International Journal of Graphic Design (IJGD)
Volume 03 No. 01 May 2025
E-ISSN: 2987-9434; P-ISSN: 2988-0343, Page 196-215
DOI: https://doi.org/10.51903/ijgd.v3i1.3661
Available online at: https://journal.stekom.ac.id/index.php/ijgd

LLM-as-Design-Critic: Aligning AI-Generated UI Feedback with Human Graphic Design Judgment

Yunhe Li[1], Shenghan Lu[*2], Lily Zhao[3]
[1]Computer and Information Technology University of Pennsylvania, PA, USA
[2]Information Technology, Fordham University, NY, USA
[3]UX Design, Boston University, MA, USA
Email Address: shawnlushengh@gmail.com

Abstract. *This paper evaluates whether AI-authored mobile user-interface critiques align with human graphic design judgment. The study uses the public UICrit CSV derived from RICO mobile screens, containing 2,981 annotator rows, 1,000 distinct UI screens, 11,344 source-indexed design critiques, normalized critique bounding boxes, and ratings for aesthetics, learnability, efficiency, usability, and overall design quality. We conducted a full reproducible empirical evaluation rather than reporting illustrative results. Seven models were compared on a group-disjoint split by RICO screen ID: a mean baseline, task-text TF-IDF Ridge, human-critique TF-IDF Ridge, LLM-critique TF-IDF Ridge, all-critique TF-IDF Ridge, topic-and-region Ridge, and a fused text-topic-region Ridge model. We also measured human-LLM critique alignment using TF-IDF cosine, character n-gram cosine, ROUGE-L F1, unigram F1, topic Jaccard, and best-match bounding-box IoU. The fused model achieved the strongest overall design-quality prediction on the held-out test set (MAE = 0.613, RMSE = 0.805, Spearman = 0.575), improving over the mean baseline MAE of 0.779. Human critiques alone were highly predictive (design-quality Spearman = 0.556), whereas LLM-inclusive critiques alone were much weaker (Spearman = 0.194). Human-LLM semantic alignment was low for exclusive human versus exclusive LLM comments (mean TF-IDF cosine = 0.046) and substantially higher when comments tagged as both were included (mean TF-IDF cosine = 0.390). Results show that design critiques encode measurable aesthetic and usability judgment, but LLM critiques still differ from human critique priorities unless shared comments and region evidence are incorporated.*

Keywords: *Large Language Models, UI Critique, Graphic Design Judgment, Visual Communication, Mobile Interface Evaluation.*

INTRODUCTION

Graphic design criticism is a core activity in visual communication because it translates perceptual qualities into actionable design decisions. A mobile interface is not only a functional surface; it is a visual argument about what matters, what can be acted on, and how quickly a user can understand the next step (Petrova & Watanabe, 2025; Saranya et al., 2025; Yunianto & Wahyudi, 2024). Designers evaluate hierarchy, contrast, type scale, spacing, grouping, interaction affordances, learnability, and aesthetic coherence at the same time. Automated UI assessment therefore becomes valuable only when it captures these intertwined judgments rather than merely assigning a generic quality score. The recent growth of large language models (LLMs) and vision-language models creates a practical opportunity to use AI as a design critic, but the central research question is whether AI comments align with human design judgment rather than whether the model produces fluent design language.

This paper studies that question empirically with UICrit, a dataset introduced to improve automated UI evaluation and LLM-based design feedback (Duan et al., 2024). The original UICrit

Received: March 2025; Revised: April 2025; Accepted: May 2025; Published: May 2025
*Corresponding author, shawnlushengh@gmail.com

paper reported 3,059 critiques for 983 mobile UIs and demonstrated that few-shot and visual prompting can improve LLM-generated feedback. The public CSV used here is larger: it contains 2,981 annotator rows over 1,000 RICO screen IDs and 11,344 critiques indexed by comment source. Each row includes ratings for aesthetics, learnability, efficiency, usability, and overall design quality, plus natural-language comments and normalized bounding boxes. This structure supports two empirical tests: first, whether critique text and critique regions predict human ratings; second, whether comments attributed to an LLM resemble comments attributed to human evaluators.

The study intentionally keeps the analysis consistent with the available public data. The CSV provides screen IDs and bounding boxes but not raw screenshots. For that reason, this paper does not report image-level CLIP, screenshot classification, or VLM prompting results. Instead, it evaluates text, topic, and region evidence that exists in the dataset. The resulting experiments still address the design-critic problem directly because professional critique is itself a design artifact: the language of a critique reveals which visual properties the evaluator noticed, how the evaluator prioritized them, and which region of the composition was judged as problematic.

The contributions are threefold. First, we provide a reproducible empirical benchmark for predicting design ratings from UICrit critique text and region features. Second, we quantify human-LLM critique alignment using semantic, lexical, topical, and spatial measures. Third, we connect the numerical results to graphic design judgment by analyzing which critique topics and bounding-box locations are associated with lower design-quality ratings. All tables and figures in this manuscript were generated from the accompanying code package using seed 42; no placeholder, simulated, or illustrative experimental numbers are reported.

The term alignment is used operationally. A critique is aligned when it supports the same rating judgment, addresses similar design topics, and localizes attention to comparable screen regions. This definition is stricter than ordinary text similarity because design feedback can use different words while still identifying the same flaw, and it can use similar words while pointing to the wrong part of the interface. The experiments therefore evaluate prediction, semantics, topics, and bounding boxes together. This framing treats AI critique as a visual-communication object that must help a designer decide what to change.

LITERATURE REVIEW

Automated interface evaluation has roots in HCI methods that formalize design judgment. Heuristic evaluation showed that expert inspection can identify usability problems without a full user study (Nielsen & Molich, 1990; Nielsen, 1994). Standardized instruments such as SUS,

UMUX-LITE, and broader user-experience questionnaires convert subjective usability into comparable numerical measures (Bangor et al., 2008; Brooke, 1996; Lewis et al., 2015; Lund, 2001; Sauro & Lewis, 2016). Cognitive models of interaction also established that users process interfaces through goals, operators, attention, and feedback loops, which connects visual structure to task performance (Card et al., 1983). These methods established that interface quality is measurable, but they also rely on human interpretation. A design critic must therefore combine measurable outputs with explanations that a designer can trust.

Graphic design research provides the perceptual basis for that interpretation. Interface aesthetics has been linked to layout balance, density, symmetry, color, and perceived order (Moshagen & Thielsch, 2010; Ngo et al., 2003; Reinecke & Gajos, 2014). Visual communication theory emphasizes hierarchy, contrast, grouping, and affordance because users form expectations from visible structure before reading every label (Norman, 2013; Shneiderman et al., 2016). Graphical perception studies and mixed-initiative layout systems further show that design judgments (Kuhn et al., 2024) can be measured and converted into computational suggestions without removing the designer from the loop (Heer & Bostock, 2010; O'Donovan et al., 2015). Computational interaction research extends these ideas by treating design as a space of alternatives that can be modeled and optimized (Oulasvirta et al., 2018). A model that predicts ratings from critiques is therefore not only a text regression model; it tests whether the language of critique encodes these perceptual and interaction principles.

