

자동 평가 지표 기반 생성형 AI 광고 이미지 비교와 프롬프트 개선 연구

심 현 준*

배재대학교 광고사진영상학과 초빙교수

Comparison of AI-Generated Advertising Images and Prompt Improvement Based on Automatic Evaluation Metrics

Hyun-Jun Sim*

Invited Professor, Department of Advertising Photography and Film, Paichai University, Daejeon 35345, Korea

[요 약]

본 연구는 광고 도메인 특화 지표인 AD-LPIPS와 AD-Score, 범용 지표인 LPIPS와 CLIPScore를 독립 테스트셋에 적용하여 생성 모델 차이, 광고사진 표현 형식별 특성, 전문가 판단과의 관계, 프롬프트 개선 가능성을 검토하였다. 국제광고사진 공모전 수상작 180장을 대상으로 동일한 역설계 프롬프트를 적용해 Midjourney V7과 Nano Banana Pro 이미지를 생성하였다. 네 지표의 결과는 2×6 혼합설계 분산분석과 전문가 3인의 2AFC 다수결 판단을 통해 비교하였다. 분석 결과, 생성 모델의 주효과는 네 지표 모두에서 유의하였고, 표현 형식의 주효과는 세 지표에서 유의하였다. 특정 지표의 우위는 명확하지 않았으나, AD 지표 기반 프롬프트 개선 결과는 전문가에게 더 자주 선택되었다. 이는 자동 평가 지표가 생성 이미지 선별과 프롬프트 개선을 지원하며, 광고 제작 과정에서 전문가 판단을 보완할 수 있음을 시사한다.

[Abstract]

This study applied the advertising-domain-specific metrics AD-LPIPS and AD-Score, along with the general-purpose metrics LPIPS and CLIPScore, to an independent test set to examine the differences between generative models, characteristics by advertising photography expression type, correspondence with expert judgment, and prompt improvement potential. The test set comprised 180 award-winning images from an international advertising photography competition. Applying the same reverse-engineered prompt, images were generated with Midjourney V7 and Nano Banana Pro. The four metric outputs were analyzed through a 2×6 mixed-design ANOVA and compared with majority-vote 2AFC judgments by three advertising photography experts. The results showed significant main effects of generative model across all four metrics and significant main effects of expression type in three metrics. Although no single metric showed explicit superiority, AD-metric-based prompt refinement was more frequently selected by experts. These findings suggest that automatic metrics can support image selection and prompt improvement while complementing expert judgment in AI-based advertising production.

색인어 : 자동 평가 지표, 전문가 다수결, 독립 테스트셋, 혼합설계 분산분석, 프롬프트 개선**Keyword** : Automatic Evaluation Metric, Expert Majority, Independent Test Set, Mixed-Design ANOVA, Prompt Improvement<http://dx.doi.org/10.9728/dcs.2026.27.9.2561>

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 12 June 2026; Revised 08 July 2026

Accepted 16 July 2026

*Corresponding Author; Hyun-Jun Sim

Tel: ※ 개인정보 표시제한

E-mail: lightjun@naver.com

I. 서 론

1-1 연구의 배경과 목적

생성형 AI를 활용한 광고 이미지 제작은 한 번의 생성으로 완결되지 않는다. 실제 제작 과정에서는 프롬프트를 작성하여 이미지를 생성하고, 생성 결과를 검토하여 광고 목적에 더 적절한 이미지를 선별하며, 부족한 부분을 다시 프롬프트에 반영해 재생성하는 과정이 반복된다. 특히 광고 이미지는 단순히 보기 좋은 이미지가 아니라 제품의 형태와 물성, 브랜드가 전달하려는 메시지, 시각적 주목성, 표현 전략의 적합성을 함께 충족해야 한다. 따라서 생성형 AI 광고 이미지 제작에서 중요한 것은 개별 결과물의 미적 완성도만이 아니라, 여러 생성 결과 중 어떤 이미지를 선택하고 어떤 방향으로 수정할 것인가를 판단하는 과정이다.

