

Expert Evaluation and Implementation Feasibility of AI-Generated Scientific Dashboard Prototypes Using Controlled Layout Prompting

Joonwoo Lee*, Seungho Kim**, Scott Uk-Jin Lee***

*Student, Dept. of Applied Artificial Intelligence, Hanyang University, Ansan, Korea

**Student, Dept. of Computer Science & Engineering, Hanyang University, Ansan, Korea

***Professor, Dept. of Computer Science & Engineering, Hanyang University, Ansan, Korea

[Abstract]

Generative image models can rapidly produce UI prototypes, but uncontrolled sampling reduces comparability and may omit functional requirements. This study proposes a controlled generation and evaluation procedure using predefined layout archetypes for a scientific dashboard visualizing underwater multi-sensor fusion and BOTMA tracking simulations. Twelve prototypes were generated by crossing three model groups with four layout archetypes and evaluated by three blinded experts using quality ratings and requirement coding. Sixteen implementation paths were also tested via automated browser checks. Results showed good reliability for quality ratings ($ICC(2,3)=0.839$) but low agreement for requirement coding (Fleiss' $\kappa=0.248$). OpenAI achieved the highest quality scores, while Gemini showed the highest requirement coverage. The findings suggest that layout archetypes enable traceable structural variation, while static prototypes should be complemented with validation and verification.

▶ **Key words:** Generative-AI, UI Prototype, Dashboard Design, Info Visualization, Expert Evaluation

[요 약]

생성형 이미지 모델은 UI 프로토타입을 빠르게 생성할 수 있지만, 무작위 생성은 비교 가능성을 낮추고 기능적 요구사항 반응을 보장하지 않는다. 수중 다중센서 정보융합 및 BOTMA 측정 추적 시뮬레이션 대시보드를 대상으로, 사전 정의된 레이아웃 원형을 활용한 통제적 프로토타입 생성 및 평가 절차를 제안한다. 세 개의 이미지 모델 그룹과 네 개의 레이아웃 원형을 교차하여 프로토타입을 생성하고, 평가자가 블라인드 방식으로 품질 평정과 요구사항 코딩을 수행하였다. 또한 구현 경로에 대해 자동화된 브라우저 검사를 수행하였다. 품질 평정은 양호한 신뢰도를 보였으나, 요구사항 코딩 일치도는 낮았다. OpenAI 그룹은 품질 점수에서, Gemini 그룹은 요구사항 커버리지에서 가장 높은 값을 보였다. 본 연구는 레이아웃 원형이 생성 결과의 구조적 다양성을 추적 가능하게 하며, 정적 이미지 프로토타입은 검토 및 검증과 함께 사용되어야 함을 시사한다.

▶ **주제어:** 생성형 인공지능, 사용자 인터페이스 프로토타이핑, 대시보드 설계, 정보 시각화, 전문가 평가

- First Author: Joonwoo Lee, Corresponding Author: Scott Uk-Jin Lee
- *Joonwoo Lee (joonywlee@hanyang.ac.kr), Dept. of Applied Artificial Intelligence, Hanyang University
- **Seungho Kim (ohgnues@hanyang.ac.kr), Dept. of Computer Science & Engineering, Hanyang University
- ***Scott Uk-Jin Lee (scottlee@hanyang.ac.kr), Dept. of Computer Science & Engineering, Hanyang University
- Received: 2026. 07. 15, Revised: 2026. 08. 02, Accepted: 2026. 08. 19.

I. Introduction

생성형 이미지 모델은 자연어 입력을 기반으로 시각적 산출물을 빠르게 생성할 수 있으며, 최근 사용자 인터페이스 설계와 프로토타이핑 과정에서도 활용 가능성이 주목받고 있다. 특히 과학 대시보드와 같이 다양한 정보 요소를 동시에 배치해야 하는 시스템에서는 초기 설계 단계에서 여러 시각적 대안을 빠르게 탐색하는 것이 중요하다.

그러나 생성형 이미지 모델을 UI 프로토타입 생성에 그대로 활용할 경우 몇 가지 한계가 발생한다. 첫째, 단순히 이미지를 여러 번 생성하는 방식은 산출물 간 비교 가능성이 낮다. 두 이미지가 서로 다르게 보이더라도, 그 차이가 레이아웃 구조에서 비롯된 것인지, 스타일이나 색상 차이 때문인지, 또는 모델의 무작위적 생성 결과 때문인지 분리하기 어렵다. 둘째, 생성된 이미지는 시각적으로는 완성도 높아 보일 수 있으나 실제 구현에 필요한 기능적 요구사항을 충분히 포함하지 않을 수 있다. 예를 들어 대시보드처럼 보이는 이미지는 중앙 지도, 차트, 상태 패널과 같은 큰 시각 영역은 잘 표현할 수 있지만, 프레임 이동, 표적 선택, 센서 선택, 측정값 의미와 같은 상호작용적 요구사항은 모호하게 나타낼 수 있다.

