International Journal of Graphic Design (IJGD)
Volume 3 No. 1 May 2025
E-ISSN: 2987-9434; P-ISSN : 2988-0343, Page 196-213
DOI: https://doi.org/10.51903/ijgd.v3i1.3723
Available online at: https://journal.stekom.ac.id/index.php/ijgd

LLM-Compatible Visual Brief Cards for AI Infrastructure Capacity Dashboards: A UI/UX Framework for Turning Forecast Risk into Graphic Design Decisions

Ge Liu[1], Shilu He*[2], Hailey Wong[3]
[1]Computer Science, USC, CA, USA
[2]Mathematics, UW-Madison, WI, USA
[3]Design and Technology, Parsons School of Design, NY, USA
Email Address: gliusde@gmail.com

Abstract. *AI infrastructure dashboards frequently condense GPU inventory, demand forecasts, uncertainty, and admission-control policies into dense operational reports, making it difficult for planners to identify critical risks and appropriate design actions. This study introduces VB-Card, an LLM-compatible UI/UX framework that converts capacity-forecast evidence into structured visual brief cards containing risk indicators, capacity summaries, uncertainty cues, inventory evidence, policy explanations, and actionable recommendations. To evaluate the framework independently of vendor-specific model variation, a deterministic visual-brief generator was employed. The evaluation integrates the OpenCOLE graphic-design field contract with the Alibaba PAI GPU cluster trace. GPU demand, inventory, rolling P50/P90 forecasts, uncertainty measures, and stock-out probabilities were derived from 1,037,084 GPU-requesting task records and 1,897 machine specification records, then paired with 23,419 OpenCOLE instances. Five approaches were compared: a baseline text report, a generic visual card, an OpenCOLE keyword brief, a risk-badge card, and VB-Card. On the held-out test set, VB-Card achieved the highest overall score (0.820; 95% bootstrap CI: 0.820–0.821), outperforming the baseline text report (0.633), OpenCOLE keyword brief (0.522), and risk-badge card (0.538). Ablation results show that uncertainty, policy context, inventory evidence, and visual hierarchy each contribute to performance. The findings demonstrate that capacity risk can be translated into consistent, data-grounded visual-brief decisions, although real-world UI effectiveness requires user-based validation.*

Keywords : *Capacity Dashboard, GPU Infrastructure, Information Visualization, Risk Communication, UI/UX.*

INTRODUCTION

Capacity dashboards for AI infrastructure are now read by people who must make decisions under uncertainty: platform operators, model-training owners, SRE teams, finance planners, and executives who approve accelerator purchases. A GPU-capacity forecast can estimate P50 and P90 demand, but a dashboard still has to answer a design question: what should be visually emphasized so the user immediately understands the risk, the stock-out probability, the inventory gap, and the policy action? Traditional reports often place this information in paragraphs or tables. Such reports can be semantically complete while still failing as visual communication because they do not define hierarchy, grouping, perceptual salience, or explanation placement.

This paper treats the dashboard card as a graphic-design object rather than a passive container for prediction results. The research question is whether forecast risk, GPU inventory, uncertainty, and policy explanation can be transformed into a compact visual brief that is aligned with OpenCOLE-style graphic design fields. OpenCOLE is relevant because it formalizes automatic graphic design generation around intention, description, keywords, background mood, object

Received: Januari 2025; Revised: February 2025; Accepted: April 2025; Published: May 2025
*Corresponding author, gliusde@gmail.com

captions, and textual headings (CyberAgent, 2024; Inoue et al., 2024). These are the kinds of structures a designer, renderer, or LLM-enabled tool needs when turning a technical risk signal into a visually organized UI card.

The problem is especially visible in AI infrastructure because the same card must satisfy multiple audiences. An on-call engineer wants to know whether SLO workloads are protected, a research lead wants to know whether training jobs should be delayed, and a finance planner wants to know whether the inventory signal justifies procurement. A single textual score cannot serve all of these interpretations. The interface must organize evidence so that severity, evidence, and action are visible in a predictable order. VB-Card therefore treats the visual brief as a translation layer between technical operations and design communication.

The proposed framework uses a capacity-risk policy to choose a severity label, color token, heading, subheading, uncertainty visual, inventory visual, policy note, and action footer. It does not optimize a forecasting model. Instead, it makes the later design step explicit: it converts forecast evidence into design decisions. This distinction is important for UI/UX research because the design failure of a dashboard often emerges after prediction, when the interface communicates the result too weakly, overstates certainty, or leaves policy rationale outside the user's visual field. The evaluation combines two data roles. OpenCOLE supplies the design-brief schema and split structure, while the Alibaba PAI GPU trace supplies the capacity variables. The PAI trace contains production MLaaS GPU workload information from a large heterogeneous cluster and provides task-level GPU requests as well as machine-level GPU capacity (Alibaba Cluster Trace Program, 2021; Weng et al., 2022). This pairing allows the study to evaluate visual-brief outputs against real infrastructure evidence rather than using capacity quantities that exist only as template variables.

The contributions are threefold. First, the paper defines a visual-brief card framework for AI infrastructure dashboards that links forecast quantities to graphic-design components. Second, it reconstructs the capacity-risk benchmark from PAI GPU trace tables while preserving OpenCOLE's visual-brief field structure. Third, it revises the evaluation so that evidence accuracy, risk-policy agreement, uncertainty communication, action coverage, and layout consistency are measured separately. The result is a UI/UX and graphic design contribution, not a claim about forecasting superiority or deployed operator performance.

LITERATURE REVIEW

Information visualization research has long shown that visual encodings determine whether people can compare quantities, detect outliers, and make decisions under time pressure. Graphical perception studies established that position, length, and aligned marks support more accurate comparisons than less ordered visual channels (Cleveland & McGill, 1984). Dashboard design

literature similarly stresses that the interface must guide attention toward the few variables that matter at the moment of action (Few, 2006). A capacity dashboard is therefore not a neutral surface. If a P90 shortfall, a probability of stock-out, and an emergency policy are all presented with equal visual weight, the interface has already made a poor design choice.

