

Visual Performance Evaluation Model Construction and Generative AI Prompt Template Design for OA Form Interfaces in Smart Office Scenarios

Ying Liu, Zhigang Hu

Shaanxi University of Science and Technology, Xi'an, Shaanxi, China

Abstract: OA form interfaces serve as critical nodes for human-computer interaction in smart office scenarios, and their visual design quality directly affects user operational efficiency and cognitive experience. Addressing the gaps in existing evaluation research—namely the lack of systematic weight derivation and operational scoring criteria tailored to the B-end scenario of OA forms, and the absence of executable visual design standards for generative AI tools—this study integrates literature review, vendor case analysis, questionnaire survey (n=386), and user interviews (n=12) to construct a visual performance evaluation system for OA form interfaces comprising 5 dimensions and 15 indicators. The Analytic Hierarchy Process (AHP) is employed to derive indicator weights: Information Architecture (0.309) > Layout Rationality (0.234) > Visual Clarity (0.215) > Interaction Usability (0.168) > Aesthetic Consistency (0.074). Pairwise comparison weights exhibit significant ranking differences relative to direct scoring, revealing that information architecture and layout constitute the structural priority dimensions of visual performance. Furthermore, graded checklist scoring criteria are developed for the 15 secondary indicators, and a Visual Performance Evaluation (VPE) linear weighted model is constructed (inter-rater reliability ICC = 0.87). The checklist criteria are transformed into 22 prompt parameters through direct mapping, combinatorial mapping, and constraint mapping, forming a standardized AI prompt template. A controlled experiment (n=54) demonstrates that using the template increases the mean VPE score of AI-generated interfaces by 27.8% ($p < 0.001$), with more pronounced improvements in visual dimensions whose parameters exhibit higher executability. This study provides an operational quantitative evaluation tool for OA system interface

design optimization and offers a methodological reference for translating design evaluation knowledge into generative AI prompts.

Keywords: Smart Office; OA Form Interface; Visual Performance; Analytic Hierarchy Process; Checklist Scoring Criteria; AI-Assisted Design

1. Introduction

Against the backdrop of deepening digital transformation, smart office has evolved from traditional process electrification to an intelligent collaborative office model centered on user experience. According to iResearch's "2024 China Collaborative Office Platform Industry Research Report," the market size of China's collaborative office software has exceeded 33 billion yuan, and OA systems have achieved widespread adoption in enterprise-level organizations [1]. As the fundamental carrier for information collection, approval workflow, and data accumulation, the visual design quality of OA forms directly affects users' cognitive load, operational efficiency, and satisfaction. Taking the frequently used expense reimbursement form as an example, practical issues such as dense field arrangement, ambiguous hierarchical structure, and absent visual guidance often force users to repeatedly locate fields and verify hierarchical relationships during form filling, significantly increasing cognitive load—particularly in high-information-density forms. However, notable deficiencies persist in the evaluation practice of OA form interface design. Existing studies predominantly focus on usability evaluation of general Web interfaces or mobile interaction design, with scarce research targeting visual performance evaluation specifically for OA forms. In weight derivation, existing B-end interface evaluation studies largely adopt expert direct scoring or empirical weighting, lacking structured constraints in the

judgment process and thus difficult to ensuring weight consistency and reproducibility. Meanwhile, generative AI tools have deeply penetrated the interface design workflow, yet their output quality is highly dependent on prompt executability and professional standards, and design evaluation knowledge has not yet been standardized into instructions directly consumable by AI tools. Visual performance originates from the field of human factors engineering; introduced into the domain of OA form interface design, this study defines it as "the visual cognitive efficiency and operational accuracy with which users complete information recognition, understanding, and filling tasks in an OA form interface."

Building on the above analysis, this study aims to construct a visual performance evaluation model for OA form scenarios with reproducible weights and operational scoring, and to translate it into executable AI prompt templates. This goal entails four key challenges: first, the indicator system must be systematically constructed to cover dimensions of visual cognition, information architecture, interaction, layout, and aesthetics; second, weight derivation must incorporate structured constraints and consistency checks to ensure reliable judgment logic; third, scoring criteria must be operationalized into independently checkable specification items to reduce subjective arbitrariness; and fourth, evaluation knowledge must be translated into parameterized instructions executable by AI tools, achieving a closed loop from "evaluation criteria to generation instructions."

Addressing these challenges, this study adopts the Analytic Hierarchy Process (AHP) as the unified weight derivation method: AHP organizes decision elements in a hierarchical structure, transforms expert judgments into quantifiable relative weights through pairwise comparisons, and ensures reliable judgment logic through embedded consistency ratio testing ($CR < 0.10$) [2], which has been widely applied in information systems and interface evaluation. The specific approach is as follows: a 5-dimension, 15-indicator evaluation system is constructed through triangulation of literature, vendor cases, questionnaires, and interviews; AHP is used to derive primary and secondary indicator weights; graded checklist scoring criteria are developed for the 15 secondary indicators; a VPE linear weighted evaluation

model is constructed and its discrimination and reliability are validated; finally, the checklist criteria are translated into a standardized AI prompt template through three mapping mechanisms, and its effectiveness is tested through a controlled experiment.

The main contributions of this study are as follows: first, a visual performance evaluation indicator system and graded checklist scoring criteria for OA form scenarios are constructed, providing an operational quantitative evaluation tool for B-end form interfaces; second, AHP is employed to derive indicator weights, revealing the ranking differences between AHP weights and independent weighting and the underlying weighting mechanism correction effect; third, a VPE linear weighted evaluation model is constructed, validating its discrimination (inter-form grade difference of 1.17 points) and inter-rater reliability ($ICC = 0.87$); fourth, a translation path from "evaluation criteria to prompt parameters" is established, a five-segment standardized AI prompt template is designed, and its effectiveness in improving the visual performance of AI-generated interfaces is verified through a controlled experiment.

