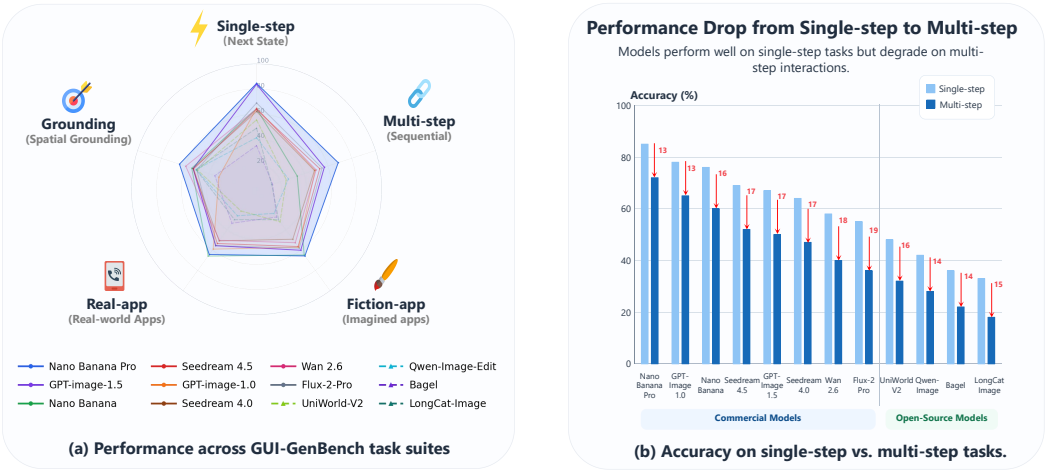


Figure 4: **Benchmark performance and single-step versus multi-step comparison.** Left: performance across the five task suites on the Chinese subset, with commercial models shown by solid lines and open-source models by dashed lines. Right: single-step and multi-step performance, with red arrows indicating the differences.

GPT-Image-1.5 (OpenAI, 2025) leads on the English subset with 63.16. Commercial models generally outperform the evaluated open-source models, including UniWorld-V2 (Li et al., 2025b) and Bagel (Deng et al., 2025), although the size of the gap varies by task. Figure 4 summarizes the task-level profiles. Because the two language subsets contain independently sampled tasks, their score differences do not isolate the effect of language.

Single-step versus multi-step generation. Strong single-step scores do not consistently transfer to multi-step planning. Nano Banana Pro (Google, 2025b) and GPT-Image-1.5 (OpenAI, 2025) both exceed 80 points on single-step tasks in each language subset, but receive lower multi-step scores. Across models, multi-step scores are often below 60 and, for some open-source models, below 10. The trend is not universal: GPT-Image-1 scores 64.33 on English multi-step tasks versus 60.92 on single-step tasks.

Multi-step generation requires both plausible individual screens and coherent dependencies between them. A trajectory can therefore fail despite visually convincing frames. However, the task suites differ in more than trajectory length, and all five frames are generated in one pass. The observed score gap should not be interpreted as a controlled estimate of error accumulation or as direct evidence about the models' internal reasoning mechanisms.

Coordinate grounding. Dimension-level scores reveal a weakness that aggregate scores can obscure. Under Gemini-3-Pro-Native, Nano Banana Pro (Google, 2025b) obtains the highest grounding GOAL score, approximately 23.9 on the normalized 0–100 scale, while all four open-source models, including Qwen-Image-Edit (Wu et al., 2025), score below 10 (Appendix Table A1). These are rubric scores, not action-success rates. By contrast, consistency and visual-quality scores can remain much higher, indicating that plausible appearance does not ensure a correct response to the coordinate-specified action. Appendix Figure A1(b) reports the grounding GOAL comparison, and Figure 5 illustrates an incorrect transition despite a recognizable interface.

4.3 Human–VLM Score Correlation

We compare VLM scores with human judgments for five generators spanning a broad performance range: Nano Banana Pro (Google, 2025b), GPT-Image-1 (OpenAI, 2025), Qwen-Image-Edit (Wu et al., 2025), Bagel (Deng et al., 2025), and LongCat-Image (Team et al., 2025a). For each generator, we select 10 samples from each of the 10 language–task combinations, giving 100 samples per generator. Four human experts independently apply the same scoring criteria as the VLM judges. We average their ratings before computing Pearson correlation with the VLM scores.

The initial comparison for Nano Banana Pro (Google, 2025b) and GPT-Image-1 (OpenAI, 2025) yields $r = 0.9926$ and $r = 0.9833$, respectively (Appendix Figure A1(a)). Extending the comparison to lower-scoring generators gives $r = 0.9647$ for Qwen-Image-Edit, $r = 0.9613$ for Bagel, and $r = 0.9577$ for LongCat-Image. The mean across the five correlations is 0.9719, with a minimum of 0.9577. Thus, high correlation is not confined to the two strongest generators in this study. Correlation measures association, however, not absolute score agreement, and does not establish that judges detect every interface error. Appendix B gives the calculation and limitations.

4.4 Sensitivity to Scoring Choices

Dimension weights. Equal weighting treats functional and visual dimensions symmetrically. To examine whether emphasizing function changes the model comparison, we use the alternative weights reported in Appendix A: 30% for GOAL, 30% for LOGIC, 20% for CONS, and 10% each for UI and QUAL. Under Gemini-3-Pro-Native, the resulting ranking of the 12 models has Spearman correlation $\rho = 0.979$ ($p < 10^{-7}$) with the equal-weight ranking. The eight leading models retain their order, and only adjacent positions change. Thus, the leading-model comparison is unchanged under this particular increase in functional emphasis. The result does not imply invariance to other weight choices or judges.

Judge variation. The detailed tables also show why judge identity matters. For Nano Banana Pro, the multi-step Overall score is 69.07 under Gemini-3-Pro-Native, 87.68 under GPT-4o, and 91.71 under Qwen3-VL-235B-A22B-Thinking (Appendix Tables A1–A3). Mapping ratings to a common 0–100 range does not remove these differences. We therefore retain the judge-specific results alongside the task-level summary, so that readers can distinguish a generator comparison under a given evaluator from changes attributable to the evaluator. Correlation with human ratings and sensitivity to dimension weights address different aspects of evaluation; neither replaces inspection of judge-dependent scores.

5 Discussion and Analysis

5.1 From Plausible Screens to Reliable Transitions

Table 1 and Figure 4 suggest that local transition quality is an incomplete measure of a generative GUI environment. Leading models (Google, 2025b; OpenAI, 2025) exceed 80 points on single-step tasks, yet show weaker performance on multi-step trajectories. The grounding results expose a related distinction between preserving appearance and executing the intended action. Together, these observations motivate inspecting functional dimensions alongside aggregate scores. They do not establish a universal ordering of task difficulty or show that a model relies on memorization rather than interaction dynamics.

GUI-GenBench evaluates generated visual states, not executable software or downstream agent learning. Its diagnostic value is to reveal where a plausible screen ceases to be a faithful response. Whether filtering synthetic trajectories using these diagnostics improves agent training remains an open empirical question.

5.2 Qualitative Analysis of Failure Cases

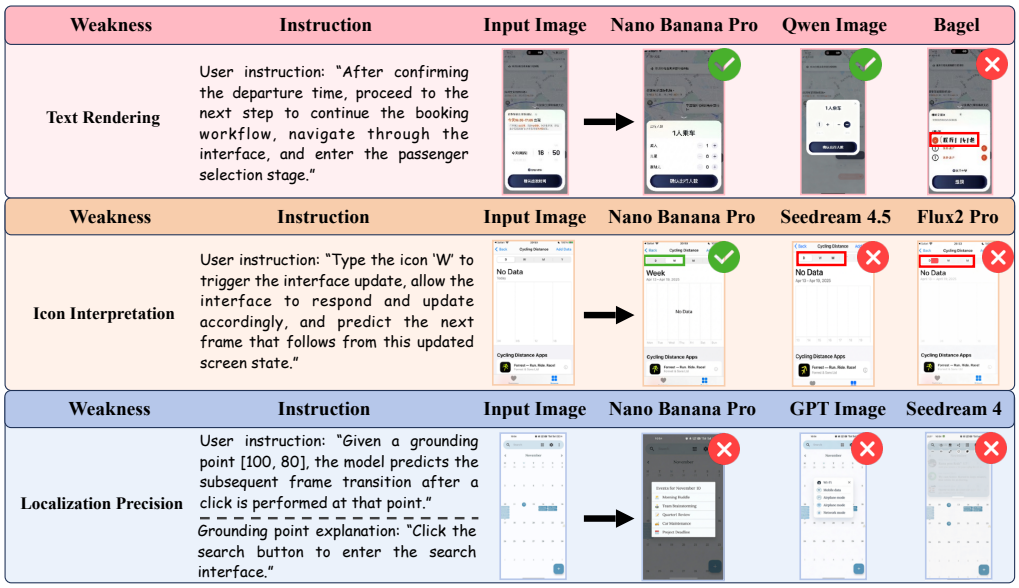


Figure 5: **Representative failure patterns in generated GUI states.** Top: text rendering in a passenger-selection interface. Middle: interpreting the "W" control to switch to a weekly view. Bottom: executing a coordinate-based instruction targeting the search control. Green checks and red crosses indicate successful and failed responses in the shown examples.