Visual analytics also emphasizes interaction and sensemaking. Shneiderman's (1996) visual information seeking mantra and Heer and Shneiderman's (2012) interaction taxonomy describe how users move from overview to detail and from pattern to action. In a compact card, the available space is too small for complex interaction, so the static visual hierarchy becomes more important. VB-Card adapts this principle by assigning the strongest hierarchy to risk badge and capacity heading, placing numerical evidence in the subheading, using the forecast ribbon and inventory bar as middle-layer evidence, and placing the policy action in a footer. The layout is designed to support recognition before deliberation.

Graphic design literature gives the same argument a compositional vocabulary. Contrast, repetition, alignment, and proximity are not merely aesthetic rules; they are mechanisms for grouping evidence and reducing ambiguity (Graham, 2002; Williams, 2014). Tufte (1983) argued that quantitative displays should preserve data integrity while removing non-informative clutter, and Ware (2012) showed that perception depends on preattentive organization. For a capacity-risk card, this means that a red badge, a P90-versus-inventory line, a confidence ribbon, and a policy footer should not be scattered. They should form a coherent reading path.

Risk communication and cognitive load research provide the second foundation. Cognitive load theory argues that working memory is limited and that instructional materials should reduce extraneous load (Sweller, 1988). Miller's (1956) classic account of memory limits is often oversimplified, but it remains useful as a warning against overloading a card with too many ungrouped tokens. In design terms, risk information should be chunked into badge, metric, explanation, and action. Norman (2013) also argues that good design reveals affordances and mappings: the interface should make the next action visible without requiring the user to infer policy from raw numbers.

Human-centered AI (Chen & Chan, 2023) literature adds a third requirement: explanations must support action and accountability, not merely describe a model. Guidelines for human-AI interaction recommend that systems show confidence, explain what they can and cannot do, and support efficient correction (Amershi et al., 2019). Model cards and datasheets similarly argue for structured documentation of assumptions, intended use, and limitations (Gebru et al., 2021; Mitchell et al., 2019). In a capacity dashboard, the policy explanation is the interface-level analog of documentation. It states why the system recommends freezing admissions, reserving burst capacity, or maintaining normal allocation.

Work on interpretable machine learning also warns that explanation must be evaluated in relation to use, not as a purely technical label (Doshi-Velez & Kim, 2017; Lipton, 2018; Ribeiro et al., 2016). A policy note that is accurate but visually detached from the risk badge may not help a planner. For this reason, VB-Card measures action coverage and layout consistency separately. The method can fail semantically if it omits policy language, and it can fail visually if the policy is not represented as a component in the card.

Finally, the paper builds on automatic graphic design generation. Layout generation (Kuhn et al., 2024) research has modeled how text, images, and graphic elements can be placed under design constraints (Li et al., 2021; O'Donovan et al., 2014). More recent systems use language and multimodal models to decompose a vague design intention into layers, captions, and editable design objects (Jia et al., 2024). OpenCOLE extends this direction by making automatic graphic design generation more reproducible with public resources (Inoue et al., 2024). The present work does not attempt to generate full poster imagery. Instead, it uses the same design-brief logic for a narrower UI/UX problem: translating complex infrastructure risk into card-level visual decisions.

The gap is therefore specific. Existing dashboard research explains visual hierarchy, and AI explanation research explains why model output should be documented. Graphic design generation research explains how intentions can become visual structures. Yet little work connects all three for AI infrastructure capacity risk, where a single card must combine forecast uncertainty, GPU inventory, severity, and policy. VB-Card addresses that gap by treating capacity risk as a design brief generation problem.

METHODS

A. Dataset contract and capacity trace

The empirical object is a capacity-card benchmark that combines an OpenCOLE visual-brief field contract with the Alibaba PAI GPU trace. OpenCOLE provides the design fields and split sizes: 23,419 rows in total, with 19,093 training rows, 1,951 validation rows, and 2,375 test rows. The fields used are id, intention, description, keywords, captions_background, captions_objects, headings_heading, and headings_sub_heading. These fields define what a graphic should communicate, which visual objects appear, what background mood is intended, and what headings or subheadings should be rendered.

The capacity variables come from the PAI trace rather than from OpenCOLE. The PAI trace is a production MLaaS GPU-cluster trace from Alibaba PAI. The tables used here are the task table, job table, group-tag table, and machine specification table. The task table supplies task launch time, instance count, requested GPU percentage, and GPU type. The machine specification table supplies GPU type and physical GPU capacity. The group-tag table supplies workload labels when available. Table 1 lists the data sources and roles used in the revised benchmark.

Table 1. Data sources and roles in the revised benchmark

Data source	Rows or units used	Relevant fields	Role in this study
OpenCOLE field files	23,419 rows; train 19,093; validation 1,951; test 2,375	id, intention, description, keywords, captions_background, captions_objects, headings_heading, headings_sub_heading	Visual-brief schema and split structure
PAI job table	1,055,501 rows	job_name, status, start_time, end_time	Job context and scheduling state
PAI task table	1,037,084 GPU-requesting rows after filtering	task_name, inst_num, start_time, plan_gpu, gpu_type	Admission-demand calculation by GPU type and time window
PAI machine specification	1,897 rows	machine, gpu_type, cap_gpu	Inventory by GPU type
PAI group-tag table	1,055,032 rows	workload, gpu_type_spec	Workload labels for brief wording

B. Capacity-risk construction

GPU inventory was computed by summing cap_gpu in the machine specification table for each GPU type. The resulting inventory values were 2,240 MISC GPUs, 1,596 P100 GPUs, 994 T4 GPUs, 832 V100 GPUs, and 1,080 V100M32 GPUs. Requested GPU demand for each task was computed as inst_num multiplied by plan_gpu divided by 100, following the trace definition that plan_gpu is expressed as a percentage of one GPU. Demand was then summed within two-hour admission windows for each GPU type.