2. Related Work

2.1 Research Status of Visual Performance

Visual performance research originates from human factors engineering. Sanders and McCormick systematically elaborated the relationship between performance indicators such as visual search and interface design [3]. Chen et al. analyzed the impact of logistics APP navigation layout on visual search performance through eye-tracking experiments [4]. Yang et al. employed the analytic hierarchy process to construct a visual saliency evaluation system for decision support system interfaces [5]. Wu's team conducted a series of visual cognitive performance experiments: eye-tracking demonstrated that the fixation/saccade ratio can effectively reflect the information search efficiency of interface layout [6], and pupil dilation amplitude can serve as a sensitive measure of visual cognitive performance [7]. Color hierarchy coding experiments showed that hierarchical coding significantly reduces search time [8]. Building on this foundation, the team further constructed a comprehensive cognitive performance evaluation model for industrial manufacturing system information interaction

interfaces, establishing an evaluation indicator system from dimensions of visual search quality and information processing level, and achieving scheme optimization through AHP weighting combined with vague sets and TOPSIS. These studies established quantitative correlations between design features and cognitive performance, but their research objects concentrated on industrial monitoring and complex system interfaces, and have not yet been migrated to the B-end high-frequency interaction scenario of OA forms.

2.2 Research Status of Interface Evaluation Methods and Weight Derivation

Interface evaluation methods include heuristic evaluation, usability testing, and comprehensive evaluation models: Nielsen's ten usability principles represent a classic framework [9,10], and ISO 9241-11 defines usability from effectiveness, efficiency, and satisfaction [11], the Analytic Hierarchy Process (AHP) proposed by Saaty transforms expert pairwise comparison judgments into consistency-testable quantifiable weights, providing a structured scheme for multi-criteria decision-making, and its subsequent development ANP breaks through the same-layer independence assumption to handle indicator dependencies [12]. In terms of method integration and extension, Ou Yang et al. first proposed the integration method of DEMATEL and ANP (DANP), driving the supermatrix to solve networked weights through the total influence matrix [13-15], which has been successively applied to green building evaluation [16], mobile health application experience evaluation [17], and industrial chain leader enterprise evaluation [18]. Such methods are suitable for characterizing networked interactions among indicators, but entail high data collection and computational complexity. The checklist scoring criteria in this study have operationalized each indicator into independently checkable specification items, the indicator system presents a clear hierarchical structure, and expert pre-surveys indicate that cross-influences within the sub-criteria layer are sparse. Therefore, AHP with clear structure and easy expert consensus operation is adopted for weight derivation, while networked modeling is reserved for future expansion.

In the field of interface design evaluation, AHP and its combined methods have been widely applied: Benaïda integrated Nielsen's principles

with the USE questionnaire to form a comprehensive heuristic framework [19], Ouyang et al. evaluated in-vehicle interface aesthetics using AHP-CRITIC-WRSR [20], Ayağ and Samanlıoğlu constructed an ERP selection model based on AHP-TOPSIS [21]. Zhu et al. built a three-level user experience framework for mobile banking [22]. Additionally, domestic scholars have established AHP evaluation systems across multiple interface types, such as aesthetic image evaluation of human-machine interface morphological element layout [23], human-machine interface design evaluation for oilfield water injection system platforms [24], and human-factor suitability evaluation system for smartphone APP interaction interfaces [25]. These studies demonstrate the mature applicability of AHP in interface evaluation, but most stop at weight derivation and scheme ranking, failing to systematically combine visual performance evaluation with operational checklist scoring criteria, nor have they been migrated to B-end form scenarios.

2.3 OA System Interface Design and AI-Assisted Design

According to the iResearch report, the domestic collaborative office platform market presents a multi-competitive landscape of OA vendors, internet platforms, and ERP vendors, and OA systems have achieved widespread adoption in enterprise-level organizations [1]. In terms of evaluation research on B-end form interfaces, Wu et al. constructed a WeChat mini-program form usability evaluation system through eye-tracking experiments combined with the entropy weight method [26], providing a reference for interface evaluation in form scenarios, but their indicator system did not account for cross-influences among indicators. In terms of AI-assisted design, Liu and Chilton proposed at the CHI conference that the design quality of prompt engineering directly affects the output effectiveness of text-to-image models [27], providing a basis for the parameterized translation of design knowledge. Recent prompt engineering research further indicates that the structural degree and refinement strategy of prompts significantly affect generation quality: Goloujeh et al. revealed through semi-structured interviews the complete journey of users in writing, evaluating, and refining prompts and the alignment obstacles therein [28], systems such as

lasting 30–45 minutes, around three themes: information acquisition efficiency, operational fluency, and visual comfort.

3.2 Research Workflow and Technical Route

This study follows the technical route of "literature review—case analysis—survey and interview—AHP weight calculation—scoring criteria development—VPE model validation—AI prompt translation" (see Figure 2), with AHP as the unified weight calculation framework.