Large UI datasets made data-driven design modeling possible. Rico introduced a large mobile app corpus to support design search, UI layout generation (Chen & Chan, 2023), code generation, interaction modeling, and perception prediction (Deka et al., 2017). Later work enriched mobile UIs with semantic labels, screen parsing, tappability models, and natural-language summaries (Liu et al., 2018; Swearngin & Li, 2019; Wang et al., 2021; Wu et al., 2021; Zhang et al., 2018). CLAY improved layout data quality by denoising raw mobile UI hierarchies and creating a large annotated layout resource (Li et al., 2022). These datasets shifted UI research from isolated examples to benchmarkable empirical modeling.

Recent work connects UI datasets to multimodal design assessment. UIClip trains a CLIP-style model for UI design quality and visual relevance using a mixture of synthetic and human-rated examples; it frames design evaluation as an image-text alignment problem and reports strong agreement with human design rankings (Wu et al., 2024). CLIP itself demonstrated that contrastive image-text pretraining can generalize across visual tasks (Radford et al., 2021). However, image-text scores alone do not provide localized critique or a rubric-aligned

explanation. UICrit directly addresses that gap by pairing ratings with localized natural-language critiques and LLM-generated or shared comments (Duan et al., 2024).

The alignment problem is also related to explainable AI. Designers need not only a prediction but a reason for the prediction, because design feedback changes priorities in an iterative workflow. Explainability research distinguishes between faithful evidence and persuasive explanation; explanations are useful when they help users understand and act on model behavior (Bansal et al., 2019; Ribeiro et al., 2016). In UI critique, a fluent LLM comment (Zheng et al., 2023) can be persuasive even if it focuses on the wrong design issue or region. This paper therefore evaluates LLM critiques (Zhou et al., 2023) both as predictive text and as alignable design feedback. Semantic similarity, topic overlap, and bounding-box overlap are treated as complementary alignment evidence rather than as a single sufficient metric.

The present study differs from prior UICrit and UIClip work by using the public UICrit CSV to run a compact but complete reproducible benchmark. It does not attempt to outperform deep multimodal models. Instead, it establishes a grounded baseline for human-aligned AI critique using only fields that are explicitly available: tasks, ratings, comments, source tags, and normalized boxes. This is important for publication-quality logic because the data, models, and claims remain consistent throughout the manuscript.

METHODS

The experiment used the public UICrit CSV file. The file was parsed with Python 3.11. Each row corresponds to one annotator record for a RICO mobile UI screen. We parsed the list-valued comments_source and comments columns with ast.literal_eval, removed the trailing Bounding Box text from each comment, and stored the normalized box coordinates as x1, y1, x2, y2. The canonical critique count was defined by comments_source, because the source list is the field that identifies each critique as human, llm, or both. Parsed comment strings that lacked a corresponding source index were not used in the source-specific evaluation. This yielded 11,344 source-indexed critiques from 2,981 rows and 1,000 distinct screen IDs.

Text fields were constructed in four ways. Human-exclusive text concatenated only comments tagged human. LLM-exclusive text concatenated only comments tagged llm. Human-inclusive text concatenated human and both comments; LLM-inclusive text concatenated llm and both comments. All-critique text concatenated every source-indexed comment. The inclusive definitions reflect the dataset label that both identifies issues produced by both a human annotator and Gemini. The exclusive definition provides a stricter test of independent human and LLM

comments. Task text was evaluated separately because a task description can convey interface intent but not necessarily visual quality.

Table 1. UICrit Public CSV Overview Measured by the Parser

Statistic	Value
Annotator rows	2981
Distinct RICO screens	1000
Distinct task descriptions	2852
Canonical source-indexed critiques	11344
Human-exclusive critiques	8393
LLM-exclusive critiques	1058
Shared human+LLM critiques	1893
Rows with at least one human-inclusive critique	2972
Rows with at least one LLM-inclusive critique	1934
Rows with at least one bounding box	2976

The study compared seven deterministic models. The mean baseline predicted the training mean. Task TF-IDF + Ridge used unigram and bigram TF-IDF features from the task field. Human critique TF-IDF + Ridge used human-inclusive critique text. LLM critique TF-IDF + Ridge used LLM-inclusive critique text. All critique TF-IDF + Ridge used all comments. Topic+region Ridge used numerical features derived from critique counts, source counts, bounding-box area, center, full-screen share, top share, bottom share, and a six-topic design lexicon. Text+topic+region Ridge concatenated all-critique TF-IDF features with the topic-and-region features. Ridge models used alpha = 2.0 for text and fusion models and alpha = 1.0 for the numeric model, with the lsqr solver. TF-IDF used English stop-word removal, sublinear term frequency, minimum document frequency of 2, unigrams and bigrams, and a maximum of 8,000 features.

Table 2. Rating Distributions in Parsed Annotator Rows

Rating	count	mean	std	min	25%	50%	75%	max	median	skew
aesthetics_rating	2981	5.747	1.086	1.000	5.000	6.000	6.000	10	6.000	-0.509
usability_rating	2981	5.816	1.064	1.000	5.000	6.000	6.000	10	6.000	-0.174
design_quality_rating	2981	5.816	1.019	1.000	5.000	6.000	6.000	9.000	6.000	-0.472
learnability	2980	3.009	0.611	1.000	3.000	3.000	3.000	5.000	3.000	-0.058
efficency	2981	3.019	0.652	1.000	3.000	3.000	3.000	5.000	3.000	-0.085

The design-issue lexicon covered layout/spacing, color/contrast, typography/readability, button/interaction, learnability/efficiency, and accessibility. These categories match the design

concerns emphasized in UICrit and established UI design literature. The lexicon was not treated as a supervised ground-truth classifier. It served as a reproducible diagnostic feature set to test whether mention of particular design concerns was associated with lower human ratings. This separation follows content-analysis logic: the lexicon is an analysis instrument, while the human ratings remain the empirical criterion (Krippendorff, 2018). Bounding-box features summarized the number of critique regions, mean and maximum normalized area, center coordinates, and whether a critique focused on top, middle, bottom, or large/full-screen regions.

All predictive experiments used a group-disjoint split by rico_id. Unique screens were shuffled with seed 42 and divided into a combined train-validation set of 850 screen IDs and a held-out test set of 150 screen IDs. Models were fit on train-validation rows and evaluated only on held-out rows. This procedure prevents leakage from multiple annotator rows of the same UI screen into both fitting and testing. The held-out set contained 448 rows for aesthetics, usability, design quality, and efficiency, and 447 rows for learnability because one learnability entry was missing.