Figure 5 pairs instructions and reference screens with generated outputs to illustrate three failure modes. The examples identify specific errors; they do not estimate how frequently those errors occur across the benchmark.

Text rendering (top row). The booking task requires a transition to passenger selection while preserving readable text. Nano Banana Pro (Google, 2025b) and the Qwen Image output are marked as successful; Bagel produces corrupted characters in the highlighted region. A recognizable layout therefore does not guarantee readable content, and the failure is not shared by every model in this example.

Icon interpretation (middle row). The instruction targets the "W" control on a cycling-distance screen. Nano Banana Pro generates the weekly view, whereas the highlighted Seedream 4.5 and Flux-2-Pro outputs fail to realize that transition. The required change depends on the control's functional meaning, not merely its visual appearance.

Coordinate grounding (bottom row). The point $[100, 80]$ targets the search control, but the generated outputs show unrelated menus or overlays. This mismatch between the requested target and the resulting state complements the low grounding GOAL scores in Section 4. A visually plausible screen remains functionally incorrect when it depicts a different interaction.

AI Use Statement

Generative AI tools assisted with migration to the ICLR 2027 LaTeX template, manuscript organization and language editing, figure-layout drafts, and refinement of the written interpretation of existing results. VLMs also serve as evaluators in the benchmark, as described in Section 4. The authors are responsible for verifying the AI-assisted material and for the final text, claims, and artifacts.

REFERENCES

Ruichuan An, Sihan Yang, Renrui Zhang, Zijun Shen, Ming Lu, Gaole Dai, Hao Liang, Ziyu Guo, Shilin Yan, Yulin Luo, et al. Unictokens: Boosting personalized understanding and generation via unified concept tokens. *arXiv preprint arXiv:2505.14671*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Rogerio Bonatti, Dan Zhao, Francesco Bonacci, Dillon Dupont, Sara Abdali, Yinheng Li, Yadong Lu, Justin Wagle, Kazuhito Koishida, Arthur Bucker, et al. Windows agent arena: Evaluating multi-modal os agents at scale. *arXiv preprint arXiv:2409.08264*, 2024.

Chunyang Chen, Ting Su, Guozhu Meng, Zhenchang Xing, and Yang Liu. From ui design image to gui skeleton: a neural machine translator to bootstrap mobile gui implementation. In *Proceedings of the 40th International Conference on Software Engineering*, pp. 665–676, 2018.

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

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pp. 1691–1703. PMLR, 2020.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Li YanTao, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents. In *ACL 2024 (Volume 1: Long Papers)*, pp. 9313–9332, 2024.

Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *International conference on machine learning*, pp. 2048–2056. PMLR, 2020.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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

Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. *arXiv preprint arXiv:2410.13863*, 2024.

Aryan Garg, Yue Jiang, and Antti Oulasvirta. Controllable gui exploration. *arXiv preprint arXiv:2502.03330*, 2025.

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

Google. A new era of intelligence with Gemini 3, 2025a. URL `https://blog.google/products/gemini/gemini-3`.

Google. Introducing nano banana pro, 2025b. URL https://blog.google/innovation-and-ai/products/nano-banana-pro/.

Ziyu Guo, Xinyan Chen, Renrui Zhang, Ruichuan An, Yu Qi, Dongzhi Jiang, Xiangtai Li, Manyuan Zhang, Hongsheng Li, and Pheng-Ann Heng. Are video models ready as zero-shot reasoners? an empirical study with the mme-cof benchmark. *arXiv preprint arXiv:2510.26802*, 2025.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15733–15744, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

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

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

Ruozi Huang, Yonghao Long, and Xiangping Chen. Automaticly generating web page from a mockup. In *SEKE*, pp. 589–594, 2016.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Kristian Kolthoff, Felix Kretzer, Lennart Fiebig, Christian Bartelt, Alexander Maedche, and Simone Paolo Ponzetto. Zero-shot prompting approaches for llm-based graphical user interface generation. *arXiv preprint arXiv:2412.11328*, 2024.

Kristian Kolthoff, Felix Kretzer, Christian Bartelt, Alexander Maedche, and Simone Paolo Ponzetto. Guide: Llm-driven gui generation decomposition for automated prototyping. In *2025 IEEE/ACM 47th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, pp. 1–4. IEEE, 2025.

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

Jianan Li, Jimei Yang, Aaron Hertzmann, Jianming Zhang, and Tingfa Xu. Layoutgan: Synthesizing graphic layouts with vector-wireframe adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2388–2399, 2020.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Mobileworldbench: Towards semantic world modeling for mobile agents. *arXiv preprint arXiv:2512.14014*, 2025a.

Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Feize Wu, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, et al. Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. *arXiv preprint arXiv:2510.16888*, 2025b.

Musen Lin, Minghao Liu, Taoran Lu, Lichen Yuan, Yiwei Liu, Haonan Xu, Yu Miao, Yuhao Chao, and Zhaojian Li. GUI-ReWalk: Massive data generation for GUI agent via stochastic exploration and intent-aware reasoning. *arXiv preprint arXiv:2509.15738*, 2025a.

Weifeng Lin, Xinyu Wei, Ruichuan An, Tianhe Ren, Tingwei Chen, Renrui Zhang, Ziyu Guo, Wentao Zhang, Lei Zhang, and Hongsheng Li. Perceive anything: Recognize, explain, caption, and segment anything in images and videos, 2025b. URL https://arxiv.org/abs/2506.05302.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025a.

Xinyi Liu, Xiaoyi Zhang, Ziyun Zhang, and Yan Lu. Ui-e2i-synth: Advancing gui grounding with large-scale instruction synthesis. *arXiv preprint arXiv:2504.11257*, 2025b.

Dezhao Luo, Bohan Tang, Kang Li, Georgios Papoudakis, Jifei Song, Shaogang Gong, Jianye Hao, Jun Wang, and Kun Shao. Vimo: A generative visual gui world model for app agents. *arXiv preprint arXiv:2504.13936*, 2025.

Mohammad Amin Mozaffari, Xinyuan Zhang, Jinghui Cheng, and Jin LC Guo. Ganspiration: balancing targeted and serendipitous inspiration in user interface design with style-based generative adversarial network. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pp. 1–15, 2022.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

OpenAI. The new chatgpt images is here, 2025. URL https://openai.com/index/new-chatgpt-images-is-here/.

Alec Radford et al. Learning transferable visual models from natural language supervision. *ICML*, 2021.

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

Philipp Sandhaus, Mohammad Rabbath, and Susanne Boll. Employing aesthetic principles for automatic photo book layout. In *International Conference on Multimedia Modeling*, pp. 84–95. Springer, 2011.

ByteDance Seed. Seedream 4.5, 2025. URL https://seed.bytedance.com/en/seedream4_5.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Andrey Sobolevsky, Guillaume-Alexandre Bilodeau, Jinghui Cheng, and Jin LC Guo. Guilget: Gui layout generation with transformer. *arXiv preprint arXiv:2304.09012*, 2023.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8406–8416, 2025.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, et al. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025a.

NextStep Team, Chunrui Han, Guopeng Li, Jingwei Wu, Quan Sun, Yan Cai, Yuang Peng, Zheng Ge, Deyu Zhou, Haomiao Tang, et al. Nextstep-1: Toward autoregressive image generation with continuous tokens at scale. *arXiv preprint arXiv:2508.10711*, 2025b.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng

Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jialiang Wei, Anne-Lise Courbis, Thomas Lambolais, Binbin Xu, Pierre Louis Bernard, and Gérard Dray. Boosting gui prototyping with diffusion models. In *2023 IEEE 31st International Requirements Engineering Conference (RE)*, pp. 275–280. IEEE, 2023.