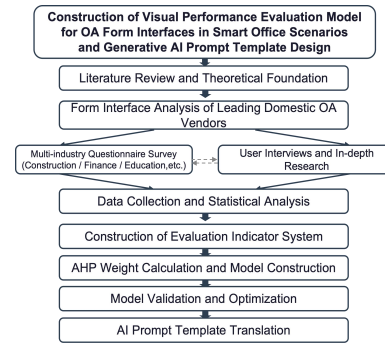


Figure 2. Research Technical Route

Table 1. Secondary Indicator Interpretation and Graded Scoring Benchmarks

Indicator	Interpretation	5-Point Benchmark (Optimal)	1-Point Baseline (Failure)
C11 Font Hierarchy Clarity	Differentiation and readability of font size hierarchy system	Hierarchy ≥ 3 levels (title $\geq 14\text{px}$ / body 12px / auxiliary 10px), level difference $\geq 2\text{px}$, clear font specification	No hierarchy distinction or only 1 level
C12 Color Contrast	Luminance contrast value between text and background	Body text $\geq 7:1$ (WCAG AAA), large title $\geq 4.5:1$	Body text $< 2:1$, difficult to read
C13 Icon Recognizability	Semantic clarity and independent recognizability of functional icons	Unified style, $\geq 24 \times 24\text{px}$, understandable without text labels	Icons missing, chaotic, or completely unrecognizable
C21 Information Grouping Logic	Whether field grouping conforms to user cognitive model	Conforms to mental model, high cohesion within groups and low coupling between groups, clear grouping titles and boundaries	No grouping or completely inconsistent with user cognition
C22 Label Description Clarity	Semantic accuracy and ambiguity-free degree of field labels	Concise labels (≤ 6 characters) without ambiguity, key fields with filling prompts	Labels missing or severely ambiguous
C23 Hierarchy Relationship Visualization	Clarity of visual expression of master-slave / parallel relationships	Multi-method (indentation + color blocks + dividing lines) comprehensive differentiation of master-slave and parallel	No hierarchy visualization, all flatly arranged
C31 Operation Path Simplicity	Number of filling steps and rationality of jumps	Single-column vertical linear layout, reasonable Tab order, steps $\leq$ field count + 1	Chaotic path, users frequently lost
C32 Feedback Timeliness	Speed and sufficiency of system response	Real-time validation ($< 200\text{ms}$), immediate status feedback, complete progress indication	Most operations without feedback or lag $> 5\text{s}$
C33 Error Prevention Effectiveness	Mechanism design for preventing input errors	Dedicated controls fully covered, global validation + submission prevention, error localization to field	No error prevention mechanism whatsoever
C41 Space Utilization	Rationality of space allocation in form area	Width $550\text{--}860\text{px}$, label ratio $40\%\text{--}45\%$, spacing $12\text{--}16\text{px}$, no crowding or blank space	Extremely unreasonable space allocation, severely affecting readability
C42 Visual Flow Guidance	Effectiveness of elements in guiding eye movement	Breadcrumb + section title + progress indicator three-level guidance, reasonable F/Z-type visual flow	No visual guidance whatsoever, relying entirely on experience
C43 Element Alignment Standard	Alignment precision of labels and controls in the grid	Based on 8px grid, labels / input boxes / buttons / dividing lines all aligned, unified grid	Chaotic alignment, no grid baseline
C51 Visual Style Unity	Consistency of global design language	Component style / corner radius / spacing / shadow / color completely unified	Multiple design languages mixed, chaotic style
C52 Color Coordination	Visual harmony degree of color scheme	Main colors ≤ 3 , accent colors $< 10\%$, color wheel coordination, cold-warm balance	Severe color conflict, visually uncomfortable
C53 Brand Recognizability	Recognition effect of brand elements in the interface	Logo + brand color + brand font + brand discourse four elements complete	No brand recognition elements whatsoever

4. Construction of Visual Performance Evaluation Indicator System for OA Form Interfaces

4.1 Indicator Extraction and Screening

This study extracts the indicator candidate set based on triangulation of literature, cases, and survey data. The primary indicators (C1–C5) are preliminarily determined through literature and vendor case pre-screening, namely Visual Clarity, Information Architecture, Interaction Usability, Layout Rationality, and Aesthetic Consistency. Subsequently, the same 8-member expert panel conducts two rounds of Delphi

screening for secondary indicators: the first-round interquartile range (IQR) mean is 1.72, decreasing to 0.95 in the second round with significant convergence; 6 candidate indicators with large attribution disputes (IQR > 2.0) are deleted, retaining 15 secondary indicators. In terms of qualitative data, visual feature coding of 12 form samples from 4 vendors covers five dimensions: information density, grouping method, navigation hierarchy, color quantity, and contrast; 12 user interviews yield 42 open codes and 12 axial codes through three-level coding, summarized as three core categories: "information acquisition efficiency, operational fluency, visual comfort,"

corresponding one-to-one with the 5 primary dimensions.

4.2 Indicator System Interpretation and Checklist Scoring Criteria

To ensure operational consistency in scoring, 1–5 point graded checklist scoring criteria are developed for the 15 secondary indicators, integrating objective parameters (contrast, font hierarchy, pixel spacing) with subjective judgment criteria. The checklist criteria design follows three principles: operability (each score level corresponds to clear boundaries), combination of subjective and objective criteria (quantitative thresholds and descriptive criteria used jointly), and reasonable gradients (graded based on specifications such as WCAG and 8px grid and expert consensus). Table 1 lists the interpretation, 5-point optimal benchmark, and

1-point failure baseline for each indicator; scores 4–2 are determined by continuous spectrum position.

5. Indicator Weight Calculation Based on AHP

5.1 AHP Method and Primary Indicator Judgment Matrix

Primary indicator weights are derived by 8 experts (3 visual communication design professors, 3 user experience designers, 2 OA product design directors) conducting pairwise comparisons of the 5 primary indicators according to Saaty's 1–9 scale. Individual judgments are aggregated into a group judgment matrix through the geometric mean method, weights are solved using the eigenvector method, and the consistency test yields $CR = 0.034 < 0.10$ (Table 2).

Table 2. Primary Indicator AHP Judgment Matrix and Weights

Criterion	C1 Visual Clarity	C2 Information Architecture	C3 Interaction Usability	C4 Layout Rationality	C5 Aesthetic Consistency	Weight
C1 Visual Clarity	1.00	0.72	1.42	0.70	3.61	0.215
C2 Information Architecture	1.39	1.00	1.56	1.62	4.62	0.309
C3 Interaction Usability	0.70	0.64	1.00	0.50	3.13	0.168
C4 Layout Rationality	1.43	0.62	2.00	1.00	1.84	0.234
C5 Aesthetic Consistency	0.28	0.22	0.32	0.54	1.00	0.074

Note: $\lambda_{\max} = 5.150$, $CR = 0.034 < 0.10$, passing the consistency test; weights are eigenvector normalization results. Secondary indicator judgment matrices and consistency testing procedures are detailed in Appendix A.