이러한 반복 과정에서 자동 평가 지표는 생성 결과를 사후적으로 평가하는 데 그치지 않고, 비교와 선별을 보조하거나 다음 생성을 위한 수정 방향을 줌으로써 신호로 쓰일 수 있다. 기존의 LPIPS, CLIPScore와 같은 범용 지표는 이미지 간 시각적 차이나 텍스트-이미지 간 의미 적합성을 수치화함으로써 생성 결과를 비교하는 데 유용하다. 그러나 이들 지표는 대부분 자연 이미지나 일반 이미지-텍스트 자료를 기반으로 설계되어 광고사진 전문가가 중요하게 판단하는 광고적 적합성, 제품 설명력, 표현 형식의 충실성, 시각적 설득력을 직접 측정하는 데에는 한계가 있다. 이에 따라 광고사진 도메인에서는 범용 지표와 함께 전문가 판단을 반영한 도메인 특화 지표가 실제 외부 자료에서도 어떤 방식으로 작동하는지 검토할 필요가 있다.

자동 평가 지표가 광고 제작 과정에서 보조 도구로 활용되기 위해서는 두 가지가 확인되어야 한다. 하나는 학습에 사용하지 않은 외부 광고사진 자료에서도 지표가 생성 모델 간 차이를 안정적으로 포착하는가이다. 내부 검증에서 일정한 성능을 보인 지표라 하더라도, 학습 자료와 출처·표현 방식·품질 조건이 다른 자료에 적용될 때 동일한 양상을 보인다고 단정할 수는 없다. 따라서 독립 테스트셋에서 지표의 산출값을 확인하는 과정은 도메인 특화 자동 평가 지표의 적용 가능성을 검토하는 데 필요한 절차이다.

다른 하나는 자동 지표의 판정이 광고사진 전문가의 판단과 어떤 관계를 보이는가이다. 자동 지표가 측정하는 시각 차이, 의미 적합성, 다차원 품질 예측은 전문가가 광고적으로 적절하다고 보는 판단과 부분적으로만 겹친다. 따라서 자동 지표와 전문가 판단 사이의 단순 일치율만을 확인하는 것으로는 지표의 활용 조건을 충분히 설명하기 어렵다. 전문가 판단이 특정 생성 모델 쪽으로 치우치는 경우에는 다수 범주 기준선도 높은 일치율을 보일 수 있으므로, 균형 정확도나 우연 일치 보정 지표를 함께 살펴야 한다. 또한 자동 지표와 전문가 판단의 불일치는 반드시 오류가 아니라, 광고 이미지 평가가 시각적 유사도, 의미 적합성, 품질 예측, 표현 전략 판단이

결합된 다층적 평가라는 사실을 보여주는 단서가 될 수 있다.

본 연구의 목적은 광고사진 전문가의 판단을 반영하여 구성된 도메인 특화 자동 평가 지표와 범용 지표를 학습에 사용하지 않은 독립 테스트셋에 적용하고, 두 상용 생성 모델의 비교와 표현 형식별 산출 양상, 전문가 판단과의 관계를 검토하는 데 있다. 여기에 더해 자동 지표를 프롬프트 수정의 피드백 신호로 사용하는 개선 루프의 가능성을 예비적으로 검토한다. 본 연구는 자동 지표의 완전한 외적 타당성을 입증하려는 것이 아니라, 외부 자료에서 자동 지표가 작동하는 방식과 전문가 판단의 보완이 필요한 지점을 확인하는 데 분석을 한정한다. 이를 통해 자동 평가 지표가 생성형 AI 광고 이미지 제작 과정에서 비교, 선별, 프롬프트 개선을 지원하는 보조적 평가 체계로 활용될 가능성을 논의하고자 한다.

1-2 연구문제

본 연구는 자동 평가 지표의 외부 적용 가능성과 프롬프트 개선 활용 가능성을 네 가지 연구문제로 검토한다. 먼저 학습에 사용되지 않은 독립 테스트셋에서 자동 지표가 두 생성 모델의 산출 차이를 포착하는지 확인하고, 광고사진 표현 형식이 지표 산출값을 구분하는 분석 단위로 작동하는지 살펴본다. 또한 자동 지표의 우세 모델 판정이 연구에 참여하지 않은 전문가의 다수결 판단과 어느 정도 대응하는지 분석하고, 마지막으로 자동 지표가 프롬프트 수정의 피드백 신호로 활용될 수 있는지 예비적으로 검토한다.