본 연구는 이러한 문제를 다루기 위해 사전에 정의된 레이아웃 원형(layout archetype)을 활용하여 AI 생성 대시보드 프로토타입의 구조적 다양성을 통제하고, 이를 전문가 평가와 구현 검증을 통해 분석한다. 연구 대상은 수중 다중센서 정보융합 및 BOTMA 측정추적 시뮬레이션의 출력 데이터를 시각화하는 웹 기반 과학 대시보드이다. 이 시스템은 프레임별 자함·표적·센서 위치와 실제·추정 궤적을 나타내는 공간 정보, 시간에 따른 추적 오차, 현재 상태 요약, 시나리오·프레임·재생·표적·센서 제어를 하나의 화면에서 통합해야 한다. 따라서 공간·시간·상태·분석·상호작용 정보가 결합된 복합 과학 대시보드의 특성을 포함하며, 생성형 이미지 모델이 구조적 배치와 기능적 요구사항을 함께 표현할 수 있는지를 분석하기에 적합한 적용 사례이다. 분석 범위는 시뮬레이션 모델이나 추적 알고리즘의 개선이 아니라, 기존 시뮬레이션 출력 데이터를 사용자가 해석할 수 있도록 구성하는 프론트엔드 시각화 구조이다. 본 연구의 연구질문은 다음과 같다.

RQ1. 이미지 모델 그룹과 할당된 레이아웃 원형은 전문가가 평가한 프로토타입 품질과 기술적으로 어떠한 관계를 보이는가?

RQ2. 생성된 프로토타입 이미지에서 필수 대시보드 요소는 어느 정도 명확히 표현되며, 어떤 요소가 모호하거나

누락되는가?

RQ3. 독립 평가자들은 프로토타입 품질과 요구사항 존재 여부를 얼마나 일관되게 판단하는가?

RQ4. 생성 이미지 기반 구현안과 코드 우선 구현안은 동일한 결정론적 기술 검사를 통과할 수 있는가?

II. Related Work

1. Dashboard Design and Visualization Evaluation

대시보드는 다양한 정보를 제한된 화면 공간 안에 통합적으로 배치하여 사용자의 상황 인식과 분석을 지원하는 인터페이스이다. Bach 등[1]은 다수의 대시보드 사례를 분석하여 대시보드 설계 패턴을 도출하였으며, 화면 공간의 배분, 정보 조직, 시각적 위계, 상호작용 구조가 대시보드 설계의 핵심 요소임을 보였다.

시각화 평가는 목적에 따라 적절한 방법이 달라져야 한다[2,3]. 이에 본 연구는 정적 이미지의 요구사항 전달력을 전문가 평가로, 구현안의 기본 동작을 자동화 검사로 분리하여 분석하며, 최종 사용자 사용성은 평가 범위에 포함하지 않는다.

2. Generative AI and UI Prototyping

사용자 인터페이스 자동 생성 연구는 생성형 AI 이전에도 진행되어 왔다. Nichols 등[6]은 자동 생성 UI의 가능성을 보였으나, 생성된 인터페이스가 실제로 유용한지 검증하는 절차의 필요성도 함께 제기하였다. 최근에는 대규모 생성 모델을 활용하여 디자인 대안을 빠르게 생성하는 방식이 주목받고 있다. Hegemann 등[4]은 계산적 UI 설계를 다목적 제약을 가진 설계 공간 탐색 문제로 설명하였다. 이 관점에서 생성형 이미지 모델은 시각적 대안 생성 속도를 크게 높일 수 있지만, 산출된 이미지는 실제 데이터 의미, 상호작용 방식, 구현 제약을 보장하지 않는다.

또한 생성형 이미지 기반 디자인 지원은 초기 예시에 대한 고착을 유발할 수 있다. Wadinambiarachchi 등[5]은 생성형 AI가 디자인 발산 사고에 미치는 영향을 분석하며, 이미지 기반 영감이 특정 시각적 방향에 대한 의존을 강화할 수 있음을 보였다. 이러한 문제를 완화하기 위해서는 단순히 여러 이미지를 생성하고 가장 좋아 보이는 이미지를 선택하는 방식보다, 생성 이전에 구조적 조건을 정의하고 산출물을 체계적으로 보존·평가하는 절차가 필요하다. 본 연구는 이를 위해 네 가지 레이아웃 원형을 사전에 정의하고, 계획된 모든 산출물을 평가하였다.