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

Mengzhou Wu, Yuzhe Guo, Yuan Cao, Haochuan Lu, Songhe Zhu, Pingzhe Qu, Xin Chen, Kang Qin, Zhongpu Wang, Xiaode Zhang, Xinyi Wang, Wei Dai, Gang Cao, Yuetang Deng, Zhi Gong, Dezhi Ran, Linyi Li, Wei Yang, and Tao Xie. UI-Oceanus: Scaling GUI agents with synthetic environmental dynamics. *arXiv preprint arXiv:2604.02345*, 2026.

Bin Xie, Rui Shao, Gongwei Chen, Kaiwen Zhou, Yinchuan Li, Jie Liu, Min Zhang, and Liqiang Nie. Gui-explorer: Autonomous exploration and mining of transition-aware knowledge for GUI agent. In *ACL 2025 (Volume 1: Long Papers)*, pp. 5650–5667. Association for Computational Linguistics, 2025a.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.

Yuquan Xie, Zaijing Li, Rui Shao, Gongwei Chen, Kaiwen Zhou, Yinchuan Li, Dongmei Jiang, and Liqiang Nie. Mirage-1: Augmenting and updating gui agent with hierarchical multimodal skills. *arXiv preprint arXiv:2506.10387*, 2025b.

Haolong Yan, Yeqing Shen, Xin Huang, Jia Wang, Kaijun Tan, Zhixuan Liang, Hongxin Li, Zheng Ge, Osamu Yoshie, Si Li, et al. Gui exploration lab: Enhancing screen navigation in agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2512.02423*, 2025.

Xuyong Yang, Tao Mei, Ying-Qing Xu, Yong Rui, and Shipeng Li. Automatic generation of visual-textual presentation layout. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 12(2):1–22, 2016.

Chaoyun Zhang, Shilin He, Jiaxu Qian, Bowen Li, Liqun Li, Si Qin, Yu Kang, Minghua Ma, Guyue Liu, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Zhang. Large language model-brained GUI agents: A survey. *Trans. Mach. Learn. Res.*, 2025, 2025a.

Huanyu Zhang, Chengzu Li, Wenshan Wu, Shaoguang Mao, Yifan Zhang, Haochen Tian, Ivan Vulić, Zhang Zhang, Liang Wang, Tieniu Tan, et al. Scaling and beyond: Advancing spatial reasoning in mllms requires new recipes. *arXiv preprint arXiv:2504.15037*, 2025b.

Huanyu Zhang, Wenshan Wu, Chengzu Li, Ning Shang, Yan Xia, Yangyu Huang, Yifan Zhang, Li Dong, Zhang Zhang, Liang Wang, et al. Latent sketchpad: Sketching visual thoughts to elicit multimodal reasoning in mllms. *arXiv preprint arXiv:2510.24514*, 2025c.

Huanyu Zhang, Xuehai Bai, Chengzu Li, Chen Liang, Haochen Tian, Haodong Li, Ruichuan An, Yifan Zhang, Anna Korhonen, Zhang Zhang, Liang Wang, and Tieniu Tan. How well do models follow visual instructions? vibe: A systematic benchmark for visual instruction-driven image editing, 2026. URL https://arxiv.org/abs/2602.01851.

Tianming Zhao, Chunyang Chen, Yuanning Liu, and Xiaodong Zhu. Guigan: Learning to generate gui designs using generative adversarial networks. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pp. 748–760. IEEE, 2021.

Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, et al. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. *arXiv preprint arXiv:2504.02826*, 2025.

Cailin Zhuang, Ailin Huang, Yaoqi Hu, Jingwei Wu, Wei Cheng, Jiaqi Liao, Hongyuan Wang, Xinyao Liao, Weiwei Cai, Hengyuan Xu, et al. Vistorybench: Comprehensive benchmark suite for story visualization. *arXiv preprint arXiv:2505.24862*, 2025.

APPENDIX

A GUI-SCORE WEIGHTING SENSITIVITY

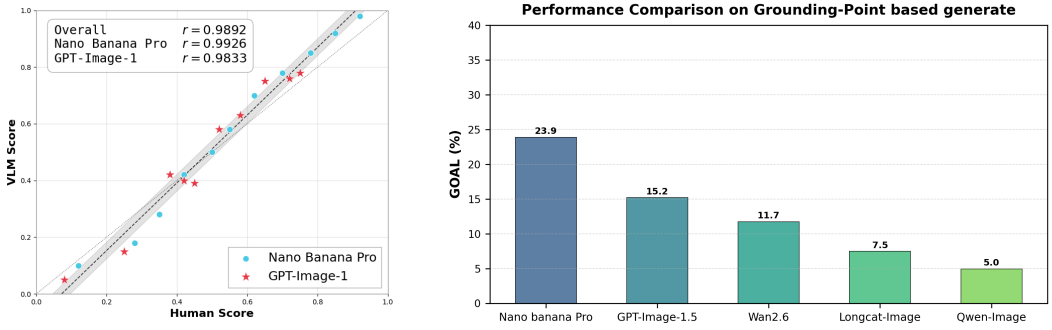
GUI-Score gives equal weight to Goal Achievement, Interaction Logic, Content Consistency, UI Plausibility, and Visual Quality. To examine sensitivity to this choice, we compare it with a function-priority scheme using weights of 30%, 30%, 20%, 10%, and 10%, respectively. We re-rank all 12 models using the Gemini-3-Pro-Native scores under each scheme. The rankings have a Spearman correlation of $\rho = 0.979$ ($p < 10^{-7}$); only adjacent positions exchange order, and the eight leading models retain their ordering. This comparison supports ranking stability for the two tested schemes under this judge, not robustness to arbitrary weights. Dimension-level scores remain important when functional correctness is the primary concern.

B EXTENDED HUMAN VALIDATION

The human study uses 100 samples per generator, with 10 drawn from each language–task combination and scored independently by four experts. For each generator, let h_i denote a human score averaged across the four experts and v_i the corresponding VLM score; $\bar{h}$ and $\bar{v}$ denote their respective means. We quantify their linear association using Pearson correlation:

$$r = \frac{\sum_i (h_i - \bar{h})(v_i - \bar{v})}{\sqrt{\sum_i (h_i - \bar{h})^2} \sqrt{\sum_i (v_i - \bar{v})^2}}. \tag{2}$$

The correlations are 0.9926 for Nano Banana Pro, 0.9833 for GPT-Image-1, 0.9647 for Qwen-Image-Edit, 0.9613 for Bagel, and 0.9577 for LongCat-Image, with a mean of 0.9719. These results show a strong linear association in the reported comparisons across generators of different quality. They do not establish absolute score calibration, agreement among human annotators, or detection of every fine-grained interface error. Inter-rater agreement and unanimous-case frequency would provide complementary evidence from the per-annotator labels. The reported correlations should therefore be interpreted as one check on the scoring framework, not a complete validation of judge reliability.



(a) Human–VLM score correlation.

(b) GOAL scores on the grounding task.

Figure A1: **Human–VLM correlation and grounding performance.** Left: the initial score-correlation comparison for Nano Banana Pro (Google, 2025b) and GPT-Image-1 (OpenAI, 2025); the text reports the extended five-model results. Right: normalized grounding GOAL scores, which assess whether generated states reflect the coordinate-specified action rather than simply preserving visual appearance.

C BENCHMARK SCOPE, DATA PROVENANCE, AND LIMITATIONS

Data provenance and language subsets. The Chinese and English subsets contain independently sampled tasks rather than paired translations. Cross-language score differences can therefore reflect differences in task composition as well as language. Screenshots in the real-app suite come from

publicly released datasets with explicit licenses; any benchmark release should include a per-sample source and license manifest. Rare Trajectory Synthesis also uses fictional applications and uncommon scenarios described through text prompts. These scenarios are intended to reduce dependence on familiar screenshots, but we cannot exclude training-data overlap for the evaluated generators.

Coverage and temporal validity. The 700 samples do not exhaust the diversity of applications, platforms, languages, or actions, and do not cover all emerging interfaces, including augmented- and virtual-reality systems. The sampled application and interaction distributions may introduce domain bias. Interface redesigns and software updates can also make captured states outdated. Results thus describe a benchmark snapshot rather than all GUI interactions. Future releases should record capture time, application version where available, source, platform, language, action type, and license.