5.2 Comparison of AHP Weights and Independent Weighting

The derived primary indicator weights are: Information Architecture (0.309) > Layout Rationality (0.234) > Visual Clarity (0.215) > Interaction Usability (0.168) > Aesthetic Consistency (0.074). To examine the effect of the pairwise comparison mechanism, Table 3 presents AHP weights alongside independent weighting weights (expert direct scoring normalized) and ranking differences.

The ranking differences reveal that the weighting mechanism substantially corrects evaluation conclusions: Layout Rationality

jumps from independent weighting rank 4 (0.170) to AHP rank 2 (0.234). In direct scoring, Visual Clarity tends to receive high scores due to the intuitiveness of "what you see is what you get," while the value of layout and architecture can only fully manifest through pairwise trade-offs; Visual Clarity drops from rank 1 (0.260) to rank 3 (0.215), as pairwise comparison forces experts to make trade-offs between dimensions, correcting the anchoring bias of direct scoring. Information Architecture and Layout Rationality collectively account for 54.3% of the weight, constituting the structural priority dimensions of OA form visual performance.

Table 3. Comparison of AHP Weights and Independent Weighting

Criterion	AHP Weight	AHP Rank	Independent Weight	Independent Rank	Rank Change	Weight Factor Analysis
C1 Visual Clarity	0.215	3	0.260	1	↓2	Direct scoring affected by "visibility anchoring"—text and contrast are most intuitive, inflating independent judgment scores; pairwise comparison falls back to rank 3 through trade-offs
C2 Information Architecture	0.309	1	0.220	2	↑1	The priority of information organization criteria emerges in pairwise comparison, as the upstream dimension determining "what users see and in what order"

C3 Interaction Usability	0.168	4	0.200	3	↓1	Moderate importance, direct scoring slightly high; AHP rank decreases by 1
C4 Layout Rationality	0.234	2	0.170	4	↑2	Spatial skeleton criteria increases in priority through trade-off comparison, jointly constituting structural dimensions with C2
C5 Aesthetic Consistency	0.074	5	0.150	5	—	Both weighting methods rank last, judgment robust; its constraints are relatively static and prompts are easy to execute

Note: Independent weighting weights are obtained by experts directly scoring the importance of the 5 primary indicators and normalizing; AHP weights are solved from the judgment matrix through the eigenvector method.

5.3 Secondary Indicator Combined Weights

Secondary indicator local weights are solved from 3×3 judgment matrices in the sub-criteria layer through the AHP eigenvector method (the CR values of the 5 matrix groups range from 0.0087 to 0.0121, all < 0.10). Taking the C1 sub-criteria layer as an example, experts consider the

importance of C11, C12, and C13 to be relatively close with only minor differences, and the eigenvector method yields local weights (0.3577, 0.3000, 0.3423). The combined weight of each secondary indicator = primary indicator AHP weight × intra-layer local weight, and the 15 results are shown in Table 4.

Table 4. Secondary Indicator AHP Combined Weight Calculation Results

Rank	ID	Secondary Indicator	Dimension	Local Weight	AHP Combined Weight
1	C21	Information Grouping Logic	Information Architecture	0.3727	0.1152
2	C22	Label Description Clarity	Information Architecture	0.3136	0.0969
3	C23	Hierarchy Relationship Visualization	Information Architecture	0.3136	0.0969
4	C41	Space Utilization	Layout Rationality	0.3529	0.0824
5	C42	Visual Flow Guidance	Layout Rationality	0.3353	0.0783
6	C11	Font Hierarchy Clarity	Visual Clarity	0.3577	0.0769
7	C13	Icon Recognizability	Visual Clarity	0.3423	0.0736
8	C43	Element Alignment Standard	Layout Rationality	0.3118	0.0728
9	C12	Color Contrast	Visual Clarity	0.3000	0.0645
10	C31	Operation Path Simplicity	Interaction Usability	0.3800	0.0640
11	C32	Feedback Timeliness	Interaction Usability	0.3200	0.0539
12	C33	Error Prevention Effectiveness	Interaction Usability	0.3000	0.0505
13	C51	Visual Style Unity	Aesthetic Consistency	0.3533	0.0261
14	C52	Color Coordination	Aesthetic Consistency	0.3333	0.0246
15	C53	Brand Recognizability	Aesthetic Consistency	0.3133	0.0232

The top three combined weights are Information Grouping Logic (C21, 0.1152), Label Description Clarity (C22, 0.0969), and Hierarchy Relationship Visualization (C23, 0.0969), all belonging to the Information Architecture dimension; Font Hierarchy Clarity (C11), which ranked first under independent weighting, drops to rank 6 (0.0769). The logic underlying this ranking reversal is that the former determines "what information users see and in what order" (upstream decision), while the latter determines "whether it can be seen clearly" (downstream presentation), and AHP pairwise comparison translates the "upstream priority" design logic into precise mathematical weights.