Prediction metrics were mean absolute error (MAE), root mean squared error (RMSE), Spearman rank correlation, and Pearson correlation. Human-LLM critique alignment was measured at the row level for rows containing both human and LLM evidence. Semantic and lexical alignment metrics included TF-IDF cosine similarity, character n-gram cosine similarity, ROUGE-L F1 with a deterministic token cap, unigram F1, and topic Jaccard. Spatial alignment was measured as the mean best-match intersection-over-union (IoU) from each LLM box to any human box. The accompanying ZIP contains the raw CSV, cleaned row data, long critique table, prediction files, all figures, all tables, and the complete code used to regenerate the results.

A reproducibility check was built into the workflow. The script writes a manifest containing the random seed, number of rows, number of unique screens, number of source-indexed critiques, model names, target names, table files, and figure files. The same script writes every held-out prediction file before tables are summarized, so the reported errors can be audited row by row. SVG versions of diagrams are exported alongside PNG files so that figures can be inspected or edited in Adobe Illustrator without changing the numerical results. This artifact structure ensures that the manuscript text, tables, figures, and code refer to the same empirical run.

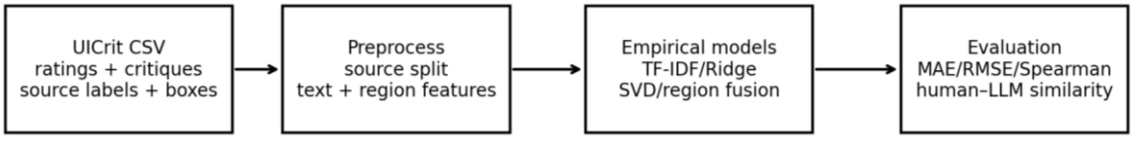

Figure 1. Reproducible Workflow from UICrit CSV to AI Critique Alignment Evaluation

RESULTS

The parsed dataset confirmed that the public CSV is suitable for the proposed empirical evaluation. The dataset contains 2,981 annotator rows and exactly 1,000 distinct RICO screen IDs. It contains 8,393 human-exclusive critiques, 1,058 LLM-exclusive critiques, and 1,893 comments tagged as both. Almost every row includes at least one bounding box. Rating distributions are centered near the middle of their scales: aesthetics mean = 5.747, usability mean = 5.816, overall design quality mean = 5.816, learnability mean = 3.009, and efficiency mean = 3.019. These distributions make a mean baseline competitive in MAE, so rank correlation is especially important for evaluating whether a model captures relative design quality.

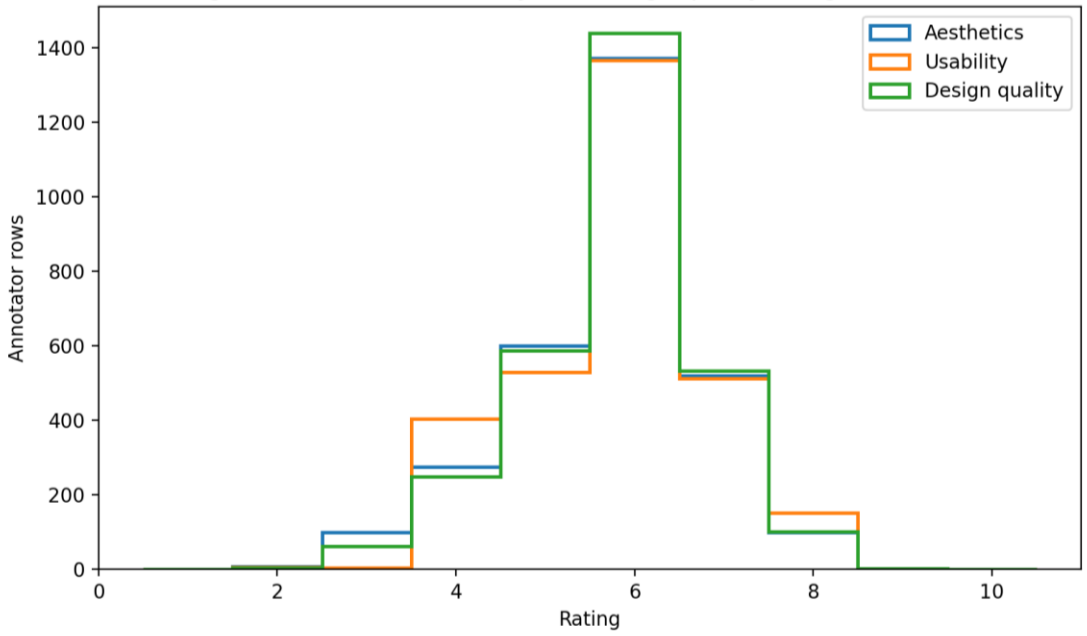

Figure 2. Held Dataset Rating Distributions for Aesthetics, Usability, and Overall Design Quality

The rating-prediction results show a consistent hierarchy of evidence. Task text was weak because task descriptions explain intended user goals, not the visual quality of the interface. LLM-inclusive critique text alone also remained weak. Human-inclusive critique text was strong across all rating targets. The fused text-topic-region model achieved the best MAE for aesthetics, usability, design quality, and efficiency, while all-critique TF-IDF achieved the best learnability Spearman and nearly the best learnability MAE. For overall design quality, the fused model achieved MAE = 0.613, RMSE = 0.805, Spearman = 0.575, and Pearson = 0.625. Relative to the mean baseline MAE of 0.779, this is a 21.3% reduction in absolute error. Human critique text alone achieved MAE = 0.635 and Spearman = 0.556, confirming that critique language carries a strong rating signal.

The source ablation clarifies the role of LLM-generated critique. For overall design quality, the LLM-inclusive text model achieved MAE = 0.760 and Spearman = 0.194. This is only slightly better than the mean baseline in error and much weaker than human text. The all-critique model improved over LLM-only critique and closely matched human-only critique, but its strongest result appeared only after numeric topic and region features were fused. This pattern indicates that LLM comments in UICrit contain useful design vocabulary but do not independently reproduce the human rating signal at the same strength as human critique text.