Table 2. Capacity variables derived from PAI trace and mapped into visual-brief fields

Capacity Variable	Source	Computation	Visual-Brief Usage
GPU pool and region	PAI gpu_type; deterministic time-window region label	GPU type from task and machine tables; region label groups card comparisons	Heading and intention
Available inventory	PAI machine specification	Sum of cap_gpu by gpu_type	Inventory object and subheading
P50 and P90 demand	PAI task table	Rolling quantiles of two-hour requested GPU demand	Subheading and forecast ribbon
Uncertainty width	PAI task table	(P90 - P10) / inventory over preceding 12 windows	Forecast uncertainty ribbon
Stock-out probability	PAI task and machine tables	Share of preceding windows where demand exceeded inventory	Probability chip and policy note
Workload label	PAI group-tag table and task role names	Semantic tag when present; task role fallback	Keywords and policy explanation
Risk level and color	Risk policy	Thresholds applied to P90 gap, P50 utilization, and stock-out probability	Risk badge and background accent
Recommended action	Risk policy	Action matched to severity class	Action footer and policy note

For every window, P50, P90, and P10 demand were computed from the preceding 12 windows, corresponding to the previous 24 hours. Stock-out probability was the fraction of those 12 prior windows in which launched demand exceeded inventory. Uncertainty was represented

as (P90 - P10) divided by inventory. Table 2 shows how these derived variables enter the visual-brief fields.

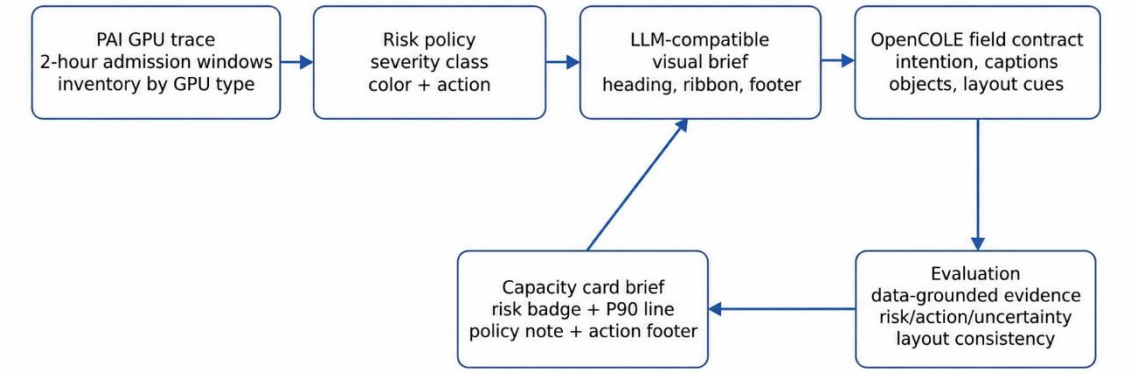

Figure 1. VB-Card pipeline connecting PAI capacity evidence, risk policy, LLM-compatible visual brief fields, OpenCOLE-style design structure, and evaluation

C. Risk policy

Risk level was computed from P90 demand, available inventory, P50 utilization, and stock-out probability. Critical risk was assigned when the P90 gap exceeded 14% of inventory or stock-out probability exceeded 0.70. High risk was assigned when the P90 gap exceeded 2% or stock-out probability exceeded 0.50. Medium risk was assigned when P50 utilization exceeded 0.40 or stock-out probability exceeded 0.30. Low risk was assigned otherwise. The 0.40 medium threshold was used because the trace windows represent admission demand rather than continuous GPU utilization; a moderate admission load can still warrant monitoring before the pool approaches a shortage. Table 3 gives the visual encoding policy.

Table 3. Visual encoding policy for risk levels

Risk Level	Color Token	Threshold Pattern	Action Footer
Low	green	P90 demand within inventory and stock-out probability <= .30	Maintain standard allocation
Medium	amber	P50 admission demand > .40 of inventory or stock-out probability > .30	Monitor and queue non-urgent jobs
High	orange	P90 demand near/above inventory or stock-out probability > .50	Rebalance and reserve capacity
Critical	red	P90 gap > 14% or stock-out probability > .70	Escalate and protect SLO workloads

D. Visual brief representation

The visual brief representation is intentionally narrow. A row does not ask the generator to draw a finished dashboard screenshot; it asks the generator to decide which design primitives must appear. This mirrors a common editorial workflow in graphic design: a brief specifies tone, hierarchy, objects, and copy before a designer or renderer produces the final artifact.

For capacity dashboards, the primitives were chosen because each corresponds to a user question: the badge answers how severe the risk is, the heading answers where and what pool is

affected, the subheading answers how large the gap is, the ribbon answers how uncertain the forecast is, the inventory bar answers how the forecast compares with capacity, the policy note answers why action is recommended, and the footer answers what to do next. Figures 1, 2, and 3 summarize this flow, card structure, and risk matrix.

Figure 2. Card mockup showing the risk badge, heading hierarchy, forecast ribbon, inventory object, policy note, and action footer

E. Role of LLMs

The system is described as LLM-compatible rather than as a direct comparison of LLM outputs. The generator used in the benchmark is deterministic and produces a structured visual brief that can be passed to an LLM, UI renderer, or designer. This choice isolates the value of the visual-brief contract and avoids confounding the results with stochastic sampling, prompt drift, or proprietary model changes. The study therefore evaluates whether a capacity-risk policy can be expressed as a reliable visual-brief structure, not whether one commercial LLM is better than another.