Evaluation limitations. High human–VLM correlation does not exclude shared judge biases or missed interface errors. Normalizing all judges to the same numerical range also does not make their scores equally calibrated. We do not use CLIP-family similarity as a primary metric: small text, icon, or button-state errors can matter functionally despite high global similarity. Localization overlap metrics such as IoU are relevant where spatial annotations exist, but do not capture whole-interface correctness. Such metrics and human–human agreement would provide complementary evidence. Finally, our single-shot trajectory protocol does not test closed-loop agent interaction or the effect of generated data on downstream training.

Safety implications. Trajectories containing fabricated controls, incorrect transitions, or grounding errors should not be treated as faithful execution traces. Training agents on such outputs may reinforce invalid or unsafe behavior. GUI-GenBench provides a diagnostic of generated content, not a guarantee of deployment safety. Synthetic trajectories require filtering and independent validation before use in agent training or safety-critical interaction.

D Evaluation Framework

Figure A2 summarizes three stages. **Generation** produces a screen or trajectory from the task instruction and, where specified, a reference state. **Scoring** applies the task-specific rubrics to GOAL, LOGIC, CONS, UI, and QUAL using GPT-4o (Hurst et al., 2024), Gemini-3-Pro-Native (Google, 2025a), and Qwen3-VL-235B-A22B-Thinking (Bai et al., 2025). **Analysis** summarizes scores by task and dimension, compares the judges, and examines correlation with human ratings. Keeping these stages distinct separates generation failures from variation in how evaluators score them.

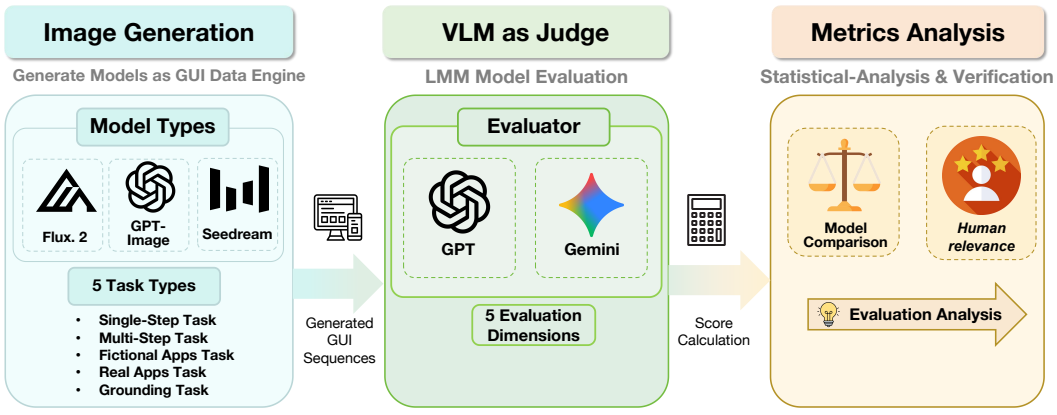


Figure A2: **GUI-GenBench evaluation framework.** Generated screens and trajectories are scored by three VLM judges using task-specific rubrics. Analysis summarizes task- and dimension-level scores and compares VLM scores with human ratings.

E Detailed Performance on GUI-GenBench Using Different Judges

The following tables report all 12 generators under each VLM judge, allowing model comparisons to be examined separately for each evaluator:

- Table A1: Detailed scores assigned by Gemini-3-Pro-Native (Google, 2025a) as the judge.
- Table A2: Detailed scores assigned by GPT-4o (Hurst et al., 2024) as the judge.
- Table A3: Detailed scores assigned by Qwen3-VL-235B-A22B-Thinking (Bai et al., 2025) as the judge.

For every task suite, each table reports GOAL, LOGIC, CONS, UI, QUAL, and an Overall score. Raw ratings are linearly mapped to $[0, 100]$ before aggregation. This common reporting range does not imply identical calibration across judges; the separate tables expose evaluator-dependent variation rather than treating any one judge as ground truth.

Table A1: Detailed GUI-Score results on GUI-GenBench with **Gemini-3-Pro-Native** (Google, 2025a) as the judge.

Task	Metric	Nano Banana Pro	Nano Banana	GPT-Image -1.5	GPT-Image -1.0	Seedream 4.5	Seedream 4.0	Wan 2.6	Flux-2-Pro	Bagel	UniWorld	Qwen -Image-Edit	Longcat -Image
single-step	GOAL	**95.04**	69.60	87.65	86.50	56.60	54.58	58.94	65.08	24.35	41.00	25.60	36.07
	LOGIC	82.04	58.50	**86.50**	78.80	49.25	45.29	52.17	53.60	29.12	43.33	27.93	35.47
	CONS	75.24	69.70	72.65	58.10	66.23	65.68	**81.81**	77.46	54.76	58.77	54.93	53.73
	UI	79.50	51.60	**84.76**	58.00	50.19	47.87	53.50	48.43	28.44	47.93	40.07	42.60
	QUAL	**90.22**	73.50	79.97	32.00	61.89	68.39	64.65	79.70	30.95	54.00	52.00	45.73
	Overall	**84.41**	64.58	82.31	62.68	56.83	56.36	62.21	64.85	33.52	49.01	40.11	42.72
multi-step	GOAL	**86.63**	47.61	79.13	77.45	68.25	60.14	46.89	67.18	3.85	2.29	1.98	1.78
	LOGIC	**56.75**	30.46	36.94	35.31	36.63	38.59	25.54	52.71	8.87	3.73	6.70	2.96
	CONS	**56.95**	45.69	36.39	34.08	38.66	38.70	56.43	46.46	15.46	31.86	43.32	19.89
	UI	**54.96**	26.80	50.93	50.41	35.15	33.44	37.56	29.79	12.10	20.17	29.17	16.08
	QUAL	**90.08**	62.03	86.23	85.82	66.80	48.04	64.88	63.18	13.54	40.41	31.60	10.41
	Overall	**69.07**	42.52	57.92	56.61	49.10	43.78	46.26	51.86	10.76	19.69	22.55	10.22
fictional-app	GOAL	40.51	40.40	**46.20**	38.78	40.00	39.33	36.49	45.93	15.80	4.00	0.40	17.01
	LOGIC	**51.03**	48.28	43.13	40.61	39.73	39.80	42.54	48.60	29.80	4.60	0.81	20.90
	CONS	57.95	60.00	53.20	53.88	61.00	64.68	61.17	50.67	41.20	**88.07**	62.36	85.00
	UI	50.26	45.45	**68.72**	61.63	42.60	39.40	41.24	39.00	27.60	28.03	27.95	25.42
	QUAL	89.49	89.70	**98.20**	95.71	92.40	59.87	74.16	85.93	19.20	30.47	32.39	20.90
	Overall	57.85	56.77	**61.89**	58.12	55.15	48.62	51.12	54.03	26.72	31.03	24.78	33.85
real-app	GOAL	28.73	32.80	26.96	34.75	40.00	**43.80**	33.16	34.15	18.20	2.06	0.40	17.07
	LOGIC	44.40	45.20	36.36	39.80	37.20	43.27	41.10	**47.24**	33.80	2.06	0.40	20.00
	CONS	67.13	65.33	56.83	56.77	55.00	67.87	59.07	50.71	63.60	72.23	67.67	**82.79**
	UI	48.00	46.80	**70.24**	58.79	41.20	40.60	39.24	39.25	34.20	22.54	27.60	25.99
	QUAL	90.60	92.40	96.16	**97.84**	89.80	55.20	68.78	84.25	21.80	21.79	32.80	20.88
	Overall	55.77	56.51	57.31	**57.59**	52.64	50.15	48.27	51.12	34.32	24.14	25.77	33.35
grounding	GOAL	**23.88**	21.80	15.19	15.51	16.12	15.42	11.74	11.29	5.86	4.00	4.95	7.47
	LOGIC	36.94	41.00	37.39	37.55	**51.84**	44.58	35.65	36.13	17.17	18.20	14.43	15.35
	CONS	77.96	71.20	51.20	45.31	64.29	67.71	**83.26**	69.68	62.42	69.80	80.21	74.95
	UI	77.35	52.00	66.90	61.77	48.57	48.96	68.04	50.32	41.62	70.40	**80.41**	74.34
	QUAL	**92.24**	72.80	85.57	81.50	75.51	67.08	83.26	79.35	53.94	79.20	83.09	73.13
	Overall	**61.67**	51.76	51.25	48.33	51.27	48.75	56.39	49.35	36.20	48.32	52.62	49.05

Table A2: Detailed GUI-Score results on GUI-GenBench with **GPT-4o** (Hurst et al., 2024) as the judge.