6. Evaluation Model Construction and Validation

6.1 Model Construction

Based on AHP weights and checklist scoring

criteria (Table 1), this study constructs a Visual Performance Evaluation (VPE) model for OA form interfaces (see Figure 3), adopting the linear weighted method:

$$VPE = \sum W_i \times S_i \quad (1)$$

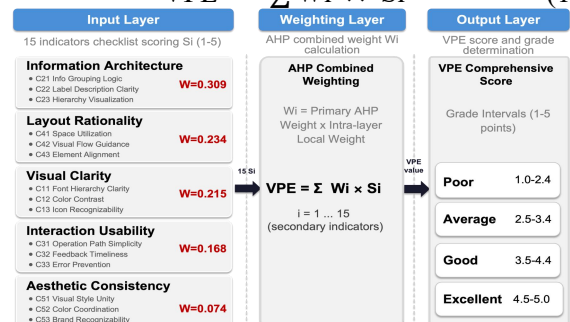


Figure 3. OA Form Interface Visual Performance Evaluation (VPE) Model Architecture

where VPE is the comprehensive evaluation value; W_i is the AHP combined weight of the i -th secondary indicator (Table 4, sum = 1); S_i is the score of that indicator (determined according

to Table 1 criteria, 1–5 points). VPE ranges from 1 to 5 points: 4.5–5.0 Excellent, 3.5–4.4 Good, 2.5–3.4 Average, 1.0–2.4 Poor (grades are determined by rounding VPE to one decimal place and against the interval, e.g., 3.44 is recorded as 3.4 and judged as Average).

6.2 Model Validation

Three actual OA forms are selected for

Table 5. Model Validation Primary Dimension Scores and VPE Comprehensive Results

Evaluation Object	C1 Visual Clarity/(Weight 0.215)	C2 Information Architecture/(Weight 0.309)	C3 Interaction Usability/(Weight 0.168)	C4 Layout Rationality/(Weight 0.234)	C5 Aesthetic Consistency/(Weight 0.074)	VPE Total	Grade
Form A (Construction)	3.4	3.6	3.2	3.5	3.3	3.44	Average
Form B (Finance)	4.2	4.1	4.0	3.9	3.8	4.04	Good
Form C (University)	2.8	3.0	2.7	2.9	2.8	2.87	Average

Furthermore, the "loss" ($5-S_i$) of each indicator is multiplied by the AHP weight for weighted shortboard analysis (Table 6), locating the indicators that exert the greatest drag on the total score. Indicators with larger weights produce stronger pulling effects of losses on the total score, and thus have higher improvement priority.

The validation results indicate that: (1) the model has sufficient discrimination, with Form B (4.04) significantly higher than Form C (2.87), a grade difference of 1.17 points; (2) inter-rater reliability $ICC = 0.87$ ($ICC(2,k)$ two-way random consistency model), an increase of 0.08

validation: Form A (construction enterprise expense reimbursement), Form B (financial institution approval workflow), and Form C (university leave application). Ten evaluators (5 UX designers, 5 OA administrators) judge each item according to Table 1 criteria, and the dimension scores and VPE comprehensive results are shown in Table 5.

from 0.79 in the pre-test before introducing the checklist criteria, indicating that standardized checklists effectively reduce subjective scoring arbitrariness; (3) AHP shortboard diagnosis differs from traditional methods—taking Form A as an example, Interaction Usability is identified as the main shortboard under independent weighting, while AHP weights reveal that losses in Information Architecture and Layout exert greater drag (higher weights), and optimization suggestions are accordingly adjusted to "prioritize improving information grouping logic and space utilization."

Table 6. AHP Weighted Shortboard Analysis

Evaluation Object	VPE Total	Top 3 Shortboards (Dimension: Score, Loss Contribution)	Optimization Suggestions
Form A (Construction)	3.44	1st: Information Architecture (score 3.6, loss contribution 0.433) / 2nd: Layout Rationality (score 3.5, loss contribution 0.351) / 3rd: Visual Clarity (score 3.4, loss contribution 0.344)	Prioritize improving information grouping logic and space utilization (Information Architecture loss contribution 0.433 ranks first, Layout Rationality 0.351 second), focus on optimizing field grouping and label description
Form B (Finance)	4.04	1st: Information Architecture (score 4.1, loss contribution 0.278) / 2nd: Layout Rationality (score 3.9, loss contribution 0.257) / 3rd: Visual Clarity (score 4.2, loss contribution 0.172)	Overall quality good, suggest focusing on Information Architecture detail refinement (loss contribution 0.278 ranks first), consolidating grouping logic and filling guidance advantages
Form C (University)	2.87	1st: Information Architecture (score 3.0, loss contribution 0.618) / 2nd: Layout Rationality (score 2.9, loss contribution 0.491) / 3rd: Visual Clarity (score 2.8, loss contribution 0.473)	All dimension scores are low, suggest conducting systematic reconstruction with information grouping and layout grid as precursors (loss contributions 0.618 and 0.491 rank top), simultaneously improving Visual Clarity

7. AI-Assisted Design Prompt Template Based on the VPE System

The VPE system and checklist criteria (Table 1) essentially constitute a systematic encoding of the design knowledge of "what a good OA form interface looks like." This study translates them into AI-executable prompt parameters through

three mapping mechanisms: direct mapping (quantitative parameters such as contrast, font size, and spacing are directly translated into prompt constraints), combinatorial mapping (multiple indicators are merged into hierarchical parameters such as information grouping and visual flow), and constraint mapping (qualitative criteria are translated into layout and behavioral

constraints). The checklist criteria for the 15 indicators are collectively transformed into 22 prompt parameters, covering 6 dimensions: layout, color, font, spacing, components, and brand.

The allocation of template parameter density follows the "dual-factor principle of weight priority and parameter executability." On one hand, the AHP weight ranking ($C2 > C4 > C1 > C3 > C5$) provides priority guidance for design resource allocation in the template: in the five-segment structure, the "Scenario Constraints" and "Visual Specifications" modules have the highest parameter density, respectively carrying the highest-weight Information Architecture ($C2$, 0.309) and Layout Rationality ($C4$, 0.234) constraints, while the "Brand Parameters" module is the most streamlined, corresponding to the lowest-weight Aesthetic Consistency ($C5$, 0.074). On the other hand, the static executability of parameters determines improvement efficiency: for example, although Aesthetic Consistency has the lowest weight, its corresponding parameters (color values, number

of colors used) can be directly and precisely executed by AI, so its improvement magnitude may instead be the largest; while Interaction Usability involves dynamic behavioral logic that is difficult to fully constrain through static prompts, its improvement magnitude is relatively limited. This dual-factor perspective simultaneously explains the dimensional differences in experimental results and provides direction for prompt engineering optimization.