Badge	Color token	Threshold pattern	Action footer
LOW	green	P90 within inventory; stock-out <= .30	**Maintain standard allocation**
MEDIUM	amber	P50 utilization > .40 or stock-out watch-list	**Monitor and queue non-urgent jobs**
HIGH	orange	P90 near/above inventory or stock-out > .50	**Rebalance and reserve capacity**
CRITICAL	red	P90 gap > 14% or stock-out > .70	**Escalate and protect SLO workloads**

Figure 3. Risk badge design matrix used by the visual brief generator

F. Compared methods

Five methods were evaluated on the held-out test rows. The baseline text report writes an accurate paragraph but does not define a visual hierarchy. The generic visual card produces a neutral card without risk-specific evidence. The OpenCOLE keyword brief uses lexical cues from the design fields but does not enforce risk color or policy mapping. The risk-badge card displays risk severity and color but omits much of the uncertainty and policy rationale. The proposed VB-Card includes the full component set. Table 4 defines the comparison.

Table 4. Experimental methods

Method	Main Output	Expected Strength	Expected Weakness
Baseline text report	Paragraph with forecast, inventory, risk, probability, and action	High data-grounded content	No card hierarchy or risk color
Generic visual card	Neutral card with status language	Simple UI structure	Weak risk specificity and weak evidence
OpenCOLE keyword brief	Keyword-selected heading and cue list	Keyword preservation	No enforced color-policy design
Risk-badge card	Risk heading, severity color, compact inventory	Strong badge and color validity	Weak policy and uncertainty coverage
Proposed VB-Card	Badge, heading, P90 subheading, ribbon, inventory bar, policy note, action footer	Integrated technical evidence and visual hierarchy	Requires structured forecast inputs

G. Evaluation metrics

The revised metrics separate data-grounded correctness from structural completeness. Semantic alignment is token-F1 between generated card text and the capacity-card reference text. Keyword F1 compares generated wording with the reference technical keywords. Evidence accuracy checks whether the P50 demand, P90 demand, inventory, and stock-out probability values appear in the output. Risk accuracy checks whether the risk class appears. Color validity checks whether the output color matches the risk policy.

Table 5. Evaluation metrics and operational tests

Metric	Operational Definition	Weight in Overall Score
Semantic alignment	Token-F1 between generated card text and reference capacity-card brief	0.15
Keyword F1	Token-F1 between generated wording and reference technical keywords	0.10
Evidence accuracy	P50, P90, inventory, and stock-out probability values appear	0.20
Risk accuracy	Risk label is present in generated output	0.15
Color validity	Generated color equals risk color	0.10
Action coverage	Recommended action and policy rationale are represented	0.10
Uncertainty coverage	P90, uncertainty/ribbon, and stock-out probability are represented	0.10
Layout consistency	Required components and hierarchy order pass	0.05
Readability	Normalized Flesch-style score for card text	0.05

Action coverage checks whether the recommended action and policy rationale appear. Uncertainty coverage checks whether P90, uncertainty, and stock-out probability are represented. Layout consistency checks whether the required visual-brief components and their order are present. Readability uses a normalized Flesch-style score. Table 5 gives the weights used in the overall score.

Table 6. Dataset Descriptive Statistics by Split

Split	Rows	Avg. Intention Words	Avg. Keywords	Critical Rate	High Rate	Median Inventory	Median P90	Median Stock-Out	Median Uncertainty
test	2,375	33.07	15.0	18.0%	8.1%	1,080	846	0.00	0.518
train	19,093	33.08	15.0	18.0%	8.1%	1,080	848	0.00	0.518
validation	1,951	33.09	15.0	18.0%	8.0%	1,080	829	0.00	0.510

H. Statistical procedure

The held-out test split was the primary evaluation set. Mean scores were computed for each method. Overall-score confidence intervals used 300 bootstrap resamples of test rows. Paired differences were computed by matching each method's output on the same test row and subtracting baseline scores from the proposed VB-Card score. Ablation tests were conducted on the test split by removing one design component at a time.

Table 7. Risk-Level Distribution by Split

Split	Low	Medium	High	Critical
train	10,282 (53.9%)	3,828 (20.0%)	1,539 (8.1%)	3,444 (18.0%)
validation	1,051 (53.9%)	391 (20.0%)	157 (8.0%)	352 (18.0%)
test	1,279 (53.9%)	476 (20.0%)	192 (8.1%)	428 (18.0%)

RESULTS

A. Dataset checks

The revised benchmark preserved the OpenCOLE split sizes while replacing template-only capacity variables with PAI-derived capacity evidence. Table 6 shows that the test split contained 2,375 rows, an average capacity-card intention length of 33.07 content words, and 15 technical keywords per row. The median available inventory was 1,080 GPUs and the median P90 admission demand was 846 GPUs on the test split. Critical-risk cases represented 18.0% of the test set, high-risk cases 8.1%, medium-risk cases 20.0%, and low-risk cases 53.9%, as shown in Table 7 and Figure 4. These proportions provide positive and negative cases for severity-sensitive design decisions without making the evaluation depend on a single risk state.

B. Main comparison

Table 8 reports the full test-set comparison and Figure 5 shows the same overall-score pattern visually. The proposed VB-Card achieved the highest overall score, 0.820. Its strongest advantages were evidence accuracy, risk accuracy, color validity, action coverage, uncertainty

coverage, and layout consistency. The baseline text report scored 0.633 overall because it contained the numerical evidence, risk label, probability, and action, but had no risk-color encoding and weak visual hierarchy. The OpenCOLE keyword brief preserved some lexical evidence but did not consistently bind that evidence to risk color or policy rationale. The risk-badge card achieved perfect color validity but did not represent the uncertainty ribbon or policy rationale sufficiently.