Task	Metric	Nano Banana Pro	Nano Banana	GPT-Image -1.5	GPT-Image -1.0	Seedream 4.5	Seedream 4.0	Wan 2.6	Flux-2-Pro	Bagel	UniWorld	Qwen -Image-Edit	Longcat -Image
single-step	GOAL	**90.73**	74.57	88.37	80.41	63.37	60.56	71.38	73.10	31.40	47.30	37.40	40.28
	LOGIC	**94.25**	81.10	92.04	84.39	68.91	66.92	77.18	78.07	38.00	52.20	42.40	44.68
	CONS	**84.52**	79.49	78.27	68.88	70.40	70.32	84.34	83.25	55.90	62.20	64.60	61.44
	UI	**91.69**	78.09	88.57	82.96	66.00	65.55	75.84	78.48	42.20	55.40	55.60	51.00
	QUAL	78.95	69.54	**78.98**	77.14	55.37	58.71	69.68	77.06	37.80	53.20	56.90	48.56
	Overall	**88.03**	76.56	85.25	78.76	64.81	64.41	75.68	77.99	41.06	54.06	51.38	49.19
multi-step	GOAL	89.83	72.30	**95.44**	95.31	88.92	76.96	77.21	83.74	18.31	26.21	15.63	14.01
	LOGIC	84.99	72.80	87.16	**87.45**	80.43	70.14	73.03	78.38	26.83	36.56	32.28	29.53
	CONS	**83.12**	77.70	82.77	82.04	77.70	68.34	81.89	80.51	31.47	56.11	68.36	44.75
	UI	**89.80**	76.60	89.16	**89.80**	84.39	71.55	80.88	79.49	27.27	52.60	55.82	33.87
	QUAL	90.68	78.40	92.32	**93.98**	85.83	64.68	79.02	76.97	25.37	56.11	42.07	21.31
	Overall	87.68	75.56	89.37	**89.72**	83.45	70.33	78.41	79.82	25.85	45.52	42.83	28.69
fictional-app	GOAL	40.36	42.07	**51.60**	49.21	48.21	40.47	40.14	47.61	23.00	14.53	6.33	25.21
	LOGIC	62.28	61.48	**70.07**	65.15	62.13	55.46	60.89	64.53	43.40	24.87	17.07	46.94
	CONS	79.77	76.14	**85.33**	79.17	79.87	71.20	76.91	72.20	52.60	77.67	68.57	69.58
	UI	83.26	79.26	**91.28**	87.83	81.46	67.72	72.38	72.01	46.87	57.47	45.80	54.20
	QUAL	78.10	73.07	87.87	**92.32**	82.46	52.00	66.80	65.26	35.47	43.40	27.83	40.00
	Overall	68.75	66.40	**77.23**	74.74	70.83	57.37	63.42	64.32	40.27	43.59	33.12	47.19
real-app	GOAL	33.40	39.24	43.73	**43.83**	43.73	41.34	41.42	41.26	22.80	9.11	4.27	25.00
	LOGIC	55.80	59.32	61.72	59.06	57.20	55.47	50.25	**62.87**	50.67	19.42	17.53	43.80
	CONS	79.00	80.48	**81.66**	81.22	76.33	69.67	76.75	76.28	64.13	71.10	69.27	70.80
	UI	81.60	81.56	**88.40**	87.35	78.00	64.93	71.75	76.54	50.73	46.77	42.33	58.80
	QUAL	78.40	78.77	88.15	**92.40**	80.33	48.99	61.92	69.67	34.53	32.85	22.47	41.80
	Overall	65.64	67.87	72.73	**72.77**	67.12	56.08	60.42	65.32	44.57	35.85	31.17	48.04
grounding	GOAL	**35.60**	26.07	20.40	16.33	21.55	17.73	15.85	17.40	8.47	7.89	6.73	5.82
	LOGIC	**67.00**	66.13	53.80	55.80	58.38	58.14	57.21	58.87	54.73	61.00	63.87	62.07
	CONS	78.20	85.00	75.82	59.33	75.35	68.04	84.69	79.80	67.07	86.45	**91.53**	89.03
	UI	81.40	77.67	76.03	67.73	69.23	62.89	77.41	71.93	55.13	85.45	**91.47**	88.23
	QUAL	80.40	76.53	80.74	75.93	75.76	63.30	82.18	77.53	55.27	87.17	**90.33**	84.01
	Overall	68.52	66.28	61.36	55.02	60.05	54.02	63.47	61.11	48.13	65.59	**68.79**	65.83

Table A3: Detailed GUI-Score results on GUI-GenBench with **Qwen3-VL-235B-A22B-Thinking** (Bai et al., 2025) as the judge.

Task	Metric	Nano Banana Pro	Nano Banana	GPT-Image -1.5	GPT-Image -1.0	Seedream 4.5	Seedream 4.0	Wan 2.6	Flux-2-Pro	Bagel	UniWorld	Qwen -Image-Edit	Longcat -Image
single-step	GOAL	**93.64**	70.54	84.70	75.53	67.24	65.85	67.65	66.84	27.14	38.72	33.23	35.05
	LOGIC	**92.45**	72.86	81.27	71.20	66.02	64.41	69.13	68.27	30.44	38.32	35.32	34.78
	CONS	**83.28**	66.40	64.60	55.43	59.66	57.68	70.31	69.97	37.58	36.50	44.07	34.88
	UI	**94.25**	65.62	83.98	70.14	62.41	61.16	65.02	63.57	31.85	40.17	44.78	35.59
	QUAL	**98.04**	80.51	94.85	86.80	72.59	76.03	77.37	80.58	41.21	55.19	60.30	48.48
	Overall	**92.33**	71.19	81.88	71.82	65.58	65.03	69.90	69.85	33.64	41.78	43.54	37.76
multi-step	GOAL	92.35	60.97	93.61	**96.22**	89.39	83.45	72.70	83.80	4.62	6.44	4.22	3.70
	LOGIC	88.53	60.44	84.43	**89.46**	77.59	75.09	65.92	76.15	17.65	23.95	19.70	16.69
	CONS	**84.04**	71.32	75.50	79.22	73.23	71.26	76.32	72.29	19.69	42.06	57.16	30.29
	UI	**95.96**	60.77	90.68	92.31	75.00	73.23	74.48	64.74	16.12	33.32	51.10	21.16
	QUAL	97.65	71.89	96.34	**98.20**	86.87	80.17	87.13	77.29	17.65	48.06	53.59	21.16
	Overall	**91.71**	65.08	88.11	91.08	80.42	76.64	75.31	74.85	15.15	30.77	37.15	18.60
fictional-app	GOAL	28.13	34.48	**43.33**	34.68	39.19	34.27	27.50	34.13	15.69	9.73	8.07	18.40
	LOGIC	41.93	49.90	**58.33**	46.87	48.26	45.67	45.28	48.70	30.24	19.60	13.93	31.39
	CONS	49.47	**58.05**	52.87	43.30	49.26	43.73	49.51	43.33	25.86	53.07	20.53	36.04
	UI	53.47	55.69	**70.20**	63.91	55.64	45.27	47.64	40.67	22.36	34.00	19.47	29.93
	QUAL	76.53	72.73	**97.00**	94.75	82.21	60.47	73.12	63.00	20.88	34.00	24.47	29.03
	Overall	49.91	54.17	**64.35**	56.70	54.91	45.88	48.61	45.97	23.01	30.08	17.29	28.96
real-app	GOAL	30.47	30.47	34.73	37.33	36.67	**39.00**	27.95	35.08	17.58	8.40	4.33	21.73
	LOGIC	43.47	47.32	**50.73**	49.40	45.87	46.27	47.03	48.56	33.60	14.95	12.80	31.33
	CONS	58.33	**62.21**	53.13	49.93	48.73	42.80	51.30	47.81	34.75	37.75	17.60	43.96
	UI	59.80	59.53	**73.07**	65.47	50.47	41.53	50.88	48.42	26.06	22.39	15.13	34.93
	QUAL	75.60	78.52	**97.93**	93.67	80.13	56.00	67.03	66.94	20.88	22.73	16.60	34.53
	Overall	53.53	55.61	**61.92**	59.16	52.37	45.12	48.84	49.36	26.57	21.24	13.29	33.30
grounding	GOAL	24.38	20.89	15.68	18.72	19.31	14.41	16.05	16.58	12.44	12.50	6.69	9.08
	LOGIC	36.50	36.56	32.26	30.54	**37.66**	32.21	34.45	31.81	22.34	22.85	15.00	18.16
	CONS	46.20	42.89	33.77	25.10	40.96	38.76	**48.21**	46.44	35.79	40.83	42.04	42.13
	UI	**69.49**	58.11	56.85	44.30	56.22	52.55	67.22	56.98	40.07	59.03	68.66	61.35
	QUAL	**89.12**	79.38	81.10	70.40	76.29	72.76	86.08	73.56	53.11	81.39	84.23	80.21
	Overall	**53.14**	47.57	43.93	37.81	46.09	42.14	50.40	45.07	32.75	43.32	43.32	42.19