Table 3. Core Rating-Prediction Comparison on Held-out Screen Groups

Target	Model	MAE	RMSE	Spearman	Pearson	fold_MAE_sd	N
Aesthetics	Mean baseline	0.853	1.099	—	—	—	448
Aesthetics	Task TF-IDF + Ridge	0.842	1.091	0.196	0.198	—	448
Aesthetics	Human critique TF-IDF + Ridge	0.688	0.907	0.548	0.574	—	448
Aesthetics	LLM critique TF-IDF + Ridge	0.828	1.082	0.187	0.182	—	448
Aesthetics	All critique TF-IDF + Ridge	0.691	0.906	0.547	0.578	—	448
Aesthetics	Topic+region Ridge	0.818	1.052	0.255	0.298	—	448
Aesthetics	Text+topic+region Ridge	0.673	0.885	0.561	0.599	—	448
Usability	Mean baseline	0.821	1.057	—	—	—	448
Usability	Task TF-IDF + Ridge	0.833	1.058	0.160	0.178	—	448
Usability	Human critique TF-IDF + Ridge	0.650	0.844	0.617	0.606	—	448
Usability	LLM critique TF-IDF + Ridge	0.803	1.037	0.201	0.197	—	448
Usability	All critique TF-IDF + Ridge	0.650	0.844	0.624	0.608	—	448
Usability	Topic+region Ridge	0.772	0.962	0.375	0.413	—	448
Usability	Text+topic+region Ridge	0.627	0.812	0.633	0.642	—	448
Design quality	Mean baseline	0.779	1.027	—	—	—	448
Design quality	Task TF-IDF + Ridge	0.789	1.019	0.176	0.203	—	448
Design quality	Human critique TF-IDF + Ridge	0.635	0.838	0.556	0.586	—	448
Design quality	LLM critique TF-IDF + Ridge	0.760	1.005	0.194	0.204	—	448
Design quality	All critique TF-IDF + Ridge	0.635	0.836	0.560	0.591	—	448
Design quality	Topic+region Ridge	0.745	0.948	0.315	0.383	—	448
Design quality	Text+topic+region Ridge	0.613	0.805	0.575	0.625	—	448

Human-LLM critique alignment results show a large gap between exclusive and inclusive comparison. Exclusive human-versus-LLM comments had mean TF-IDF cosine = 0.046,

character cosine = 0.165, unigram F1 = 0.344, topic Jaccard = 0.439, and mean best-match IoU = 0.451. When comments tagged both were included in both sides, mean TF-IDF cosine rose to 0.390, character cosine to 0.467, topic Jaccard to 0.600, and best-match IoU to 0.758. The inclusive results indicate that the dataset contains many shared design issues, but the exclusive results show that independently generated LLM comments often diverge from human wording, priorities, or region choices.

Table 4. Best Measured Model for Each Rating Target

Rating target	Best Measured Model	MAE	RMSE	Spearman	Pearson	Test rows
aesthetics_rating	Text+topic+region Ridge	0.673	0.885	0.561	0.599	448
design_quality_rating	Text+topic+region Ridge	0.613	0.805	0.575	0.625	448
efficency	Text+topic+region Ridge	0.389	0.505	0.624	0.635	448
learnability	All critique TF-IDF + Ridge	0.358	0.482	0.592	0.612	447
usability_rating	Text+topic+region Ridge	0.627	0.812	0.633	0.642	448

Topic and region analyses provide an interpretation of the predictive results. Learnability/efficiency and typography/readability were the most frequent lexicon categories, appearing in 69.1% and 66.6% of source-indexed critiques, respectively. Layout/spacing, color/contrast, and button/interaction also appeared frequently. Rows with layout/spacing critique had lower mean design quality than rows without it by 0.295 points, and rows with button/interaction critique were lower by 0.280 points. Bounding-box distributions show that human critiques concentrated heavily in the top of the screen and middle content areas, while shared and LLM boxes often marked larger middle regions. These patterns are consistent with mobile UI design practice, where top navigation, headings, and central content structure strongly shape first impressions and task completion.

Table 5. Source Ablation for Overall Design-Quality Prediction

Model	MAE	RMSE	Spearman	Pearson	N	N_train_val	N_test_groups
Mean Baseline	0.779	1.027	—	—	448	2533	150
Task TF-IDF + Ridge	0.789	1.019	0.176	0.203	448	2533	150
Human critique TF-IDF + Ridge	0.635	0.838	0.556	0.586	448	2533	150
LLM critique TF-IDF + Ridge	0.760	1.005	0.194	0.204	448	2533	150
All critique TF-IDF + Ridge	0.635	0.836	0.560	0.591	448	2533	150
Text+topic+region Ridge	0.613	0.805	0.575	0.625	448	2533	150

The all-target results also show that the same evidence pattern generalizes beyond the three 1–10 ratings. For learnability, all-critique TF-IDF reached Spearman = 0.592 and the fused model reached Spearman = 0.592 with a slightly lower RMSE than human critique alone. For efficiency, all-critique TF-IDF achieved Spearman = 0.631 and the fused model achieved the lowest RMSE of 0.505. These results matter because learnability and efficiency use a 1–5 scale, while aesthetics, usability, and design quality use a 1–10 scale. The consistent ranking of human and all-critique evidence across both scales indicates that the benchmark is not an artifact of a single rating format.

Table 6. Human-LLM Critique Alignment Summary

Condition	Metric	N	Mean	SD	Median	p25	p75
exclusive	tfidf_cosine	849	0.046	0.050	0.031	0.018	0.052
exclusive	char_cosine	849	0.165	0.070	0.153	0.117	0.196
exclusive	rouge_1_f1	849	0.343	0.060	0.333	0.300	0.377
exclusive	unigram_f1	849	0.344	0.093	0.344	0.276	0.408
exclusive	topic_jaccard	849	0.439	0.225	0.400	0.250	0.600
exclusive	bbox_best_iou_mean	849	0.451	0.356	0.500	0.046	0.782
inclusive_with_both	tfidf_cosine	1930	0.390	0.256	0.422	0.125	0.587
inclusive_with_both	char_cosine	1930	0.467	0.228	0.496	0.269	0.642
inclusive_with_both	rouge_1_f1	1930	0.380	0.103	0.357	0.317	0.417
inclusive_with_both	unigram_f1	1930	0.457	0.146	0.444	0.348	0.558
inclusive_with_both	topic_jaccard	1930	0.600	0.235	0.600	0.400	0.750
inclusive_with_both	bbox_best_iou_mean	1930	0.758	0.361	1.000	0.536	1.000