The standardized prompt template adopts a five-segment structure of "Role Setting—Scenario Constraints—Visual Specifications—Component Requirements—Brand Parameters" (the complete template is shown in Appendix B). Designers can generate complete prompts by filling in keywords such as enterprise name, brand color, industry, and form type. The correspondence between each module, evaluation dimensions, and indicator IDs is shown in Table 7, ensuring full-chain traceability from "AHP weight → scoring criteria → AI instructions."

Table 7. Prompt Template Module and Indicator Mapping

Template Module	Carried Dimension (Weight)	Corresponding Indicators	Main Checklist Parameters Carried
Role Setting	—	—	Designer role and task background
Scenario Constraints	Information Architecture C2 (0.309)	C21, C22	Business grouping, label conciseness, scenario and user
Visual Specifications	Layout Rationality C4 (0.234), Visual Clarity C1 (0.215)	C41, C42, C43, C11, C12, C13, C23	Grid, spacing, alignment, font size, contrast, icons, visual flow, grouping
Component Requirements	Interaction Usability C3 (0.168)	C31, C32, C33	Control specifications, feedback validation, error prevention, attachment upload
Brand Parameters	Aesthetic Consistency C5 (0.074)	C51, C52, C53	Brand color, font, number of colors used, global unity

To validate template effectiveness, a between-group controlled experiment is conducted with 54 designers having more than 3 years of B-end experience: the experimental group (27 participants) uses the standardized template, while the control group (27 participants) writes freely without access to the VPE system or template; both groups use Midjourney V6 to generate 2 leave approval form interface images each (see Figure 4), a total of 108 pieces, with the mean score of each designer's two images as the analysis unit ($n = 27$ per group); 5 evaluators not involved in the experiment blindly score the randomly mixed 108 interfaces, and between-group differences are tested using independent samples t-tests (Welch correction when variances are unequal), with results shown in Table 8. It should be noted that static interface images cannot present interactive behaviors such as real-time validation and dynamic feedback;

therefore, the scoring of the Interaction Usability dimension ($C3$) is based on inferential evaluation of visible interactive features presented in the interface, such as control types, validation states, and feedback prompts, and its conclusions are presented as exploratory findings.

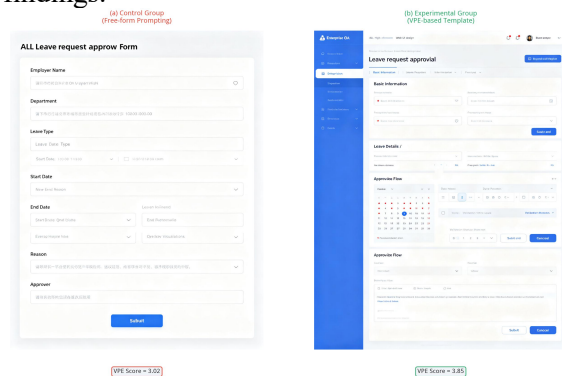


Figure 4. Representative Examples of Generated Leave-Approval Form Interfaces

Table 8. Prompt Template Validation Experiment Results

Evaluation Dimension	Experimental Group VPE	Control Group VPE	Improve ment	Experimental Group σ	Control Group σ	t-value	p-value
C1 Visual Clarity	3.82	3.12	+22.4%	0.34	0.61	5.17	< 0.001
C2 Information Architecture	3.68	2.85	+29.1%	0.38	0.65	5.44	< 0.001
C3 Interaction Usability	3.41	2.81	+21.4%	0.42	0.59	4.12	< 0.001
C4 Layout Rationality	3.91	2.94	+33.0%	0.29	0.52	7.83	< 0.001
C5 Aesthetic Consistency	3.83	2.79	+37.3%	0.27	0.55	8.51	< 0.001
VPE Comprehensive	3.73	2.92	+27.8%	0.31	0.58	6.41	< 0.001

The experimental group's mean VPE of 3.73 (Good) is significantly higher than the control group's 2.92 (Average), with a comprehensive improvement of 27.8% ($p < 0.001$). Dimensional improvement magnitude presents gradient differences: Aesthetic Consistency is the largest (+37.3%), as its corresponding parameters (color values, number of colors used) can be directly executed; Interaction Usability is the smallest (+21.4%), involving dynamic behavioral logic that is difficult to fully constrain through static prompts. The standard deviations of all dimensions in the experimental group (0.27–0.42) are smaller than those in the control group (0.52–0.65), indicating that the template simultaneously improves the mean quality and compresses output dispersion. The effect sizes (Cohen's d) for all dimensions and the comprehensive score range from 1.1 to 2.4, indicating that the improvement magnitude are statistically large effects. The complete prompt template content is as follows: The following is a five-segment standardized prompt template built based on the VPE system and AHP weights. After replacing the parameters within [] with actual enterprise information, it can be used in AI image generation tools such as Midjourney and DALL·E. Template parameter density is allocated according to AHP weights and parameter executability; indicator IDs are annotated in parentheses to ensure full-chain traceability from "AHP weight $\rightarrow$ scoring criteria $\rightarrow$ AI instructions."

[Role Setting] You are a senior B-end user experience designer, specializing in enterprise-level OA system form interface design, with a deep understanding of user operational habits and visual cognitive patterns in smart office scenarios.

[Scenario Constraints] (Corresponding to C2 Information Architecture, AHP weight 0.309)

Please design a [Form Type] interface for [Enterprise Name] This form is used for [Usage Scenario Description], and the target users are [User Group]. The interface should adapt to

[PC / Mobile] browsing, with an overall style matching the characteristics of the [Industry Type] industry. Fields are grouped by business process, with each group containing 3–6 logically related fields, and groups are carrying in card containers.