F TASK-SPECIFIC SCORING RUBRICS

Figures A3, A4, A5, and A6 reproduce the rubric cards used by the VLM judges. The five task suites share the GOAL, LOGIC, CONS, UI, and QUAL dimensions, while the descriptions specify how each dimension applies to the task. Performance tiers range from "Incomplete" to "Exceptional":

- Figure A3: single-step transitions, including whether the generated screen realizes the requested action.
- Figure A4: multi-step planning, including coherent progression and consistency across five-step trajectories.
- Figure A5: fictional-app generation and real-app trajectories, including structural plausibility and consistency in less familiar settings.
- Figure A6: grounding, including whether the resulting state corresponds to the coordinate-specified interaction.

The cards specify the qualitative basis of the ordinal ratings. They should be read alongside the generation protocol and the judge-specific results, rather than as guarantees of error-free evaluation.

JUDGE RUBRICS -- Type1: Single-Step

GOAL (Goal Achievement): Does the generated state achieve the caption-described change?

5: Caption change is achieved completely and unambiguously. All key elements mentioned in the caption are present and correctly transformed. No ambiguity about whether the change matches the description.

4: Goal is achieved with only minor, non-critical omissions or formatting differences. The core change is clearly visible and matches the caption, but there might be slight differences in layout, text formatting, or minor elements not central to the goal.

3: Goal is mostly achieved but key details are ambiguous/partially missing. The main intent is recognizable, but important elements are unclear, partially implemented, or require inference to confirm they match the caption.

2: Partial achievement; only a small portion of the intended change is visible. Most of the expected transformation is missing or incorrect, though some elements loosely relate to the caption.

1: Barely related; the change does not match caption semantics. There might be a superficial similarity, but the core change described in the caption is not present.

0: Complete failure; no relevant change or totally wrong change. The generated image shows either no change from the initial state or a change completely unrelated to the caption.

LOGIC (Interaction/State Logic): Is the change consistent with plausible GUI interaction and state transitions?

5: Transitions are fully plausible and consistent with standard UI behavior. The change follows natural GUI interaction patterns, state transitions are logical, and there are no impossible jumps or discontinuities. The transformation looks like a real UI would behave.

4: Mostly plausible; minor implausibility but still credible. The overall transition makes sense, but there might be slight inconsistencies that don't break the believability of the interaction (e.g., slightly accelerated animation, minor state inconsistency).

3: Some implausible elements; still partially coherent. The change shows recognizable UI logic in parts, but contains elements that don't follow standard interaction patterns or have logical gaps that require assuming missing intermediate steps.

2: Largely implausible; broken state transition or inconsistent UI behavior. The change shows significant logical problems, such as abrupt "teleportation" between unrelated states, contradictory UI element behaviors, or changes that would require impossible user actions.

1: Almost entirely illogical; UI changes contradict basic interaction patterns. The transformation violates fundamental UI principles, shows elements appearing/disappearing without cause, or demonstrates behaviors that no real GUI would exhibit.

0: Impossible; severe "teleportation" or nonsensical transformation. The change is completely disconnected from any plausible UI interaction, showing random or meaningless state changes that cannot be explained by any reasonable user action sequence.

CONS (Consistency/Preservation): Are unrelated regions preserved from the Initial Image?

5: Unaffected regions are preserved nearly perfectly. Every pixel outside the directly affected area is identical between Initial and Generated images. No drift, no color changes, no subtle alterations in preserved elements.

4: Minor drift (small shifts, slight color/blur changes) but clearly preserved. The preserved regions remain recognizable and functionally identical, though there might be nearly imperceptible changes due to compression or anti-aliasing artifacts.

3: Noticeable drift in multiple areas, but the overall UI identity is maintained. Elements outside the target region may shift slightly, show minor color variations, or have subtle blur, but the general layout and all UI components remain identifiable and in approximately correct positions.

2: Significant unintended changes outside the target region. Background elements show obvious movement, text changes in supposedly unchanged areas, or layout modifications that weren't part of the intended change. The preservation is clearly imperfect.

1: Widespread unintended changes; most of the UI is altered. Large sections of the UI that should remain unchanged show visible differences, with elements moving, changing appearance, or being corrupted across the screen.

0: Entire screen is corrupted or replaced. No meaningful preservation of unchanged regions; the entire UI appears transformed, broken, or replaced with unrelated content.

UI (UI Plausibility/Integrity): Are UI elements plausible (no hallucinations, broken layout, impossible widgets/states)?

5: UI components are coherent, correctly structured, and look native. All elements follow platform conventions, proper layering is maintained, states are correct, no visual anomalies, and text is perfectly rendered. The interface looks like it could be a real, professional GUI.

4: Mostly coherent; minor layout/element issues. The UI generally follows platform conventions with only subtle anomalies. There might be minor layering issues, slight misalignment, or small state inconsistencies, but overall it resembles a usable interface.

3: Several element/layout issues but still resembles a usable UI. Some UI elements appear slightly off (wrong sizes, minor misalignment, occasional text artifacts) but remain recognizable and the interface still appears functional despite the issues.

2: Many hallucinations or severe layout problems. Multiple UI components show significant problems: incorrect states, major layering errors, impossible layouts, or generated elements that don't exist in real UIs. The interface integrity is compromised.

1: UI is mostly broken; elements are nonsensical. Most UI components are incorrect, showing impossible states, major layering violations, or hallucinated controls that make no sense in the context. Very few elements appear correct or usable.

0: UI is unusable or completely hallucinated. The interface shows no recognizable UI elements, contains completely impossible layouts, hallucinated controls that violate all design principles, or appears as random visual noise rather than a coherent interface.

QUAL (Visual Quality): Visual readability (text/icon clarity, artifacts).

5: Crisp, readable, no obvious artifacts. Text and icons have sharp edges, perfect color accuracy, no compression artifacts, and professional rendering quality throughout the image.

4: Very readable; minor blur/artifacts. Text and icons remain clear and legible, with only slight compression artifacts or minimal edge blurring that doesn't affect readability.

3: Readable but noticeable blur/artifacts in important areas. Text is mostly readable but shows some artifacts, icons may be slightly blurry, and there might be minor color distortions in critical regions.

2: Significant quality degradation; text/icons hard to read. Obvious artifacts throughout the image, significant blurring of important elements, color issues, or compression noise that interferes with reading text or recognizing icons.

1: Severe artifacts; most text is unreadable. Major quality problems throughout, extensive blurring, strong color distortions, or heavy compression artifacts make most text illegible and icons unrecognizable.

0: Completely unusable image quality. Extreme artifacts, total blur, meaningless color shifts, or severe corruption make the image completely unreadable and unusable for any UI evaluation purpose.

Figure A3: GUI-Score rubric for Single-Step Visual Transition.

JUDGE RUBRICS -- Type2: Multi-Steps

GOAL (Final Goal Achievement): Does Frame 5 achieve the global goal?

5: Perfect achievement - Frame 5 shows EXACTLY what Global Goal describes, all objectives fully met with no ambiguity. The final state precisely matches the intended outcome described in the goal.

4: Strong achievement - Final goal achieved with only minor omissions or slight deviations that don't affect core objectives. Frame 5 accurately reflects the goal with only non-critical details missing or slightly different.

3: Partial achievement - Some aspects of final goal met, but important elements are missing or incorrectly implemented. Frame 5 shows progress toward the goal but key components are absent, wrong, or ambiguous.