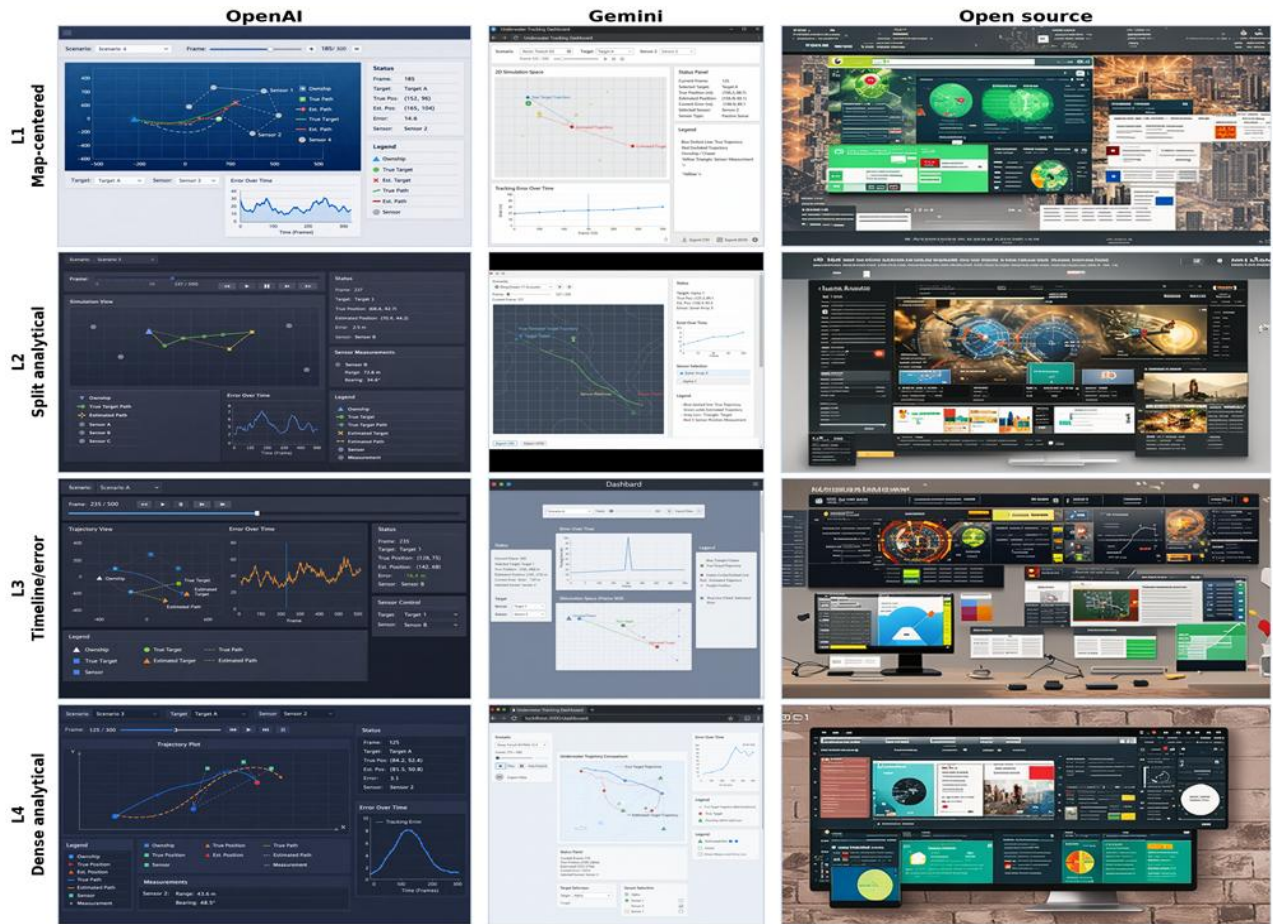


Fig. 1. Prototype Matrix by Model Group and Layout Archetype

3. Expert Evaluation and Inter-rater Reliability

전문가 평가는 UI 프로토타입과 시각화 산출물을 구조적으로 평가하는 데 유용하지만, 평가 결과를 해석하기 위해서는 평가자 간 신뢰도 분석이 필요하다. 본 연구는 5점 척도 품질 평정의 신뢰도를 평가하기 위해 Koo와 Li[7]의 보고 지침에 따라 ICC 모형과 평가 단위를 명시하였다. 또한 세 개의 명목 범주로 이루어진 요구사항 존재 여부 코딩에는 Fleiss의 카파[8]를 사용하였다. 다만 카파 계수는 범주 분포와 평가자 경향에 영향을 받을 수 있으므로, 본 연구는 정확 일치율도 함께 보고한다.

III. Methodology

1. Target System and Dashboard Requirement

본 연구의 대상 시스템은 수중 다중센서 정보융합 및 BOTMA 측정추적 시뮬레이션의 출력 데이터를 시각화하는 웹 기반 대시보드이다. 프론트엔드는 시뮬레이션 모델을 직접 구현하거나 수정하지 않으며, 기존 백엔드에서 제공되는 프레임 단위 데이터를 표시한다.

대시보드는 2차원 공간의 자함 또는 추적 플랫폼, 실제 추정 표적 및 궤적, 센서 위치·측정 지표, 시간별 추적 오차, 현재 상태, 시나리오·프레임·재생·표적·센서 제어와 범례를 포함해야 한다. 이를 중앙 2D 공간, 실제·추정 표적 위치 및 궤적, 센서 위치·측정 지표, 추적 오차, 상태 패널, 다섯 가지 제어 기능, 범례, 데스크톱 구조의 16개 요구사항으로 정의하였다.

2. Prototype Generation Design

본 연구는 세 개의 이미지 모델 그룹과 네 개의 레이아웃 원형을 교차하여 총 12개의 프로토타입 이미지를 생성하였다. 모델 그룹은 OpenAI, Gemini, 오픈소스 모델로 구성하였다. OpenAI 그룹은 gpt-image-1.5, Gemini 그룹은 gemini-2.5-flash-image, 오픈소스 그룹은 stabilityai/sd-turbo를 사용하였다. 네 가지 레이아웃 원형은 표1과 같다.