[Visual Specifications] (Corresponding to C1 Visual Clarity, C4 Layout Rationality)

Layout and Grid: Form width 550–860px, all elements aligned based on an 8px grid system. Labels are left-aligned, accounting for 40%–45%. Field spacing ≥ 12 px, inter-group spacing ≥ 16 px. Single-column vertical linear layout, filling from top to bottom (C41, C43)

Font Hierarchy: Title 14px / bold, body 12px / regular, auxiliary information 10px / light gray. Hierarchy ≥ 3 levels (title / body / auxiliary), level difference ≥ 2 px (C11)

Color Contrast: Text-to-background contrast $\geq 7:1$ (WCAG AAA level), large title $\geq 4.5:1$ (C12)

Icons and Identification: Functional icons in unified linear style, size $\geq 24 \times 24$ px, with text labels. Required fields marked with red *, clear visual flow (C13, C42, C23)

[Component Requirements] (Corresponding to C3 Interaction Usability, AHP weight 0.168)

Control Specifications: Input box height 36px, corner radius 4px, border #D9D9D9, focused state border #1890FF. Dedicated controls fully covered for dropdown selection, date selection, attachment upload, etc. (C31, C33)

Feedback Validation: Real-time validation (< 200 ms), error prompts localized to fields, success state with green checkmark, loading state with skeleton screen. Complete progress indication (C32)

[Brand Parameters] (Corresponding to C5 Aesthetic Consistency, AHP weight 0.074)

Brand Color: Primary color [Brand Color Value], no more than 2 auxiliary colors, accent color proportion $< 10\%$. Global color coordination, cold-warm balance (C52)

Font and Unity: Brand font [Font Name], global component style / corner radius / spacing / shadow completely unified. Logo placed at the

top-left corner of the page, brand discourse placed in the footer (C51, C53)

8. Discussion

(1) Methodological contribution and value of checklist criteria. This study systematically applies AHP to visual performance evaluation of OA form interfaces, ensuring weight reproducibility through pairwise comparison and consistency testing [2], the extensive validation of its variants and integration methods (ANP, DANP) in the multi-criteria decision-making field [13-15] also provides reference for the methodspectrum. The graded checklist criteria increase inter-rater reliability from 0.79 to 0.87, effectively reducing subjective scoring arbitrariness and providing an operationalized scoring paradigm for B-end interface evaluation.

(2) Practical implications of weight differences. The fact that the two weighting methods yield different rankings is itself an important finding: the resource order under independent weighting is "first make the text clear," while under AHP it is "first build the architecture and layout skeleton, and clear presentation follows naturally." This priority reversal has direct guiding value for OA vendor product roadmaps and designer revision planning.

(3) Correspondence between template effectiveness, weights, and executability. In the experiment, improvements in Layout Rationality (+33.0%) and Information Architecture (+29.1%) are relatively prominent, corresponding to the two highest-weight structural dimensions whose parameters (grid, spacing, alignment, grouping) can be directly constrained by static prompts; the largest improvement in Aesthetic Consistency (+37.3%) stems from high parameter executability, while the smallest improvement in Interaction Usability (+21.4%) stems from the difficulty of constraining dynamic behavioral logic statically. This "weight priority—parameter executability—improvement magnitude" correspondence provides empirical evidence for prompt engineering optimization and parameter organization in AI-assisted design.

9. Conclusions

This study employs the Analytic Hierarchy Process (AHP) as the weight derivation method, constructs a visual performance evaluation model for OA form interfaces, and develops graded checklist scoring criteria for 15 secondary indicators. The main conclusions are

as follows:

First, AHP weights reveal the structural priority dimensions of OA form visual performance: Information Architecture (0.309) and Layout Rationality (0.234) collectively account for more than half (54.3%) of the weight, significantly higher than the directly visible Visual Clarity (0.215). Visual optimization should take information organization and spatial skeleton as precursors, rather than starting from surface elements such as font size and contrast.

Second, significant differences exist between AHP weight ranking and independent weighting ranking (Visual Clarity ranks first), and the pairwise comparison mechanism corrects the visibility anchoring bias of direct scoring, indicating that the choice of weighting method has a substantive impact on evaluation conclusions.

Third, the VPE model combined with graded checklist criteria exhibits good discrimination and inter-rater reliability ($ICC = 0.87$), and AHP weighted shortboard analysis can provide optimization diagnosis based on the weight system.

Fourth, after translating the VPE system into a standardized AI prompt template, the mean VPE of AI-generated interfaces increases by 27.8% ($p < 0.001$), verifying the effectiveness of the "evaluation criteria to generation instructions" translation path.

This study has the following limitations: first, the AHP judgment matrix is based on subjective scoring by 8 experts, without introducing objective data calibration such as eye-tracking; Wu's team's eye-tracking experimental paradigm can serve as a future calibration baseline. Second, the sample is concentrated in 5 industries, and the 3 validation cases belong to different industries; industry characteristics may constitute confounding variables, and the cross-industry robustness of weights and conclusions requires further validation. Third, potential cross-influences among primary indicators are not explicitly modeled; subsequent work can introduce ANP or DEMATEL-ANP to achieve networked solvin. Fourth, in the template validation experiment, the control group writes freely, and it is not yet possible to fully separate the respective contributions of "structured template effect" and "AHP weight allocation effect" to VPE improvement; subsequent work can add a control condition with the same structure but non-AHP weight allocation to

further isolate the mechanism. Fifth, the experimental materials are static interface images, and the Interaction Usability dimension (C3) is based on inferential evaluation of visible interactive features, with limited representation of dynamic interaction quality. Sixth, the template is primarily adapted to image generation AI tools, and its applicability to real interactive OA forms requires further verification in conjunction with design tools. Future research can explore calibrating judgment matrices with eye-tracking data, tracking the impact of AI tool iteration on template effectiveness, and validating the unique contribution of AHP weights and cross-industry robustness through more refined experimental designs.