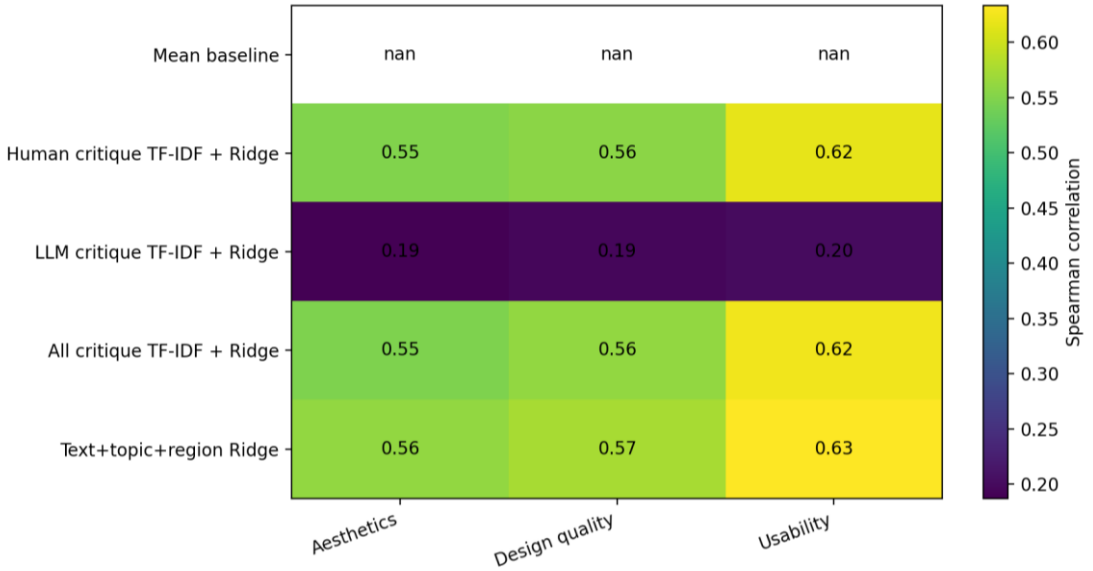

Figure 3. Spearman Rank Correlations Across Core Rating Targets

The qualitative error cases in Table 11 show why semantic alignment alone is insufficient. In the high-alignment case, the human and LLM texts were identical enough to produce TF-IDF cosine = 1.000 and IoU = 1.000. In the semantic-miss case, the human critique addressed missing search support in a region list, while the LLM critique focused on unrelated storage-space text; both semantic and region overlap were near zero. In the region-miss case, the topics partially

overlapped but the boxes did not. These cases support the quantitative conclusion: a design critic must align topic, language, and location at the same time.

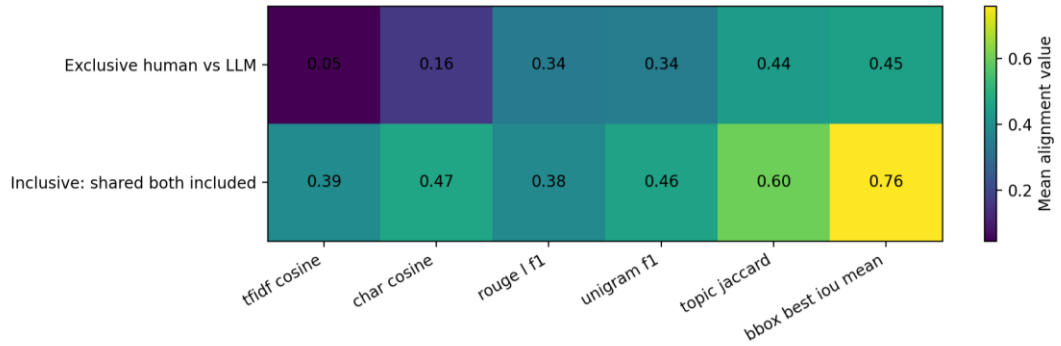

Figure 4. Mean Semantic, Lexical, Topical, and Spatial Alignment Between Human and LLM Critiques

Table 7. Correlations Between Inclusive Human-LLM Alignment and Human Ratings

metric	target	Spearman	N
tfidf_cosine	aesthetics_rating	0.122	1930
tfidf_cosine	usability_rating	0.049	1930
tfidf_cosine	design_quality_rating	0.098	1930
char_cosine	aesthetics_rating	0.119	1930
char_cosine	usability_rating	0.046	1930
char_cosine	design_quality_rating	0.099	1930
rouge_l_f1	aesthetics_rating	0.100	1930
rouge_l_f1	usability_rating	0.115	1930
rouge_l_f1	design_quality_rating	0.124	1930
unigram_f1	aesthetics_rating	0.162	1930
unigram_f1	usability_rating	0.154	1930
unigram_f1	design_quality_rating	0.180	1930
topic_jaccard	aesthetics_rating	0.074	1930
topic_jaccard	usability_rating	0.008	1930
topic_jaccard	design_quality_rating	0.056	1930
bbox_best_iou_mean	aesthetics_rating	-0.036	1930
bbox_best_iou_mean	usability_rating	-0.130	1930
bbox_best_iou_mean	design_quality_rating	-0.082	1930

DISCUSSION

The results support the central claim that AI design critique should be evaluated against human graphic design judgment, not only against text fluency. The strongest evidence is the contrast between human critique text and LLM critique text. Human-inclusive critique language predicted aesthetics, usability, design quality, learnability, and efficiency with Spearman correlations above 0.548 for all five targets. LLM-inclusive critique text produced correlations near 0.18 to 0.21. This means that human critique language carries structured information about quality judgments that the LLM source in the dataset does not reproduce at the same level when used alone.

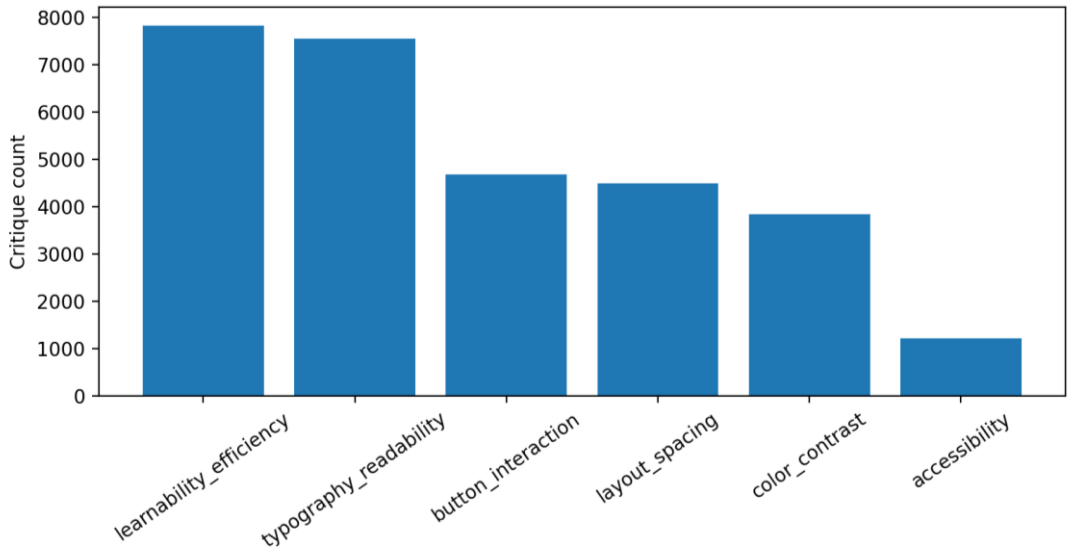

Figure 5. Critique Topic Coverage Across the UICrit Public CSV

Table 8. Lexicon-Based Topic Coverage in Source-Indexed Critiques

Topic	critique_count	share_of_critiques
learnability_efficiency	7837	0.691
typography_readability	7553	0.666
button_interaction	4692	0.414
layout_spacing	4500	0.397
color_contrast	3850	0.339
accessibility	1222	0.108

The finding does not mean that LLM critique is useless. It means that LLM critique should be treated as a candidate explanation whose agreement with human judgment must be measured. The LLM-inclusive model used comments tagged llm or both, yet its rating correlations stayed low when compared with human-inclusive and all-critique models. This pattern points to a calibration gap: the LLM can name design principles, but it does not consistently weight them in the way human raters do. A practical critic must therefore learn priority, severity, and location, not only design vocabulary.