공통 프롬프트는 데스크톱 웹 대시보드, 수중 다중센서 정보융합 및 BOTMA 측정추적 시뮬레이션 출력, 프레임 단위 위치 정보, 실제 궤적과 추정 궤적 비교, 시간에 따른

Table 1. Layout Archetypes

	Layout Archetype	Structural Intention
L1	Map-centered	Places the 2D tracking space prominently at the center with supporting analysis elements around it
L2	Split analytical	Separates the spatial visualization area from the analysis and status areas
L3	Timeline/error-centered	Emphasizes frame navigation and error-over-time analysis
L4	Dense analytical	Compactly arranges the map, controls, status, error chart, and legend

추적 오차, 센서 마커와 측정 지표, 시나리오·프레임·재생·상태·범례 제어 요소를 포함하도록 구성하였다. 네 개 레이아웃 조건의 계획된 차이는 화면 구조와 정보 배치였다. 또한 사람, 실제 바다 장면, 장식적 물고기, 영화적 SF 조종석, 판타지 UI, 모바일 앱, 포스터, 과도한 발광 효과, 읽을 수 없는 텍스트, 무기 중심 이미지 등은 부정 프롬프트로 억제하였다.

생성된 이미지의 실제 크기는 OpenAI 1536×1024, Gemini 1024×1024, 오픈소스 768×432이었다. OpenAI와 Gemini는 시드 제어가 노출되지 않았으며, 오픈소스 모델은 로컬 환경에서 시드 4108-4111, 2단계 추론, guidance scale 0으로 생성되었다. 생성된 12개 이미지는 수동 편집하지 않았고, 사후 선별 없이 모두 동결하였다. 생성된 12개 프로토타입 이미지의 전체 구성은 그림 1에 제시하였다.

모델 그룹 간 비교에는 생성 조건의 차이가 포함되어 있다. 오픈소스 조건은 로컬 하드웨어 및 메모리 제약으로 인해 더 짧은 CLIP 호환 프롬프트, 낮은 해상도 및 상이한 추론 설정을 사용하였다. 따라서 본 연구의 모델 그룹 결과는 실제로 사용된 생성 조건 묶음 간의 기술적 차이를 나타내며, 관찰된 차이를 특정 모델 또는 상용·오픈소스 모델 유형의 본질적 성능 차이로 해석하는 데에는 한계가 있다.

3. Expert Evaluation Procedure

세 명의 독립 평가자가 12개 프로토타입 이미지를 모두 평가하였다. 평가자는 데이터에서 E-001, E-002, E-003으로 익명화하였다. 평가 도구는 모델명, 모델 버전, 파일명, 생성 출처를 숨겼으며, 각 평가자에게는 평가자 ID를 기반으로 결정론적으로 무작위화된 이미지 순서를 제공하였다. 단, 레이아웃 원형에 대한 중립적 설명은 표시하였다. 레이아웃 준수 여부가 평가 대상이었기 때문이다.

각 평가자는 연구 대상 대시보드의 목적과 평가 기준을 사전에 안내받은 뒤 독립적으로 평가를 수행하였다. 평가

과정에서 평가자 간 사전 논의나 합의 절차는 두지 않았으며, 이는 독립 판단에 기반한 평가자 간 신뢰도를 확인하기 위한 것이다.

각 평가자는 7개 품질 기준을 1점에서 5점까지 평가하였다. 품질 기준은 요구사항 커버리지, 웹 대시보드 현실성, 시각적 위계, 궤적 명확성, 제어 요소 가시성, 프롬프트 준수, 구현 참고 유용성이었다. 7개 품질 기준은 대시보드 설계의 정보 조직, 시각적 위계 및 상호작용 구조[1]와 시각화 평가 논의[2,3]를 본 연구의 도메인 요구사항과 결합하여 구성하였다. 웹 대시보드 현실성·시각적 위계는 화면 구조를, 궤적 명확성·제어 요소 가시성은 정보와 상호작용 단서의 판독 가능성을 평가한다. 요구사항 커버리지·프롬프트 준수·구현 참고 유용성은 생성 결과의 충실성과 후속 구현 활용성을 평가한다. 따라서 이는 최종 사용자 사용성을 측정하는 표준화 척도가 아니라 초기 프로토타입 비교를 위한 연구 맞춤형 기준이다.

또한 평가자는 16개 요구사항 요소를 present, ambiguous, absent 중 하나로 코딩하였다. present는 추론 없이 명확히 식별 가능한 경우, ambiguous는 요소가 암시되지만 불명확하거나 잘못 표기되거나 의미가 불확실한 경우, absent는 시각적 근거가 없는 경우로 정의하였다. 이미지가 비어 있거나 손상된 경우, 중복 생성 실패인 경우, 대시보드로 인식되지 않는 경우, 중앙 시각화 영역이 없는 경우 등에는 제외 플래그를 적용할 수 있도록 하였으나, 실제 분석에서 제외된 이미지는 없었다.

4. Implementation Evaluation Procedure

정적 이미지 평가는 present, ambiguous, absent 코딩을 통해 생성 이미지 자체의 요구사항 전달력을 분석하였다. 구현 평가는 동일한 React 애플리케이션, 데이터 구조 및 기능 요구사항을 공유하는 12개 프로토타입 우선 구현안과 4개 코드 우선 대조 구현안의 기술적 동작을 검사하였다. 두 결과는 합산하지 않았으며, 구현 검사의 통과를 생성 이미지가 모든 기능을 명확히 표현했다는 의미로 해석하지 않았다.