References

- [1] iResearch. 2024 China Collaborative Office Platform Industry Research Report. Shanghai: Shanghai iResearch Market Consulting Co., Ltd., 2024.
- [2] Saaty T L. The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation. New York: McGraw-Hill, 1980.
- [3] Sanders M S, McCormick E J. Human Factors in Engineering and Design. 7th ed. New York: McGraw-Hill, 1993.
- [4] Chen X Q, Huang L Q, Li M Z. Visual Search Performance Research on Intercity Logistics APP Interface Navigation Design. Packaging Engineering, 2021, 42(8): 198-204.
- [5] Yang M H, Yang J M, Qian X S. Research on Visual Saliency Evaluation in Decision Support System Interface Design. Packaging Engineering, 2021, 42(20): 77-84.
- [6] Wu X L, Xue C Q, Gedeon T, et al. Visual Search Experiment of Information Features in Digital Monitoring Task Interfaces. Journal of Southeast University (Natural Science Edition), 2018, 48(5): 807-814.
- [7] Wu X L, Gedeon T, Xue C Q, et al. Comparison of Consistency Effects Between Fixation/Saccade Indicators and Pupil Dilation Amplitude in Information Feature Search. Journal of Computer-Aided Design & Computer Graphics, 2019, 31(9): 1636-1644.
- [8] Wu X L, Xu P P, Jiang X M. Research on Color Hierarchy Coding of Industrial Monitoring Interfaces Based on Attentional Capture. Industrial Engineering and Management, 2022, 27(4): 32-38.
- [9] Nielsen J. Usability Engineering. San Francisco: Morgan Kaufmann, 1993.
- [10] Nielsen J. 10 Usability Heuristics for User Interface Design. (1994-04-24)[2024-12-15]. <https://www.nngroup.com/articles/ten-usability-heuristics/>.
- [11] ISO 9241-11:2018. Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts. Geneva: ISO, 2018.
- [12] Saaty T L. Decision Making with Dependence and Feedback: The Analytic Network Process. 2nd ed. Pittsburgh: RWS Publications, 2001.
- [13] Ou Yang Y P, Shieh H M, Leu J D, et al. A novel hybrid MCDM model combined with DEMATEL and ANP with applications. International Journal of Operations Research, 2008, 5(3): 160-168.
- [14] Tzeng G H, Chiang C H, Li C W. Evaluating Intertwined Effects in E-learning Programs: A Novel Hybrid MCDM Model Based on Factor Analysis and DEMATEL. Expert Systems with Applications, 2007, 32(4): 1028-1044.
- [15] Tzeng G H, Huang J J. Multiple Attribute Decision Making: Methods and Applications. Boca Raton: CRC Press, 2011.
- [16] Shao Q G, Liou J J H, Weng S S, et al. Improving the Green Building Evaluation System in China Based on the DANP Method. Sustainability, 2018, 10(4): 1173.
- [17] Yu N, Huang Y T. Important Factors Affecting User Experience Design and Satisfaction of a Mobile Health App—A Case Study of Daily Yoga App. International Journal of Environmental Research and Public Health, 2020, 17(19): 6967.
- [18] Ji Y D, Zhao M, Zhao W. Evaluation of Ecological Dominance of Industrial Chain Leader Enterprises Based on Grounded Theory: A Scale Development. Science & Technology Progress and Policy, 2024, 41(6): 86-96.
- [19] Benaida M. Developing and Extending Usability Heuristics Evaluation for User Interface Design via AHP. Soft Computing, 2023, 27(14): 9693-9707.
- [20] Ouyang W M, Du H M, Jiang J J, et al. Research on Aesthetic Design of In-Vehicle Infotainment System Interfaces. Packaging Engineering, 2025, 46(12): 274-285.

- [21] Ayağ Z, Samanlıoğlu F. A hesitant fuzzy linguistic terms set-based AHP-TOPSIS approach to evaluate ERP software packages. *International Journal of Intelligent Computing and Cybernetics*, 2021, 14(1): 54-77.
- [22] Zhu D, Xu Y, Ma H, et al. Building a Three-Level User Experience (UX) Measurement Framework for Mobile Banking Applications in a Chinese Context: An Analytic Hierarchy Process (AHP) Analysis. *Multimodal Technologies and Interaction*, 2022, 6(9): 83.
- [23] Zhang N, Wang J M, Yang Y P. Evaluation Method for Aesthetic Image of Human-Machine Interface Morphological Element Layout Design. *Mechanical Science and Technology for Aerospace Engineering*, 2015.
- [24] Zhang R J, Wang M X, Guo Y P, et al. Human-Machine Interface Design Evaluation for Oilfield Water Injection System Platforms. *Journal of Machine Design*, 2021, 41(7): 118-125.
- [25] Xu F H, Zhang L. Research on Human-Factor Suitability Evaluation System for Smartphone APP Interaction Interfaces. *Chinese Journal of Ergonomics*, 2018, 24(3).
- [26] Wu J W, Sun J L, Li M. Research on Usability Evaluation of WeChat Mini-Program Interface Design Based on Entropy Weight Method. *Packaging Engineering*, 2021, 42(12): 191-196.
- [27] Liu V, Chilton L B. Design Guidelines for Prompt Engineering Text-to-Image Generative Models//*Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'22)*. New Orleans: ACM, 2022: Article 384.
- [28] Goloujeh A M, Sullivan A, Magerko B. Is It AI or Is It Me? Understanding Users' Prompt Journey with Text-to-Image Generative AI Tools//*Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'24)*. Honolulu: ACM, 2024.
- [29] Wang Z, Huang Y, Ma L, et al. PROMPTCHARM: Text-to-Image Generation through Multi-modal Prompting and Refinement//*Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'24)*. Honolulu: ACM, 2024.