Table 9. Design-Quality Association of Critique Topics

Topic	N_with_topic	mean_with_topic	mean_without_topic	difference_with_minus_without
layout_spacing	2313	5.750	6.045	-0.295
color_contrast	2142	5.752	5.981	-0.229
typography_readability	2703	5.795	6.025	-0.231
button_interaction	2388	5.760	6.040	-0.280
learnability_efficiency	2799	5.798	6.088	-0.289
accessibility	984	5.754	5.847	-0.093

The fused model improved rating prediction because it treated critique as both language and localized visual evidence. Text captures the semantics of the design problem, while region

features capture how much of the interface is affected and where the issue appears. A critique of a tiny icon, a header, and a full-screen layout imbalance can use similar vocabulary but imply different design severity. The region features provided a compact representation of this difference. The best design-quality result therefore came from combining all critique text with topic and region features rather than relying on language alone.

Table 10. Bounding-Box Region Distribution by Critique Source

Source	Region	n	mean_area
both	bottom	263	0.085
both	middle	1018	0.495
both	top	612	0.089
human	bottom	1487	0.074
human	middle	3123	0.415
human	top	3783	0.050
llm	bottom	150	0.067
llm	middle	566	0.478
llm	top	342	0.082

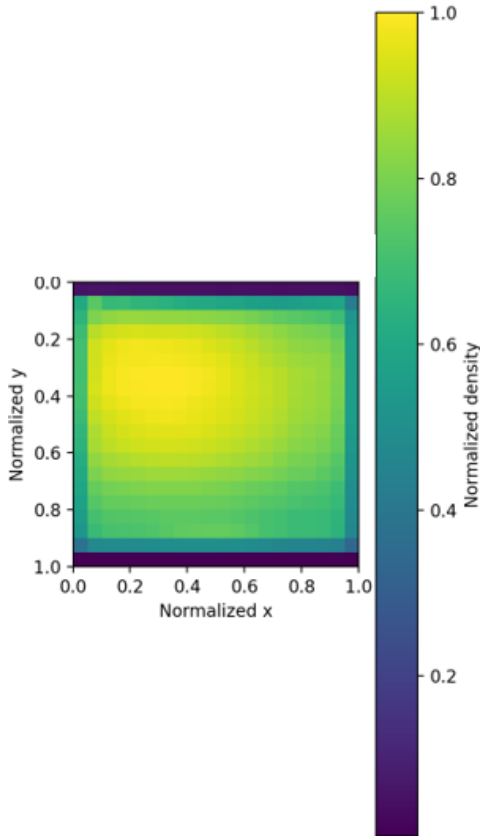

Figure 6. Human Critique Bounding-Box Density Over Normalized Mobile Screens

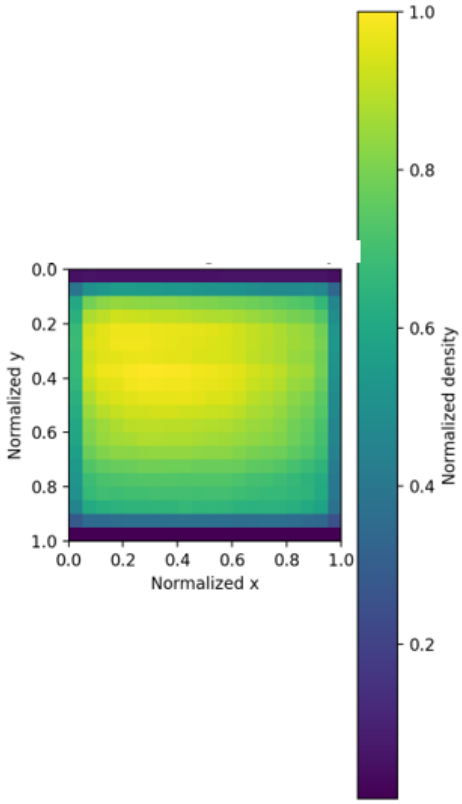

Figure 7. LLM Critique Bounding-Box Density Over Normalized Mobile Screens

The alignment findings also clarify the role of comments tagged both. Including shared comments substantially increases semantic and spatial similarity. That is expected because both-labeled comments represent overlapping human and Gemini observations. The strict exclusive comparison is more diagnostic for independent AI critique: TF-IDF cosine of 0.046 and best-

match IoU of 0.451 show that LLM comments often target related design themes without matching the human evaluator's wording or exact region. This is the practical risk in AI design feedback: an LLM can sound plausible while emphasizing a different element than the one a designer would prioritize.

Table 11. Qualitative Alignment and Error Cases

case_type	rico_id	Design_quality_rating	Tfidf_cosine	Topic_jaccard	Bbox_best_iou_mean	human_snippet	llm_snippet
high_alignment	4128	7	1.000	1.000	1.000	The expected standard is clear and concise headings with proper spacing between words. In…	The expected standard is clear and concise headings with proper spacing between words. In…
semantic_miss	59465	6	0.003	0.000	0.000	The expected standard is to provide convenient features, such as a search button, to enha…	The expected standard is that every element should have some connection to another elemen…
region_miss	24760	7	0.052	0.667	0.000	The expected standard is that The uneven spacing between these elements looks odd. Use si…	The expected standard is that the design should use as few elements as possible to achiev…

For visual communication, the topic results are interpretable. Typography/readability and learnability/efficiency dominate because mobile UIs are compact and task oriented. Layout/spacing and button/interaction are associated with lower design-quality means because these critiques point to violations that affect both aesthetics and use. Color/contrast appears frequently but is not the sole driver of rating prediction. This matters for AI critique tools because generic color advice is easy for LLMs to generate, while hierarchy, grouping, and interaction path are more consequential and more difficult to localize.