자동화된 브라우저 검사는 Chromium 149.0.7827.55, Linux 6.14.0-37-generic, Node.js 20.18.1 환경에서 수행하였다. 뷰포트 크기는 1440×1000으로 설정하였고, 모든 구현안에 동일한 121프레임 결정론적 텔레메트리 데이터를 제공하였다. 각 경로에 대해 제어 영역, 지도, 상태 패널, 오차 차트, 범례가 표시되는지 확인하였고, 프레임 슬라이더를 40으로 이동시킨 뒤 프레임 이동이 수행되는지 확인하였다. 또한 수평 오버플로, 브라우저 콘솔 오류, 문서 높이, 지도 영역의 뷰포트 점유율을 측정하였다.

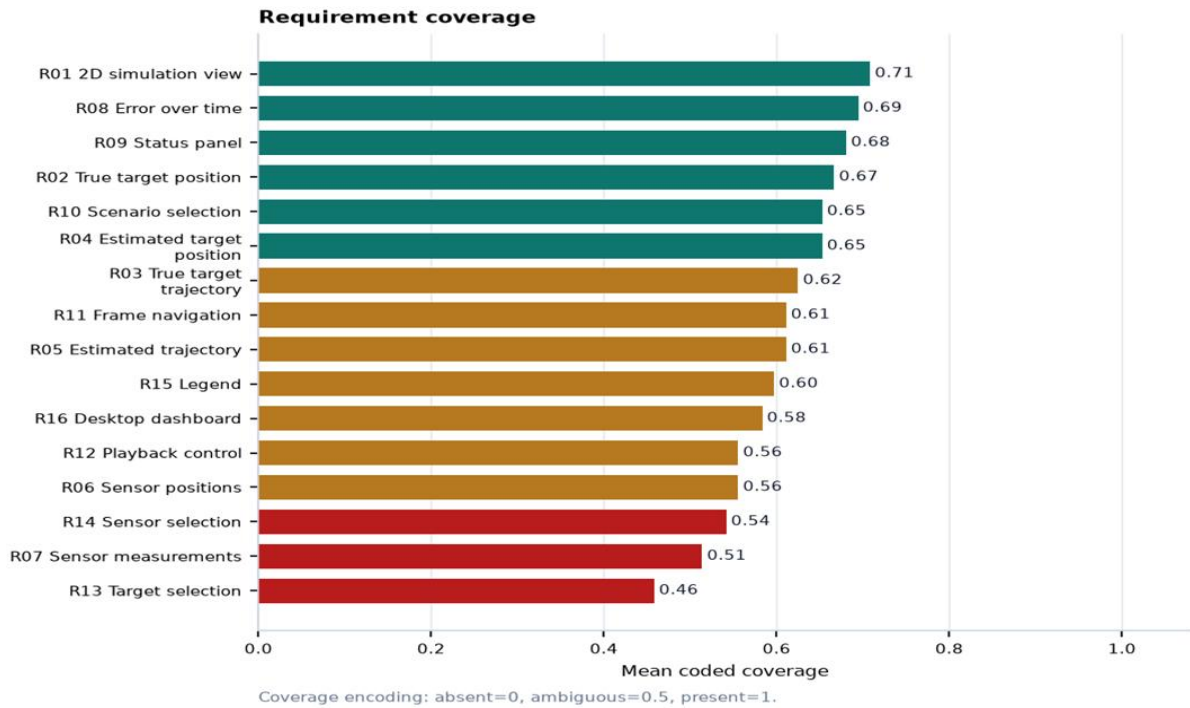


Fig. 2. Coverage of Required Dashboard Elements

5. Analysis Method

품질 평정의 평가자 신뢰도는 12개 이미지와 7개 기준의 조합인 84개 대상에 대해 계산하였다. 단일 평가자 신뢰도는 ICC(2,1), 세 평가자 평균 신뢰도는 ICC(2,3)로 보고하였다. 요구사항 코딩 신뢰도는 12개 이미지와 16개 요구사항의 조합인 192개 대상에 대해 Fleiss의 카파와 정확 일치율을 계산하였다. 신뢰구간은 대상 단위 percentile bootstrap을 사용하였고, 10,000회 재표본추출 및 결정론적 시드를 적용하였다.

요구사항 커버리지의 기술통계에는 present=1, ambiguous=0.5, absent=0으로 인코딩한 기술적 지표를 사용하였다. 이는 객관적 정확도 점수가 아니라, 평가자 판단을 집계한 서술적 커버리지 지표이다. 모델 그룹 및 레이아웃 결과는 이미지 수준 합성 평균을 기반으로 집계하였으며, 84개 기준 행을 독립적인 이미지처럼 취급하지 않았다.

Table 2. Inter-rater Reliability

Measure	Targets	Estimate	Bootstrap 95% CI
ICC(2,1), single-rater quality reliability	84	0.634	[0.536, 0.715]
ICC(2,3), mean quality reliability of three raters	84	0.839	[0.776, 0.883]
Fleiss' κ , requirement coding	192	0.248	[0.193, 0.300]
Exact requirement agreement	192	30.73%	-

IV. Results

1. Data Completeness and Inter-rater Reliability

세 평가자는 모두 12개 이미지에 대한 평가를 완료하였다. 전체 데이터는 품질 평정 252건과 요구사항 코딩 576건으로 구성되었다. 사전 정의된 제외 플래그가 적용된 이미지는 없었으며, 분석 과정에서 완결성 경고도 발생하지 않았다. 평가자 간 신뢰도 분석 결과는 표 2와 같다.