2: Weak achievement - Only minimal progress toward final goal, most key objectives not met. Frame 5 barely resembles the intended outcome, with most critical elements missing or wrong.

1: Minimal achievement - Almost no progress toward final goal, only superficial similarity. Frame 5 might share some basic UI structure but shows virtually no meaningful progress toward the stated objective.

0: No achievement - Frame 5 completely unrelated to final goal or shows wrong outcome. The trajectory ends in a state that contradicts or has no meaningful connection to the global goal.

LOGIC (Interaction Logic Coherence): Are Frame0→5 transitions plausible and consistent with GUI interaction?

5: Perfect logic - Every interaction step is natural and reasonable, with coherent step-by-step progression and no jumps. Each frame logically follows from the previous one, showing realistic UI state transitions without teleportation or impossible changes. The entire trajectory demonstrates a believable user interaction flow.

4: Strong logic - Mostly natural progression with minor logical gaps, but overall interaction makes sense. The trajectory is mostly coherent, though there might be one or two slightly abrupt transitions that are still plausible within GUI interaction patterns.

3: Partial logic - Some steps follow expected flow, but several show logical jumps or unclear causation. The trajectory has noticeable gaps where it's unclear how one frame transitions to the next, requiring assumptions about missing intermediate states or actions.

2: Weak logic - Frequent logical inconsistencies across frames, progression is confusing. Multiple frames appear disconnected, showing "teleportation" between unrelated UI states, or transitions that contradict normal GUI behavior patterns.

1: Minimal logic - Actions barely follow expected flow, most steps are disconnected. The trajectory shows little coherent progression, with frames that seem randomly ordered or only loosely related to each other, making the interaction pattern unclear.

0: No logic - Frame-to-frame changes are completely disconnected and meaningless. The trajectory appears as a random sequence of unrelated UI states with no discernible interaction logic or coherent progression.

CONS (Visual Consistency Across Frames): Are UI elements stable across all 5 frames?

5: Perfect consistency - ALL UI elements (system bars, layout, style, colors) remain completely stable across all frames, zero drift. Every unchanged component maintains identical position, appearance, and state throughout the entire trajectory. No flicker, no subtle shifts, perfect stability.

4: Strong consistency - Minor variations in some elements (slight color shifts, subtle positioning) but core UI remains stable. Background elements might show nearly imperceptible changes, but all major components stay consistent and recognizable across frames.

3: Partial consistency - Noticeable drift in some UI elements between frames, but major components remain recognizable. Some elements shift position, change color slightly, or show minor flicker, but the overall UI structure and identity is maintained.

2: Weak consistency - Significant inconsistencies across frames: layout shifts, style changes, unstable UI elements. Background components move noticeably, colors change substantially, or there's clear flicker/corruption in multiple frames affecting UI stability.

1: Minimal consistency - UI changes substantially across frames, few elements remain stable, major visual drift. Most background elements show significant movement or change, with only a few components maintaining their position/appearance throughout the trajectory.

0: No consistency - Completely different UI across frames, no recognizable stability or coherent interface. Each frame appears to show a different UI layout, style, or completely unrelated interface elements with no continuity.

UI (Element Integrity): Are UI components in each frame native-looking and coherent?

5: Perfect integrity - All UI components follow platform conventions, proper layering, correct states, no visual anomalies, text perfectly rendered in EVERY frame. Every frame shows professional-quality UI that looks indistinguishable from a real native application interface.

4: Strong integrity - UI mostly follows conventions with minor layering issues or subtle anomalies, text mostly clear across frames. There might be occasional minor issues like slight misalignment or subtle rendering problems, but overall UI quality remains high in all frames.

3: Partial integrity - Some UI elements appear slightly off (wrong sizes, minor misalignment, occasional text artifacts) but still plausible in most frames. The UI remains recognizable and generally functional despite several noticeable imperfections across the trajectory.

2: Weak integrity - Noticeable UI problems in multiple frames: incorrect states, layering errors, text clarity issues, minor hallucinations. Multiple frames show significant UI defects that affect the believability and usability of the interface.

1: Poor integrity - Significant UI breakdown across frames: impossible states, major layering violations, unreadable text, clear hallucinations. Most frames contain serious UI problems that make the interface appear broken or nonsensical.

0: No integrity - UI completely nonsensical across frames, with hallucinated controls, impossible layouts, or no recognizable interface elements. The trajectory shows no coherent UI structure, appearing as random visual elements rather than a functional interface.

QUAL (Visual Quality): Are text/icons readable and artifacts minimal across all frames?

5: Excellent quality - Crystal clear text/icons in ALL frames, sharp edges, no compression artifacts, perfect color accuracy. Every frame shows professional rendering quality with crisp, easily readable text and sharp iconography throughout.

4: Good quality - Text/icons clear and readable across frames, minor compression artifacts or slight blurring at edges only. Slight quality degradation may be visible but doesn't affect readability or overall visual quality.

3: Fair quality - Text mostly readable but shows some artifacts in multiple frames, icons slightly blurry, minor color distortions. Quality issues are noticeable and affect some regions, but important text and icons remain legible.

2: Poor quality - Text becoming difficult to read in several frames, noticeable artifacts, significant blurring, color issues. Quality degradation is substantial and affects readability of important UI elements in multiple frames.

1: Very poor quality - Text barely legible across frames, severe artifacts, major blurring throughout, strong color distortions. Most text is extremely difficult or impossible to read, with major quality problems affecting the entire image.

0: Unusable quality - Text completely unreadable in most frames, extreme artifacts, totally blurred, colors meaningless. The visual quality is so poor that the UI cannot be evaluated or used for any practical purpose.

Figure A4: GUI-Score rubric for Multi-Step Planning.

JUDGE RUBRICS -- Type3/4: Fiction-App/Real-App

GOAL (Final Goal Achievement): Does Frame5 achieve the Final Goal?

5: Perfect achievement - Frame5 shows EXACTLY what Final Goal describes, all objectives fully met with no ambiguity. The final state precisely matches the intended outcome, with all key elements correctly implemented and clearly visible.

4: Strong achievement - Final Goal achieved with only minor omissions or slight deviations that don't affect core objectives. Frame5 accurately reflects the goal with only non-critical details missing or slightly different from the exact description.

3: Partial achievement - Some aspects of Final Goal met, but significant elements are missing or incorrectly implemented. Frame5 shows progress toward the goal but key components are absent, wrong, or ambiguous, requiring inference to see the connection.

2: Weak achievement - Only minimal progress toward Final Goal, most key objectives not met. Frame5 barely resembles the intended outcome, with most critical elements missing or wrong, showing only surface-level similarity.

1: Minimal achievement - Almost no progress toward Final Goal, only superficial similarity. Frame5 might share some basic UI structure with the goal but shows virtually no meaningful progress toward the stated objective.

0: No achievement - Frame5 completely unrelated to Final Goal or shows wrong outcome. The trajectory ends in a state that contradicts or has no meaningful connection to the desired final state.

LOGIC (Action-Visual Alignment & Step Coherence): Do actions logically lead to visual changes described?

5: Perfect logic - Every action naturally and plausibly causes the visual changes described in the Visual Description, with coherent step-by-step progression. Each action-result pair is believable and follows natural GUI interaction patterns without requiring unexplained assumptions.

4: Strong logic - Actions mostly lead to expected visual changes, minor logical gaps but overall progression makes sense. The action-visual alignment is mostly correct, though one or two steps might have slight inconsistencies that don't break the overall coherence.

3: Partial logic - Some actions align with visual changes, but several steps show logical jumps or unclear causation. The progression requires making assumptions about missing intermediate actions, and some steps don't clearly follow from their described actions.

2: Weak logic - Actions and visual changes poorly connected, frequent logical inconsistencies across steps. Multiple actions don't realistically cause the described visual changes, and the progression shows significant gaps in causality that make the trajectory confusing.

1: Minimal logic - Actions barely relate to visual changes, progression is confusing or nonsensical. Most actions are disconnected from the resulting UI states, requiring major leaps of logic to understand how one step leads to the next.

0: No logic - Actions completely disconnected from visual changes, no coherent progression through steps. The action descriptions and visual results appear unrelated, showing no meaningful causal relationship throughout the trajectory.

CONS (Visual Consistency Across Frames): Are UI elements stable across all 5 frames?