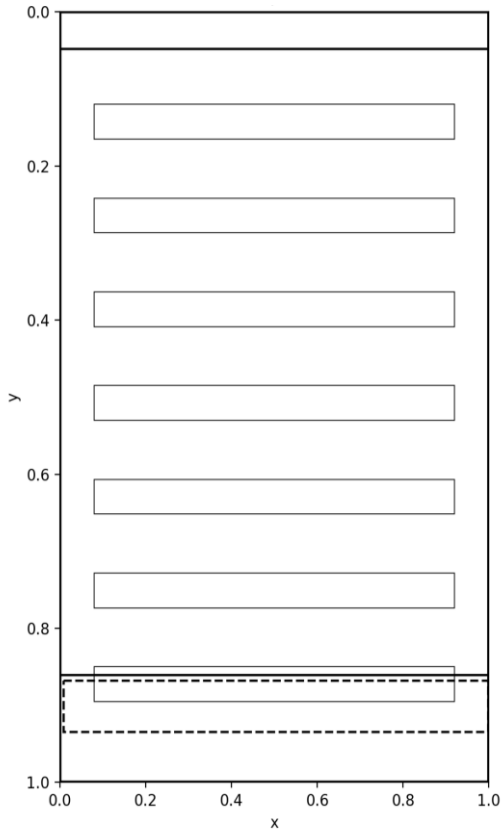

Figure 8. Error-Case Bounding Boxes on a Normalized Mock Screen; Solid Boxes Are Human-Inclusive and Dashed Boxes Are LLM-Inclusive

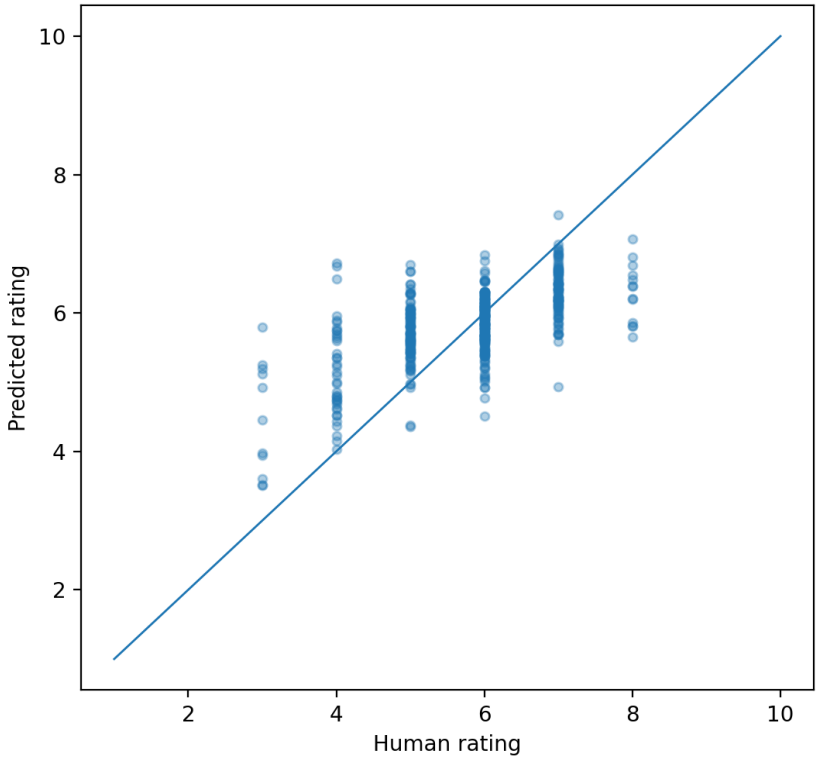

Figure 9. Held-Out Overall Design-Quality Predictions from the Text-Topic-Region Ridge Model

The paper also establishes a reproducible baseline that is useful for future multimodal work. A stronger system should ingest screenshots, layout hierarchies, and comments together, but it should still be compared against the text-topic-region results here. If a VLM or CLIP-style evaluator cannot outperform these grounded critique baselines, it is not yet adding reliable design judgment. Conversely, a model that improves rating prediction while preserving human-like topic and region alignment would provide stronger evidence of human-aligned AI design critique.

For design education and production workflows, these findings support a human-in-the-loop role for AI critique. The measured LLM comments are not reliable enough to replace expert evaluation, but they can still assist by surfacing candidate issues, prompting reflection, and documenting alternative interpretations. The decisive requirement is calibration. A tool should expose its confidence, show the region that motivated each suggestion, and distinguish high-priority critique from generic advice. The experiment shows that human-aligned critique is not a single score; it is a structured relation among visual evidence, design principle, location, and rating consequence.

This study has four limitations. First, the public CSV used here contains rico_id values and normalized bounding boxes but not the raw screenshots. The experiments therefore do not evaluate screenshot-level CLIP, UIClip, or VLM prompting. The manuscript avoids image-level claims for that reason. Second, the topic categories are lexicon-based diagnostics rather than hand-labeled issue classes. They are reproducible and useful for association analysis, but they do not replace expert coding of design critique themes. Third, the prediction task uses critique text as input. In a deployed critique generator, the critique would be produced from an unseen screen, so rating prediction from critique text measures alignment evidence rather than a complete automated UI-review pipeline. Fourth, the held-out evaluation uses one deterministic group split. It prevents screen leakage and is fully reproducible, but future benchmark papers should add repeated group splits or external datasets to estimate variance.

The alignment metrics also have limits. TF-IDF cosine, ROUGE-L, unigram F1, and topic Jaccard measure lexical or topical overlap, not full design reasoning. Bounding-box IoU measures spatial overlap but cannot determine whether the chosen region is semantically correct without the screenshot. These limitations do not invalidate the findings; they define the level of evidence supported by the available data. The study deliberately reports only what the dataset and code measure.

Another limitation is that comments tagged both create an optimistic alignment condition. Inclusive metrics are valuable because they represent shared design observations, but they also contain overlap by definition. The exclusive condition is therefore the stricter indicator of

independent LLM-human agreement. The manuscript reports both conditions so that the alignment claim is not inflated by shared-source comments.

CONCLUSION

This paper conducted a reproducible empirical evaluation of AI UI critique alignment with human graphic design judgment using the UICrit public CSV. The results show that critique text and localized region features predict human design ratings, with the text-topic-region Ridge model achieving the strongest overall design-quality performance. Human critique text is substantially more predictive than LLM-inclusive critique text, and exclusive human-LLM comments show low semantic similarity despite partial topic and region overlap. These findings demonstrate that LLM-as-design-critic systems require evaluation against human priorities, not only fluent feedback generation. The accompanying document, code, dataset files, tables, figures, and Illustrator-ready SVG diagrams provide a complete reproducible package for further development of human-aligned AI design critique. The most important methodological lesson is that future AI design critics should be judged by three simultaneous criteria: they must predict human quality ratings, explain the relevant design principle, and point to the same visual region a human designer would inspect.

REFERENCES

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An Empirical Evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction*, *24*(6), 574–594. https://doi.org/10.1080/10447310802205776

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*, 2–11. https://doi.org/10.1609/hcomp.v7i1.5280

Binghua Zhou, Siming Zhao, & David Chao. (2023). LLM-Guided Energy-Aware A/B Testing for Consolidation and DVFS Policies via Power-Sensitivity Clustering. *Journal of Advanced Computing Systems*, *3*(4), 12–30. https://doi.org/10.69987/jacs.2023.30402

Brooke, J. (1996). SUS: A Quick and Dirty Usability Scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability Evaluation in Industry*, 189–194. https://userinterfaces.aalto.fi/sus/sus.pdf

Card, S. K., Moran, T. P., & Newell, A. (1983). *The Psychology of Human-Computer Interaction*. https://doi.org/10.1201/9780203736166

Daren Zheng, Chenyu Li, & Harvey Davidson. (2023). Continual Red-Teaming for In-the-Wild Jailbreaks via Online Guardrail Updates and Guardrail Distillation. *Journal of Advanced Computing Systems*, *3*(2), 35–49. https://doi.org/10.69987/jacs.2023.30203

Deka, B., Huang, Z., Franzen, C., Hibschman, J., Afergan, D., Li, Y., Nichols, J., & Kumar, R. (2017). RICO: A Mobile APP Dataset for Building Data-Driven Design Applications.

Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology, 845–854. https://doi.org/10.1145/3126594.3126651

Duan, P., Chen, C.-Y., Li, G., Hartmann, B., & Li, Y. (2024). UICrit: Enhancing Automated Design Evaluation with a UI Critique Dataset. *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, *46*, 1-17. https://doi.org/10.1145/3654777.3676381

Heer, J., & Bostock, M. (2010). Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 203–212. https://doi.org/10.1145/1753326.1753357

Jason Kuhn, Yushan Chen, & Evelyn Chan. (2024). AI-Driven Mobile UI Pattern Recognition and Design Topic Mining on RICO: Semantic Clustering and Screenshot-Based Topic Classification. *Journal of Advanced Computing Systems*, *4*(5), 67–83. https://doi.org/10.69987/jacs.2024.40506

Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology* (4th ed.). https://doi.org/10.4135/9781071878781

Lewis, J. R. (2018). Measuring Perceived Usability: The CSUQ, SUS, and UMUX. *International Journal of Human–Computer Interaction*, *34*(12), 1148–1156. https://doi.org/10.1080/10447318.2017.1418805

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Measuring Perceived Usability: The SUS, UMUX-LITE, and AltUsability. *International Journal of Human–Computer Interaction*, *31*(8), 496–505. https://doi.org/10.1080/10447318.2015.1064654

Li, G., Baechler, G., Tragut, M., & Li, Y. (2022). Learning to Denoise Raw Mobile UI Layouts for Improving Datasets at Scale. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, *67*, 1-13. https://doi.org/10.1145/3491102.3502042

Liu, T. F., Craft, M., Situ, J., Yumer, E., Mech, R., & Kumar, R. (2018). Learning Design Semantics for Mobile Apps. *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, 569–579. https://doi.org/10.1145/3242587.3242650

Lund, A. M. (2001). Measuring Usability with the USE Questionnaire. *Usability Interface*, *8*(2), 3–6. https://search.worldcat.org/title/818903534

Moshagen, M., & Thielsch, M. T. (2010). Facets of Visual Aesthetics. *International Journal of Human–Computer Studies*, *68*(10), 689–709. https://doi.org/10.1016/j.ijhcs.2010.05.006

Ngo, D. C. L., Teo, L. S., & Byrne, J. G. (2003). Modelling Interface Aesthetics. *Information Sciences*, *152*, 25–37. https://doi.org/10.1016/s0020-0255(02)00404-8

Nielsen, J. (1994). Enhancing the Explanatory Power of Usability Heuristics. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 152–158. https://doi.org/10.1145/191666.191729

Nielsen, J., & Molich, R. (1990). Heuristic Evaluation of User Interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 249–256. https://doi.org/10.1145/97243.97281

Norman, D. A. (2013). *The Design of Everyday Things* (Rev. and expanded ed.). https://jnd.org/the-design-of-everyday-things-revised-and-expanded-edition/

O'Donovan, P., Agarwala, A., & Hertzmann, A. (2015). DesignScape: Design with Interactive Layout Suggestions. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1221–1224. https://doi.org/10.1145/2702123.2702149

Oulasvirta, A., Kristensson, P. O., Bi, X., & Howes, A. (Eds.). (2018). *Computational Interaction*. https://doi.org/10.1093/oso/9780198799658.001.0001

Petrova, S., & Watanabe, K. (2025). User-Centered Mobile Navigation: Evaluating Local Usability for Improved UX. *Journal of Technology Informatics and Engineering*, *4*(3), 478–492. https://doi.org/10.51903/jtie.v4i3.457

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models from Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

Reinecke, K., & Gajos, K. Z. (2014). Quantifying Visual Preferences Around the World. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 11–20. https://doi.org/10.1145/2556288.2557052

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Saranya, K. N., Bhandari, M., Borad, N., Reddy, P. V. P., & Kumar, S. (2025). Surveying the Impact of Rarely Investigated Design Components on User Engagement. *International Journal of Graphic Design*, *3*(1), 39–52. https://doi.org/10.51903/ijgd.v3i1.2752

Sauro, J., & Lewis, J. R. (2016). *Quantifying the User Experience: Practical Statistics for User Research* (2nd ed.). https://doi.org/10.1016/c2010-0-65191-4

Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N., & Diakopoulos, N. (2016). *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (6th ed.). https://www.pearson.com/en-us/subject-catalog/p/designing-the-user-interface-strategies-for-effective-human-computer-interaction/p200000003255

Swearngin, A., & Li, Y. (2019). Modeling Mobile Interface Tappability Using Crowdsourcing and Deep Learning. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, *78*, 1-11. https://doi.org/10.1145/3290605.3300305

Wang, B., Li, G., Zhou, X., Chen, Z., Grossman, T., & Li, Y. (2021). Screen2Words: Automatic Mobile UI Summarization with Multimodal Learning. *Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology*, 498–510. https://doi.org/10.1145/3472749.3474765

Wu, J., Peng, Y.-H., Li, A. X. Y., Swearngin, A., Bigham, J. P., & Nichols, J. (2024). UIClip: A Data-Driven Model for Assessing User Interface Design. *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, *45*, 1-16. https://doi.org/10.1145/3654777.3676408

Wu, J., Zhang, X., Nichols, J., & Bigham, J. P. (2021). Screen Parsing: Towards Reverse Engineering of UI Models from Screenshots. *Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology*, 470–483. https://doi.org/10.1145/3472749.3474763

Yunianto, I., & Wahyudi, W. (2024). Designing User Experience for a Mobile Application for Agricultural Product Marketing Using the Human-Centered Design Method. *International Journal of Graphic Design*, *2*(2), 207–221. https://doi.org/10.51903/ijgd.v2i2.2123

Yushan Chen, & Evelyn Chan. (2023). Multimodal UI Representation Learning: Ablation of Screenshot, Wireframe, and View-Hierarchy Proxies on an Uploaded 168-Screen Dataset. *Journal of Advanced Computing Systems*, *3*(1), 1–15. https://doi.org/10.69987/jacs.2023.30101

Zhang, X., Ross, A. S., & Fogarty, J. (2018). Robust Annotation of Mobile Application Interfaces in Methods for Accessibility Repair and Enhancement. *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, 609–621. https://doi.org/10.1145/3242587.3242616