품질 평정은 단일 평가자 기준에서 중간, 세 평가자 평균 기준에서 양호한 신뢰도를 보였으나 요구사항 코딩 일치도는 낮았다. 평가자별 품질 평균은 2.345-3.071, 커버리지는 0.492-0.786으로, 모호한 요소를 present로 판단하는 기준에 차이가 있음을 보여준다.

Table 3. Results by Model Group

Model Group	Images	Quality Mean	Quality SD	Requirement Coverage Mean	Coverage SD
OpenAI	4	3.631	0.488	0.813	0.076
Gemini	4	3.191	0.413	0.849	0.083
Open Source	4	1.167	0.215	0.159	0.031

2. Quality and Requirement Coverage by Model Group

표 3과 같이 OpenAI 그룹은 전체 품질 평균에서 가장

높은 값을 보였다. 반면 Gemini 그룹은 요구사항 커버리지 평균에서 가장 높은 값을 보였다. 이는 요구사항 요소를 많이 포함하는 것과 전체적인 설계 품질 또는 구현 참고 유용성이 동일하지 않음을 보여준다. 오픈소스 그룹은 두 지표 모두에서 낮은 값을 보였으나, 앞서 설명한 바와 같이 해당 조건은 더 짧은 프롬프트, 낮은 해상도, 로컬 추론 설정 등 여러 요인이 결합되어 있으므로 특정 모델 자체의 열위로 해석해서는 안 된다.

개별 이미지 수준에서는 OpenAI L4 고밀도 분석형이 가장 높은 품질 점수(4.143)를 기록하였고, OpenAI L3 타임라인/오차 중심형(3.952), Gemini L1 지도 중심형(3.714)이 뒤를 이었다. 요구사항 커버리지에서는 Gemini L1이 가장 높은 점수(0.948)를 보였고, OpenAI L4(0.906), Gemini L4(0.885)가 뒤를 이었다. 이 결과는 시각적으로 가장 우수한 구현 참고 이미지와 가장 많은 요구사항 요소를 표시한 이미지가 반드시 동일하지 않음을 보여준다.

3. Results by Layout Archetype

레이아웃 원형별 품질 점수와 요구사항 커버리지 결과는 표 4에 정리하였다. L4 고밀도 분석형은 품질 평균과 요구사항 커버리지 평균 모두에서 가장 높은 기술통계 값을 보였다. 그러나 각 레이아웃에는 서로 다른 세 모델 그룹의 이미지가 하나씩 포함되어 있으며, 이미지 간 표준편차도 크다. 따라서 본 결과는 L4가 독립적으로 우월하다는 통계적 증거가 아니라, 본 산출물 집합 안에서 고밀도 분석형이 상대적으로 높은 평균을 보였다는 기술적 관찰로 해석해야 한다.

4. Results by Quality Criterion and Requirement

7개 품질 기준 중 요구사항 커버리지가 가장 높은 평균 점수(3.056)를 보였고, 궤적 명확성(2.889)이 뒤를 이었다. 반면 구현 참고 유용성(2.361)과 프롬프트 준수(2.417)는 낮은 평균을 보였다. 이는 생성 이미지가 요구된 요소를 어느 정도 포함하더라도, 이를 실제 구현자가 참고할 수 있을 만큼 일관되고 명확한 인터페이스 구조로 제공하지 못할 수 있음을 의미한다.

Table 4. Results by Layout Archetype

Layout	Images	Quality Mean	Quality SD	Requirement Coverage Mean	Coverage SD
L1	3	2.651	1.449	0.615	0.403
L2	3	2.476	1.156	0.580	0.404
L3	3	2.571	1.458	0.573	0.352
L4	3	2.952	1.356	0.660	0.409

요구사항별 커버리지는 다음과 같은 경향을 보였다. 중앙 2D 시뮬레이션 공간은 가장 높게 나타났으며(0.708), 시간에 따른 추적 오차(0.694), 현재 프레임 상태 패널(0.681)이 뒤를 이었다. 반면 표적 선택(0.458), 센서 측정 지표(0.514), 센서 선택(0.542)은 상대적으로 낮았다. 16개 요구사항별 커버리지 차이는 그림 2에 제시하였다.

즉 생성형 이미지 모델은 중앙 지도, 그래프, 상태 패널처럼 큰 시각 영역은 비교적 잘 표현했지만, 선택 기능이 나 센서 측정 의미처럼 작고 상호작용적인 요구사항은 누락하거나 모호하게 표현하는 경향을 보였다. 이는 정적 이미지 프로토타입이 기능 명세나 상호작용 명세를 대체할 수 없음을 보여준다.

5. Implementation Check Results

총 16개 프론트엔드 경로는 모두 필수 인터페이스 영역을 표시하였고, 프레임 40으로의 이동 검사도 통과하였다. 수평 오버플로 사례와 브라우저 콘솔 오류는 발생하지 않았다. 자동화된 구현 평가 결과는 표 5와 같다.