Table 8. Main Test-Set Metrics by Method

Method	Sem.	Key.	Evid.	Risk	Color	Action	Uncert.	Layout	Read.	Overall [95% CI]
Baseline text report	0.398	0.249	1.000	1.000	0.000	1.000	0.667	0.275	0.353	0.633 [0.632, 0.633]
Generic visual card	0.130	0.176	0.024	0.043	0.000	0.000	0.000	0.220	0.416	0.080 [0.079, 0.082]
OpenCOLE keyword brief	0.227	0.320	0.532	1.000	0.147	0.502	1.000	0.481	0.209	0.522 [0.520, 0.523]
Risk-badge card	0.173	0.217	0.530	1.000	1.000	0.500	0.333	0.580	0.434	0.538 [0.537, 0.538]
Proposed VB-Card	0.551	0.218	1.000	1.000	1.000	1.000	1.000	1.000	0.320	0.820 [0.820, 0.821]

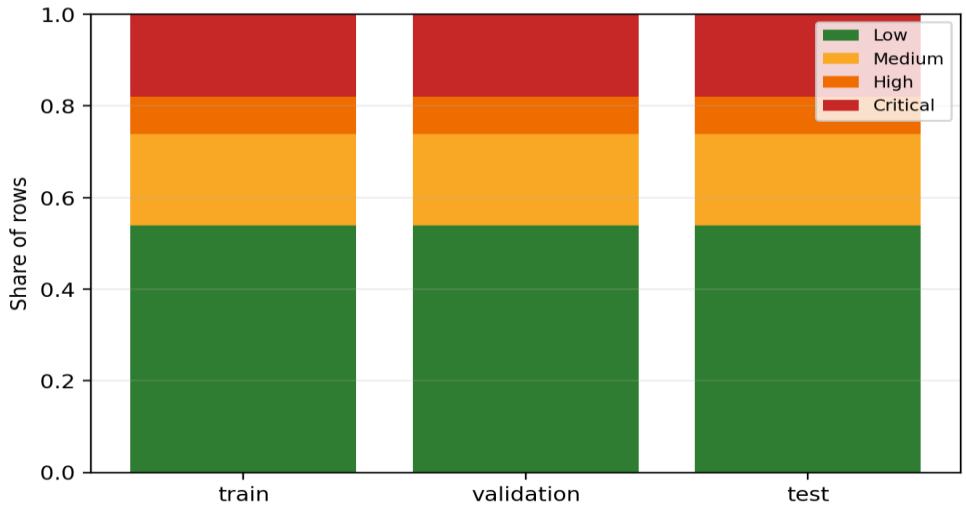

Figure 4. Risk-level distribution by split

The metric pattern separates text fidelity from interface structure. The text report is not a weak content baseline: it contains the same operational facts that a planner would expect in a capacity review. Its lower score comes from the absence of a graphic design structure. Conversely, the keyword brief preserves many terms but does not decide which terms should become a badge, a

ribbon, or a footer. The proposed method scores higher because it transforms facts into visual roles while preserving the PAI-derived evidence.

Table 9. Paired Overall-Score Differences on the Test Split

Comparison	Mean Difference	95% CI Low	95% CI High
Proposed minus Baseline text report	0.188	0.188	0.188
Proposed minus Generic visual card	0.740	0.739	0.742
Proposed minus OpenCOLE keyword brief	0.299	0.297	0.301
Proposed minus Risk-badge card	0.283	0.282	0.283

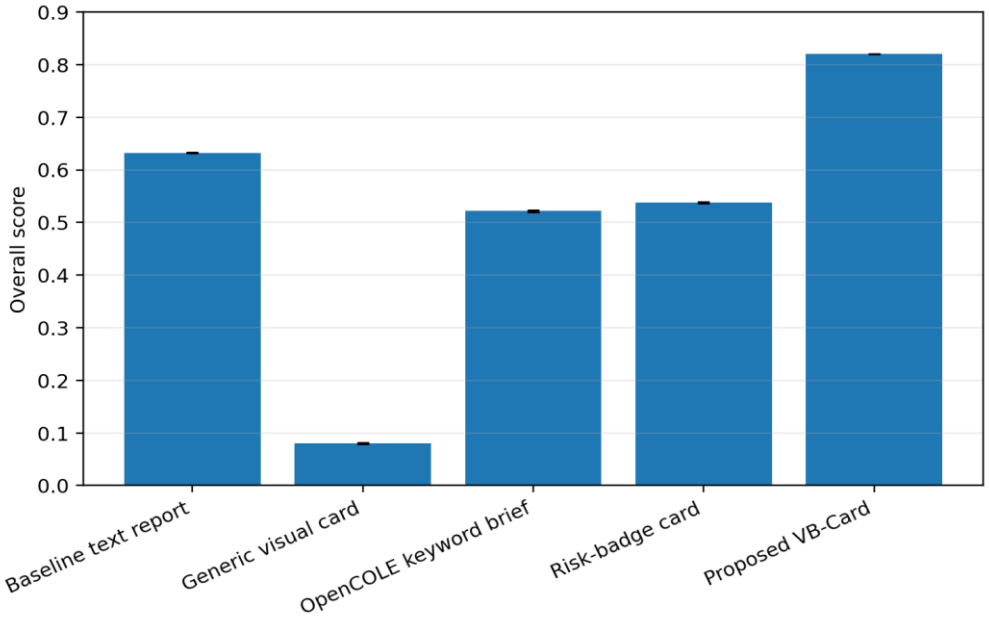

Figure 5. Test-set overall score by method with bootstrap confidence intervals

C. Ablation study

The ablation results identify which design components drive the improvement. Table 11 and Figure 8 show that removing the uncertainty ribbon reduced overall score from 0.820 to 0.725 because the output no longer represented the full uncertainty evidence. Removing policy explanation reduced the score to 0.752 because action coverage fell even when risk and color were correct. Removing the inventory object produced a smaller reduction to 0.816, and removing hierarchy constraints reduced the score to 0.814. These results support the claim that the framework works by combining evidence, uncertainty, policy explanation, and visual hierarchy rather than relying on a risk badge alone.