5: Perfect consistency - ALL UI elements (system bars, layout, style, colors) remain completely stable across all frames, zero drift. Every unchanged component maintains identical position, appearance, and state throughout the entire trajectory with perfect preservation.

4: Strong consistency - Minor variations in some elements (slight color shifts, subtle positioning) but core UI remains stable. Background elements might show nearly imperceptible changes, but all major components stay consistent and recognizable across frames.

3: Partial consistency - Noticeable drift in some UI elements between frames, but major components remain recognizable. Some elements shift position, change color slightly, or show minor flicker, but the general layout and all UI components remain identifiable.

2: Weak consistency - Significant inconsistencies across frames: layout shifts, style changes, unstable UI elements. Background components move noticeably, colors change substantially, or there's clear flicker/corruption in multiple frames affecting UI stability.

1: Minimal consistency - UI changes substantially across frames, few elements remain stable, major visual drift. Most background elements show significant movement or change, with only a few components maintaining their position/appearance throughout.

0: No consistency - Completely different UI across frames, no recognizable stability or coherent interface. Each frame appears to show a different UI layout, style, or completely unrelated interface elements with no continuity.

UI (UI Plausibility/Integrity): Are UI components in each frame native-looking and coherent?

5: Perfect integrity - All UI components follow platform conventions, proper layering, correct states, no visual anomalies, text perfectly rendered in EVERY frame. Every frame shows professional-quality UI that looks indistinguishable from a real native application interface.

4: Strong integrity - UI mostly follows conventions with minor layering issues or subtle anomalies, text mostly clear across frames. There might be occasional minor issues like slight misalignment or subtle rendering problems, but overall UI quality remains high.

3: Partial integrity - Some UI elements appear slightly off (wrong sizes, minor misalignment, occasional text artifacts) but still plausible in most frames. The UI remains recognizable and generally functional despite several noticeable imperfections.

2: Weak integrity - Noticeable UI problems in multiple frames: incorrect states, layering errors, text clarity issues, minor hallucinations. Multiple frames show significant UI defects that affect the believability and usability of the interface.

1: Poor integrity - Significant UI breakdown across frames: impossible states, major layering violations, unreadable text, clear hallucinations. Most frames contain serious UI problems that make the interface appear broken or nonsensical.

0: No integrity - UI completely nonsensical across frames, with hallucinated controls, impossible layouts, or no recognizable interface elements. The trajectory shows no coherent UI structure, appearing as random visual elements.

QUAL (Visual Quality): Are text/icons readable and artifacts minimal across all frames?

5: Excellent quality - Crystal clear text/icons in ALL frames, sharp edges, no compression artifacts, perfect color accuracy. Every frame shows professional rendering quality with crisp, easily readable text and sharp iconography.

4: Good quality - Text/icons clear and readable across frames, minor compression artifacts or slight blurring at edges only. Slight quality degradation may be visible but doesn't affect readability or overall visual quality.

3: Fair quality - Text mostly readable but shows some artifacts in multiple frames, icons slightly blurry, minor color distortions. Quality issues are noticeable and affect some regions, but important text and icons remain legible.

2: Poor quality - Text becoming difficult to read in several frames, noticeable artifacts, significant blurring, color issues. Quality degradation is substantial and affects readability of important UI elements in multiple frames.

1: Very poor quality - Text barely legible across frames, severe artifacts, major blurring throughout, strong color distortions. Most text is extremely difficult or impossible to read, with major quality problems affecting the entire image.

0: Unusable quality - Text completely unreadable in most frames, extreme artifacts, totally blurred, colors meaningless. The visual quality is so poor that the UI cannot be evaluated or used for any practical purpose.

Figure A5: GUI-Score rubrics for Zero-shot Virtual GUI and Rare Trajectory Synthesis.

JUDGE RUBRICS -- Type5: Grounding

GOAL (Effect Alignment): Does Generated Image match Expected Effect?

5: Perfect match - Effect achieved exactly as described, all mentioned elements present and correctly transformed with no ambiguity.
4: Strong match - Effect achieved with only minor omissions or slight deviations that don't affect core semantics.
3: Partial match - Some aspects match Expected Effect, but significant elements are missing or incorrectly implemented.
2: Weak match - Change vaguely relates to Expected Effect but misses most key elements or introduces wrong transformations.
1: Minimal match - Almost no correspondence between change and Expected Effect, only superficial similarity.
0: No match - Change completely contradicts or is unrelated to Expected Effect.

Goal checklist (use to justify scoring):
- Correct UI outcome happened (e.g., dialog opened, dropdown expanded, tab switched, checkbox toggled, page navigated).
- Direction is correct (open vs close, enable vs disable, selected vs unselected).
- If Expected Effect mentions specific text labels/selected items, they must be visible and correct for 4-5.
- If Expected Effect implies a very specific screen/state, wrong screen/state caps goal at 2.

LOGIC (Trigger Localization): Is change plausibly triggered by tapping given point/region?

5: Perfect localization - Change originates PRECISELY from tapped element/region with appropriate ripple effects and spatial reasoning.
4: Strong localization - Change originates from area very close to tap point, minor spatial offset but clearly connected and plausible.
3: Partial localization - Change appears in vicinity of tap but connection is ambiguous or spans multiple unrelated UI elements.
2: Weak localization - Change appears in entirely different region from tap point, showing poor spatial understanding.
1: Minimal localization - Only trace evidence that change might relate to tap location, mostly disconnected.
0: No localization - Change completely unrelated to tap location or affects random/global areas without spatial coherence.

Logic checklist:
- Change should be anchored to tapped widget/region or its plausible dependent area.
- Plausible: button→modal/dialog, tab→content panel, menu item→navigation, toggle→state change.
- Implausible: unrelated far-away changes, global theme swaps, random text/icon rewrites, unexplained page jumps.

CONS (Background Consistency): Are non-affected regions preserved relative to Initial Image?

5: Perfect consistency - EVERY pixel outside directly affected area is IDENTICAL between Initial and Generated images.
4: Strong consistency - Minor variations in non-affected regions (subtle color shifts, compression artifacts) but layout/content unchanged.
3: Partial consistency - Some background elements shift or change slightly, but major layout remains stable.
2: Weak consistency - Noticeable drift in background elements, text changes, or layout modifications in supposedly unchanged areas.
1: Minimal consistency - Background shows significant changes but retains some similar elements or overall structure.
0: No consistency - Background completely different, showing no preservation of unchanged regions.

Consistency guidance:
- A tap should not rewrite large unrelated regions. If large areas change without clear reason, cons should be ≤2.
- Minor compression/resampling artifacts are acceptable for 4; new/changed text in unrelated areas is not.

UI (UI Plausibility/Integrity): Are UI components coherent (native-looking, correct layering, no hallucinations)?

5: Perfect integrity - All UI components follow platform conventions, proper layering, correct states, no visual anomalies, text perfectly rendered and readable.
4: Strong integrity - UI mostly follows conventions with minor layering issues or subtle anomalies, text mostly clear with minimal artifacts.
3: Partial integrity - Some UI elements appear slightly off (wrong sizes, minor misalignment, occasional text artifacts) but still recognizable.
2: Weak integrity - Noticeable UI problems: incorrect component states, layering errors, text clarity issues, minor hallucinations or impossible layouts.
1: Poor integrity - Significant UI breakdown: impossible states, major layering violations, unreadable text, clear hallucinations of non-existent controls.
0: No integrity - UI completely nonsensical with hallucinated controls, impossible layouts, or no recognizable interface elements.

UI checklist:
- Correct component hierarchy (e.g., modal overlays the page; dropdown anchored to its control).
- Correct affordances and states (pressed/selected/disabled look plausible).
- No duplicated/smeared labels; consistent fonts and spacing.

QUAL (Visual Quality): Are text/icons readable and artifacts minimal?

5: Excellent quality - Crystal clear text and icons, sharp edges, no compression artifacts, perfect color accuracy, professional rendering.
4: Good quality - Text/icons clear and readable, minor compression artifacts or slight blurring at edges only.
3: Fair quality - Text mostly readable but shows some artifacts, icons slightly blurry, minor color distortions present.
2: Poor quality - Text becoming difficult to read, noticeable artifacts throughout, significant blurring, color issues.
1: Very poor quality - Text barely legible, severe artifacts, major blurring throughout, strong color distortions.
0: Unusable quality - Text completely unreadable, extreme artifacts, totally blurred, colors meaningless.

Figure A6: GUI-Score rubric for Grounding-based Generation.