Table 5. Automated Implementation Check Results

Measure	Value
Number of implementation variants	16
Mean telemetry-ready load time	633.74 ms
Mean frame-navigation latency	7.05 ms
Mean document height	1400 px
Mean map viewport-area ratio	0.412

레이아웃별로는 L3 타임라인/오차 중심형이 가장 긴 페이지 높이와 가장 큰 지도 영역 점유율을 보였다. L1 지도 중심형은 가장 짧은 페이지 높이와 가장 작은 지도 영역 점유율을 보였다. 다만 이러한 시간 및 크기 측정값은 하나의 자동화 환경에서 얻은 엔지니어링 진단 지표이며, 사용자의 실제 업무 속도, 선호도, 인지 부담 또는 사용성을 의미하지 않는다.

V. Discussion

1. Controlling Structural Diversity through Layout Archetypes

본 연구의 핵심 의의는 생성 전 레이아웃 원형을 정의함으로써 AI 생성 산출물의 구조적 차이를 추적 가능하게 만들었다는 점이다. 상용 이미지 모델에서는 시드 제어가 제한적이므로, 단순 반복 생성만으로는 산출물 간 비교 기준이 불명확해질 수 있다. 이에 본 연구는 지도 중심형, 분할 분석형, 타임라인/오차 중심형, 고밀도 분석형이라는 네

가지 원형을 사전에 설정하여 각 이미지가 어떤 구조적 의도를 바탕으로 생성되었는지 명확히 기록하였다.

다만 동일한 원형 안에서도 모델 그룹별 차이가 나타났고, 일부 이미지는 요구사항을 충분히 반영하지 못했다. 따라서 레이아웃 원형은 생성 조건을 통제하고 비교 가능성을 높이는 데 유용하지만, 생성 이미지가 해당 구조를 항상 충실히 따른다고 보기는 어렵다.

2. Separating Requirement Inclusion from Implementation Usefulness

Gemini 그룹은 요구사항 커버리지에서 가장 높은 평균을 보였지만, 전체 품질 평균에서는 OpenAI 그룹이 더 높았다. 또한 개별 이미지 수준에서도 최고 커버리지 이미지는 Gemini L1이었으나, 최고 품질 이미지는 OpenAI L4였다. 이는 생성 이미지 평가에서 단순한 요소 포함 여부와 전체 설계 품질을 구분해야 함을 보여준다.

즉, 많은 요소를 포함한 이미지라도 시각적 위계와 구현 가능한 구조가 부족하면 구현 참고 자료로서의 가치는 낮을 수 있다. 따라서 생성형 이미지 기반 UI 프로토타입은 단독 명세가 아니라 요구사항 목록, 상호작용 명세, 데이터 의미 설명과 함께 사용되어야 한다. 모든 구현안의 통과는 명시적 요구사항과 공통 데이터 구조가 제공된 조건에서의 기술적 실현 가능성을 보여주며, 생성 이미지 자체가 모든 기능을 전달했음을 의미하지 않는다.

3. Fragility of Interaction Semantics

요구사항별 분석에서 표적 선택, 센서 선택, 센서 측정 지표는 낮은 커버리지를 보였다. 이러한 요소들은 단순한 시각적 장식이 아니라 사용자가 표적과 센서를 선택하고 측정값의 의미를 해석하는 데 필요한 기능이다. 그러나 정적 이미지는 이러한 기능적 의미와 상호작용 방식을 안정적으로 전달하지 못했다. 따라서 생성형 이미지 모델은 초기 화면 배치와 시각적 방향 탐색에는 유용하지만, 상호작용 설계와 데이터 의미 설계에는 별도의 검증 절차가 필요하다. 특히 의미 해석이 중요한 시스템에서는 정적 이미지의 시각적 완성도에 과도하게 의존해서는 안 된다.

4. Meaning of Low Requirement-coding Agreement

요구사항 코딩의 Fleiss' κ 는 0.248로 낮았다. 이는 단순한 연구 약점이라기보다, 생성 이미지가 세부 요구사항을 모호하게 표현한다는 결과로 해석할 수 있다. 평가자들은 동일한 이미지를 보면서도 특정 시각 요소를 센서 선택이나 측정 지표로 볼 수 있는지에 대해 서로 다른 판단 기준을 적용하였다.

후속 연구에서는 요구사항별 조작적 정의와 present, ambiguous, absent의 긍정·경계·부정 사례를 코드북에 제시할 필요가 있다. 예를 들어 센서 선택은 단순한 센서 패널이 아니라 드롭다운, 목록, 토글 또는 선택 상태가 명확히 식별되는 경우에만 present로 판정할 수 있다. 또한 별도 이미지를 이용한 파일럿 코딩으로 코드북을 보완하고, 독립 코딩 결과로 신뢰도를 계산한 뒤 불일치 항목에 한해 합의 절차를 수행할 수 있다. 시각 요소의 존재 여부와 기능적 의미의 해석 가능성을 분리하는 것도 코딩의 명확성과 재현성을 높이는 방법이다.

5. Applicability and Limitations

본 연구의 절차는 공간·시간 정보, 상태 요약, 분석 그래프 및 제어 요소가 결합된 과학·공학 대시보드의 초기 프로토타이핑에 적용할 수 있다. 다른 도메인에서는 레이아웃 원형과 요구사항 목록을 해당 업무와 데이터 의미에 맞게 재정의해야 하므로, 본 연구의 모델별·레이아웃별 수치를 다른 시스템에 직접 일반화하는 데에는 한계가 있다.