Table 10. Overall Score by Risk Class

Method	Low	Medium	High	Critical
Baseline text report	0.630	0.636	0.631	0.637
Generic visual card	0.091	0.068	0.068	0.068
OpenCOLE keyword brief	0.536	0.500	0.500	0.513
Risk-badge card	0.543	0.534	0.524	0.534
Proposed VB-Card	0.819	0.823	0.819	0.823

D. Risk-class performance

The improvement was consistent across low, medium, high, and critical rows. Table 10 pivots the overall scores by risk class, and Figure 7 shows the same pattern visually. VB-Card scored from 0.819 to 0.823 across the four severity classes. This matters because a capacity dashboard must remain interpretable during routine periods as well as during high-pressure shortages.

Table 11. Ablation Results on the Test Split

Ablation	Sem.	Key.	Evid.	Action	Uncert.	Layout	Read.	Overall
Full VB-Card	0.551	0.218	1.000	1.000	1.000	1.000	0.320	0.820
No uncertainty ribbon	0.481	0.200	0.780	1.000	0.667	0.900	0.308	0.725
No policy explanation	0.431	0.239	1.000	0.500	1.000	0.900	0.360	0.752
No inventory object	0.513	0.246	1.000	1.000	1.000	0.980	0.318	0.816
No hierarchy constraints	0.551	0.218	1.000	1.000	1.000	0.880	0.320	0.814

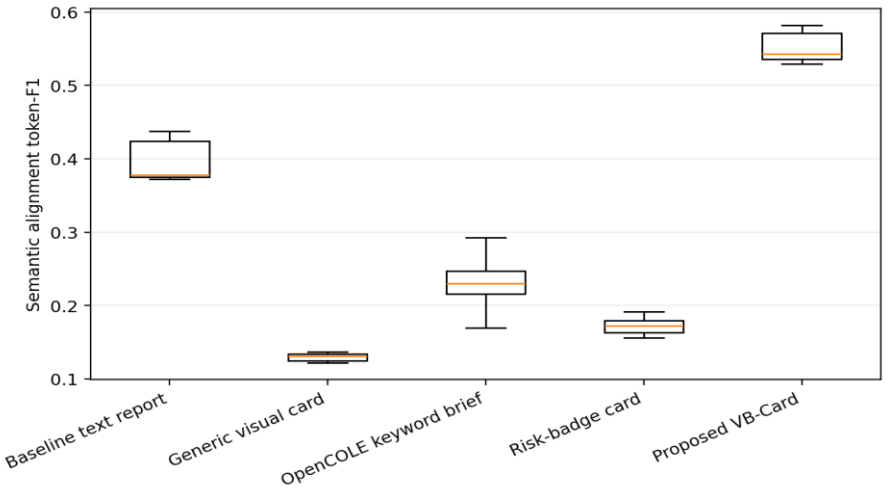

Figure 6. Distribution of semantic alignment across test rows

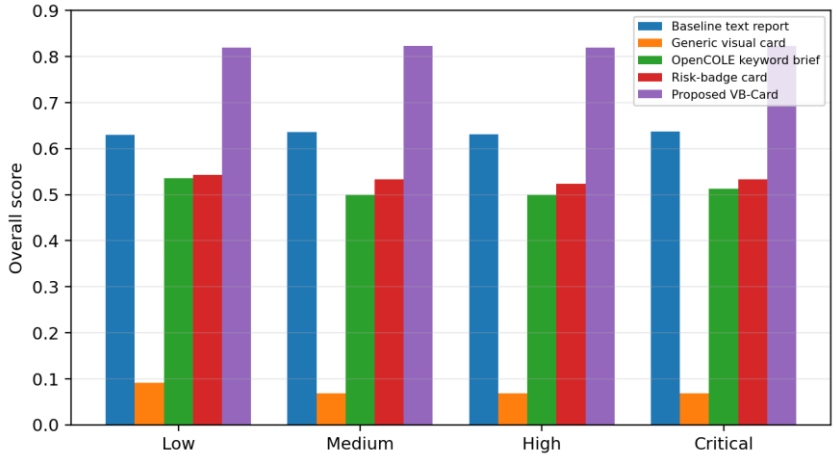

Figure 7. Overall score by risk class

E. Design-rule audit

A final audit checked whether the proposed method satisfied the card-level design requirements on the test set. Table 12 shows a 100% pass rate for risk badge, risk color, capacity

evidence, forecast uncertainty, policy rationale, and layout constraints. This audit is not the full evaluation; it is a targeted check that the proposed method did not drop the visual components that the framework requires.

Table 12. Design-Rule Audit for the Proposed VB-Card

Rule	Operational Test	Target Method	Pass Rate
Risk badge present	risk badge phrase appears	Proposed VB-Card	100.0%
Risk color valid	color token matches risk policy	Proposed VB-Card	100.0%
Capacity evidence present	P50, P90, inventory, and probability values appear	Proposed VB-Card	100.0%
Forecast uncertainty shown	P90, uncertainty, and stock-out probability represented	Proposed VB-Card	100.0%
Policy rationale shown	action and policy note represented	Proposed VB-Card	100.0%
Layout constraints met	required visual-brief components and order pass	Proposed VB-Card	100.0%