또한 본 연구는 하나의 BOTMA 사례, 세 모델 그룹, 네 레이아웃 원형, 조건별 단일 이미지 및 세 명의 평가자에 기반한다. 특히 오픈소스 조건은 하드웨어 제약으로 인해 낮은 해상도, 축약된 프롬프트 및 상이한 추론 설정을 사용했으므로, 낮은 결과를 해당 모델이나 오픈소스 모델 전반의 열위로 해석하기 어렵다. 자동화 검사 역시 기본 기능의 동작을 확인한 것으로 실제 사용성이나 인지 부담을 평가하지 않았다. 향후에는 생성 조건과 반복 횟수를 통제하고, 평가자·도메인·사용자 기반 평가를 확대할 필요가 있다.

VI. Conclusion

본 연구는 사전 정의된 레이아웃 원형을 활용하여 AI 생성 과학 대시보드 프로토타입의 구조적 변이를 통제하고 평가하는 절차를 제안하였다. 12개 이미지에 대한 블라인드 전문가 평가에서 OpenAI 그룹은 전체 품질, Gemini 그룹은 요구사항 커버리지에서 가장 높은 평균을 보였다. 생성 이미지는 중앙 지도, 오차 그래프 및 상태 패널은 비교적 잘 표현했지만, 표적·센서 선택과 측정 지표 같은 상호작용적 요소는 불명확하게 표현하였다.

구현안은 모두 자동화 검사를 통과했으나, 명시적 요구사항과 공통 구현 구조가 제공된 조건에서의 기술적 실현 가능성을 의미한다. 또한 모델별 해상도, 프롬프트 및 추론 조건이 동일하지 않았으므로 모델의 본질적 성능 차이로 일반화하는 데에는 한계가 있다. 따라서 생성형 이미지

기반 UI 프로토타입은 초기 설계 탐색에 활용하되, 요구사항 명세, 전문가 검토 및 구현 검증과 함께 사용해야 한다.

ACKNOWLEDGEMENT

This work was supported by Korea Research Institute for defense Technology planning and advancement (KRIT) Grant funded by Defense Acquisition Program Administration (DAPA) (KRIT-CT-23-026, Integrated Underwater Surveillance Research Center for Adapting Future Technologies, 2023-2029)

REFERENCES

- [1] B. Bach, E. Freeman, A. Abdul-Rahman, C. Turkay, S. Khan, Y. Fan, and M. Chen, "Dashboard Design Patterns," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 1, pp. 342-352, 2023. doi:10.1109/TVCG.2022.3209448.
- [2] C. Plaisant, "The Challenge of Information Visualization Evaluation," in *Proceedings of the Working Conference on Advanced Visual Interfaces*, pp. 109-116, 2004. doi:10.1145/989863.989880.
- [3] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale, "Empirical Studies in Information Visualization: Seven Scenarios," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 9, pp. 1520-1536, 2012. doi:10.1109/TVCG.2011.279.
- [4] L. Hegemann, Y. Jiang, J. G. Shin, Y.-C. Liao, M. Laine, and A. Oulasvirta, "Computational Assistance for User Interface Design: Smarter Generation and Evaluation of Design Ideas," in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, Article 430, 2023. doi:10.1145/3544549.3583960.
- [5] R. Wadinambiarachchi, R. M. Kelly, S. Pareek, Q. Zhou, and E. Velloso, "The Effects of Generative AI on Design Fixation and Divergent Thinking," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024. doi:10.1145/3613904.3642919.
- [6] J. Nichols, D. H. Chau, and B. A. Myers, "Demonstrating the Viability of Automatically Generated User Interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1283-1292, 2007. doi:10.1145/1240624.1240819.
- [7] T. K. Koo and M. Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155-163, 2016. doi:10.1016/j.jcm.2016.02.012.
- [8] J. L. Fleiss, "Measuring Nominal Scale Agreement among Many Raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378-382, 1971. doi:10.1037/h0031619.

Authors



Joonwoo Lee received the B.S. degree in Computational Data Science from the University of Sydney, Sydney, Australia. He is currently pursuing the integrated M.S. and Ph.D. degree in Artificial Intelligence with

the Department of Artificial Intelligence Convergence, Hanyang University, Ansan, Republic of Korea. His research interests include software engineering, software quality, software automation, and artificial intelligence for software engineering (AI4SE), with a particular focus on the evaluation and improvement of AI-assisted software development.



Seungho Kim is a Ph.D. candidate in Computer Science & Engineering at Hanyang University, Ansan, Korea. He received the B.E. degree from the School of Computer Science at Hanyang University ERICA,

Ansan, Korea, in 2024. His research interests include artificial intelligence for software engineering (AI4SE), software testing, and software quality. In particular, his work focuses on applying artificial intelligence and large language models to automated test generation, test repair, and software defect prediction. His research aims to improve the reliability, maintainability, and trustworthiness of software systems.



Scott Uk-Jin Lee received the B.E. degree in software engineering and the Ph.D. degree in computer science from The University of Auckland, Auckland, New Zealand, in 2004 and 2009, respectively.

In 2011, he joined Hanyang University ERICA, Ansan, Republic of Korea, where he is currently a Professor with the School of Computer Science and also serves as the Director of the SW-AI Convergence Education Institute. His research interests include software maintenance and evolution, software analytics, software quality, automated software engineering, and empirical software engineering.