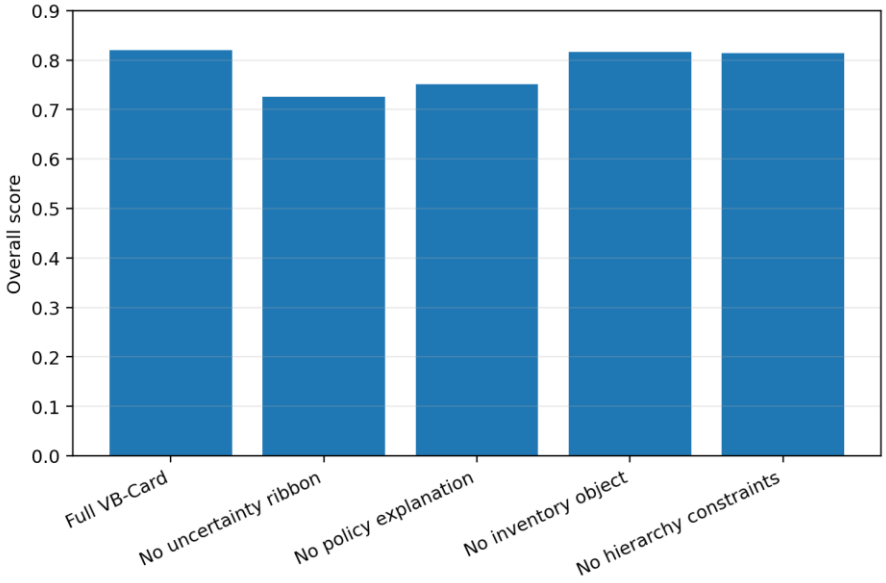

Figure 8. Ablation of VB-Card design components

DISCUSSION

The results show that visual communication of AI infrastructure capacity risk is not equivalent to textual completeness. The baseline text report included accurate quantities, a risk label, a probability, and an action. It therefore performed well on evidence, risk, action, and uncertainty coverage, yet it lost on color validity and layout consistency. This is the central UI/UX point of the study: a technically correct report can still be a weak interface artifact if it does not guide the user's eye or encode severity through a stable visual system.

The OpenCOLE keyword brief produced a different failure mode. It preserved many terms from the reference fields and therefore achieved useful keyword overlap, but it did not bind keywords to the graphic roles that make a card actionable. A keyword such as critical risk is useful only when it becomes a badge, heading, color token, and policy action. This distinction explains

why the keyword brief scored below the proposed framework despite preserving lexical evidence. It generated vocabulary, whereas VB-Card generated a design decision structure.

The risk-badge card confirms that severity encoding alone is also insufficient. It received perfect risk accuracy and color validity, but its uncertainty coverage and policy explanation were weak. In infrastructure operations, such a card can create a false sense of clarity: the user sees that risk is high or critical, but not why the risk is uncertain or which policy should be triggered. The ablation results reinforce this interpretation. Removing the uncertainty ribbon and policy note caused the largest drops, showing that explanation and uncertainty are not decorative additions. They are core visual components for responsible dashboard communication.

The graphic design implication is that LLM-compatible dashboard generation should not simply ask a model to summarize the forecast. The prompt or tool should require visual components, hierarchy roles, color tokens, and policy rationale. OpenCOLE-style fields are useful because they separate intention, description, keywords, background mood, objects, and headings. That separation allows a technical signal to be turned into a brief that a designer, renderer, or LLM-enabled workflow can operationalize. VB-Card demonstrates a domain-specific version of this principle for AI capacity planning.

The human-centered AI implication is that explanation must be embedded in the interface state. If a planner sees a critical badge but must search elsewhere for the policy rationale, the system increases cognitive load and weakens accountability. VB-Card places explanation inside the card, next to the forecast evidence and action footer. This design aligns with human-AI guidelines that recommend showing confidence, enabling understanding, and supporting user control (Amershi et al., 2019). It also parallels model documentation practices, but at the level of operational interface design (Gebru et al., 2021; Mitchell et al., 2019).

For journal positioning, the contribution belongs to UI/UX, visual communication, and AI-compatible graphic design rather than pure capacity forecasting. The forecasting variables are inputs; the research object is the transformation of those variables into visual hierarchy and semantic alignment. The empirical results support the claim that a structured visual brief can preserve technical meaning and improve design consistency. They do not, by themselves, prove faster human decisions or higher operator trust.

The results also suggest practical design guidance. First, every capacity-risk card should reserve a fixed location for severity so users can compare cards quickly across GPU pools. Second, uncertainty should be drawn as a first-class object rather than hidden in a caption, because uncertainty changes how strongly a policy action should be interpreted. Third, the policy explanation should be adjacent to the action footer so that action and rationale are read together.

Fourth, color should be constrained by a documented risk policy, not selected freely by a generative model.

For graphic design scholarship, the study contributes a domain case in which visual style and operational semantics are inseparable. A card can be visually clean but still misleading if the hierarchy suppresses uncertainty or disconnects policy from action. Conversely, a card can be technically complete but visually unusable if every detail competes for attention. The VB-Card framework offers a way to operationalize design critique with measurable fields, making it possible to compare design decisions without reducing the work to aesthetics alone.

Limitations

The first limitation is that the benchmark evaluates a text-and-structure representation of UI cards rather than a deployed dashboard used by human operators. The metrics are useful for reproducible comparison, but they are proxy measures. A future study should add human-subject evaluation with time-to-diagnosis, action selection accuracy, trust calibration, perceived workload, and qualitative interviews with capacity planners. The second limitation is that OpenCOLE is used as a design-brief field contract rather than as a source of infrastructure data. The practical capacity variables come from the PAI GPU trace, while OpenCOLE supplies the fields used to express the brief. This distinction is important: the study should not be interpreted as an evaluation of OpenCOLE poster-generation quality or as evidence that OpenCOLE itself contains capacity-dashboard cases.

The third limitation concerns the LLM component. The benchmark uses a deterministic visual-brief generator that is compatible with LLM workflows, but it does not compare commercial LLM vendors or stochastic prompt variants. This choice improves control over the experiment but underrepresents stylistic diversity and natural-language variation. A follow-up study can compare unconstrained LLM briefs, constrained LLM briefs, and deterministic VB-Card briefs on the same PAI-derived capacity cases.

The fourth limitation concerns metric design. The weighted overall score reflects a specific UI/UX theory: semantic alignment, keyword preservation, evidence accuracy, risk accuracy, color validity, action coverage, uncertainty coverage, layout consistency, and readability. Different organizations may prefer different weights. The separate metric columns are therefore more informative than the overall score alone. The fifth limitation is that the PAI trace is historical and anonymized. It is valuable because it is a real production GPU-cluster trace, but it does not expose every variable found in a modern infrastructure dashboard, such as procurement lead time, business priority, reservation contracts, or human override decisions. The present benchmark should be read as a capacity-risk communication study rather than a full operational simulator.

CONCLUSION

This paper presented VB-Card, a UI/UX framework for translating AI infrastructure capacity risk into LLM-compatible visual brief cards. The framework converts P50/P90 demand, GPU inventory, uncertainty, stock-out probability, and policy rationale into OpenCOLE-style design fields: intention, description, keywords, background mood, object captions, headings, and subheadings. The revised benchmark combines the OpenCOLE field contract with the Alibaba PAI GPU trace, so the capacity variables used in the evaluation come from production GPU workload and machine-capacity data. In a 23,419-row capacity-card benchmark with a 2,375-row held-out test split, the proposed method achieved the best test-set overall score, 0.820, and outperformed a text report, a generic card, a keyword brief, and a risk-badge card. Paired differences and ablations showed that the improvement came from integrating uncertainty, policy explanation, inventory evidence, and visual hierarchy. The study's broader conclusion is that AI capacity dashboards need design decisions, not only accurate forecasts. A card that visually binds risk, uncertainty, inventory, and policy can make a complex technical state easier to represent consistently; measuring whether it also improves human decisions requires a subsequent user study.

Author's CrediT

The study was conceived and developed by G.L. and S.H., who contributed equally to the work. G.L. led the technical implementation, experimental evaluation, and manuscript preparation. S.H. supervised the research, validated the methodology, coordinated the project, and contributed to manuscript refinement. H.W. supported the study through visualization design, user-centered interpretation, literature review, and manuscript editing. All authors have approved the final manuscript and agree to be accountable for the integrity of the work. G.L. and S.H. share equal contribution as co-first authors. S.H. is the corresponding author.

REFERENCES

Alibaba Cluster Trace Program. (2021). cluster-trace-gpu-v2020 [Data set]. https://github.com/alibaba/clusterdata/tree/master/cluster-trace-gpu-v2020

Amershi, S., Weld, D., Voros, K., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-13). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300233

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (Eds.). (1999). Readings in information visualization: Using vision to think. Morgan Kaufmann.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. Journal of the American Statistical Association, 79(387), 531-554. https://doi.org/10.1080/01621459.1984.10478080

CyberAgent. (2024). cyberagent/opencole [Data set]. Hugging Face. https://huggingface.co/datasets/cyberagent/opencole

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. https://arxiv.org/abs/1702.08608

Few, S. (2006). Information dashboard design: The effective visual communication of data. O'Reilly Media.

Few, S. (2009). Now you see it: Simple visualization techniques for quantitative analysis. Analytics Press.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daume III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92. https://doi.org/10.1145/3458723

Graham, L. (2002). Basics of design: Layout and typography for beginners. Delmar Cengage Learning.

Heer, J., & Shneiderman, B. (2012). Interactive dynamics for visual analysis. Communications of the ACM, 55(4), 45-54. https://doi.org/10.1145/2133806.2133821

Inoue, N., Masui, K., Shimoda, W., & Yamaguchi, K. (2024). OpenCOLE: Towards reproducible automatic graphic design generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.

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

Jia, P., Li, C., Yuan, Y., Liu, Z., Shen, Y., Chen, B., Chen, X., Zheng, Y., Chen, D., Li, J., Xie, X., Zhang, S., & Guo, B. (2024). COLE: A hierarchical generation framework for multi-layered and editable graphic design. arXiv. https://arxiv.org/abs/2311.16974

Knaflic, C. N. (2015). Storytelling with data: A data visualization guide for business professionals. Wiley.

Kong, W., Jiang, Z., Sun, S., Zhang, Z., & Liu, T. (2023). Aesthetics++: Refining graphic designs by exploring design principles and human preference. IEEE Transactions on Visualization and Computer Graphics, 29(6), 3060-3074. https://doi.org/10.1109/TVCG.2022.3151617

Li, J., Yang, J., Zhang, J., Liu, C., Wang, C., & Xu, T. (2021). Attribute-conditioned layout GAN for automatic graphic design. IEEE Transactions on Visualization and Computer Graphics, 27(10), 4039-4048. https://doi.org/10.1109/TVCG.2020.2999335

Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36-43. https://doi.org/10.1145/3233231

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63(2), 81-97. https://doi.org/10.1037/h0043158

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220-229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596

Munzner, T. (2014). Visualization analysis and design. CRC Press.

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 152-158). Association for Computing Machinery. https://doi.org/10.1145/191666.191729

Norman, D. A. (2013). The design of everyday things (Revised and expanded ed.). Basic Books.

O'Donovan, P., Agarwala, A., & Hertzmann, A. (2014). Learning layouts for single-page graphic designs. IEEE Transactions on Visualization and Computer Graphics, 20(8), 1200-1213. https://doi.org/10.1109/TVCG.2014.48

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). https://doi.org/10.1145/2939672.2939778

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In Proceedings 1996 IEEE Symposium on Visual Languages (pp. 336-343). IEEE. https://doi.org/10.1109/VL.1996.545307

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257-285. https://doi.org/10.1207/s15516709cog1202_4

Tufte, E. R. (1983). The visual display of quantitative information. Graphics Press.

Ware, C. (2012). Information visualization: Perception for design (3rd ed.). Morgan Kaufmann.

Weng, Q., Xiao, W., Yu, Y., Wang, W., Wang, C., He, J., Li, Y., Zhang, L., Lin, W., & Ding, Y. (2022). MLaaS in the Wild: Workload Analysis and Scheduling in Large-Scale Heterogeneous GPU Clusters. In Proceedings of the 19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22) (pp. 945-960). USENIX Association.

Williams, R. (2014). The non-designer's design book (4th ed.). Peachpit Press.

Yi, J. S., Kang, Y. A., Stasko, J., & Jacko, J. (2007). Toward a deeper understanding of the role of interaction in information visualization. IEEE Transactions on Visualization and Computer Graphics, 13(6), 1224-1231. https://doi.org/10.1109/TVCG.2007.70515

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