연구문제 1. 독립 테스트셋에서 Midjourney V7과 Nano Banana Pro의 생성 이미지는 네 자동 평가 지표 산출값에서 차이를 보이는가?

연구문제 2. 자동 평가 지표 산출값은 광고사진 표현 형식에 따라 달라지는가?

연구문제 3. 자동 평가 지표의 우세 모델 판정은 연구에 참여하지 않은 전문가 3인의 다수결 판단과 어느 정도 일치하는가?

연구문제 4. 자동 평가 지표를 피드백 신호로 사용하는 프롬프트 개선 루프는 범용 지표 기반 루프와 비교해 전문가 판단에서 더 적절한 결과를 보이는가?

II. 본 론

2-1 생성형 AI 광고 콘텐츠의 평가 문제

광고사진의 품질은 단일 척도로 환원되지 않는다. 제품의 형태와 물성을 재현하는 능력, 광고 메시지를 전달하는 능력, 관습을 변형하는 창의성, 구도·조명·색감의 조형적 완성도는 서로 다른 판단 차원을 이루며, 광고의 목적과 표현 전략에 따라 평가의 중심도 달라진다[1]. 따라서 광고사진 평가는 이미지의 시각적 완성도뿐 아니라, 해당 이미지가 어떤 표현 형식과 광고 목적 안에서 구성되었는가를 함께 고려해야 한다.

생성형 AI 광고 이미지의 평가에서도 이러한 다차원성은 그대로 유지된다.

범용 이미지 평가 지표는 자연 이미지나 일반 이미지-텍스트 쌍을 기반으로 학습되어, 광고로서의 완성도와 설득력 같은 도메인 기준을 직접 측정하도록 설계되지 않았다[2]. LPIPS는 원본과 생성 이미지 사이의 시각적 거리를, CLIPScore는 프롬프트와 이미지 사이의 의미적 대응을 확인하는 데 유용하지만, 광고사진의 품질은 원본 유사도나 텍스트-이미지 정합성만으로 충분히 설명되지 않는다. 최근 텍스트-이미지 생성 평가 연구에서도 전체적 정합성 지표가 복합 객체, 속성, 공간 관계 등 세부 구성 요소를 충분히 포착하지 못할 수 있음이 지적되고 있다[3],[4].

광고 도메인에서도 생성 이미지의 창의성과 설득성을 측정하려는 시도가 나타났으나[5], 서로 다른 자동 지표가 같은 자료에서 어떤 일치와 불일치를 보이는지, 그 판정이 전문가 판단과 어떤 관계를 갖는지는 충분히 다루어지지 않았다. 특히 자동 지표는 지각 차이, 의미 적합성, 다차원 품질 예측 중 어느 측면을 측정하는가에 따라 서로 다른 결론을 낼 수 있다. 따라서 자동 지표의 결과를 하나의 점수로 통합하기보다, 각 지표가 반영하는 평가 관점을 병렬적으로 비교할 필요가 있다[6].

또한 모델 간 비교는 전체 평균 수준뿐 아니라 표현 형식 단위에서도 검토될 필요가 있다. 광고사진 표현 형식은 이미지의 표면적 분류가 아니라, 광고 이미지가 제품, 정서, 일상성, 극적 장면, 충족감, 감각적 분위기 중 어떤 전략을 중심으로 구성되는지를 설명하는 분석 틀이다. 따라서 생성형 AI 이미지 평가에서도 표현 형식은 자동 지표의 산출값과 전문가 판단을 해석하는 맥락 변수로 기능할 수 있다. 본 연구는 이러한 문제의식에 따라 자동 지표의 모델 간 비교 결과를 전체 수준과 표현 형식 수준에서 함께 검토하고, 그 결과가 전문가 다수결 판단과 어떻게 대응하거나 어긋나는지를 분석한다.

2-2 자동 평가 지표와 전문가 판단

자동 지표와 전문가 판단의 관계는 단순한 일치 여부로 환원되지 않는다. 자동 지표가 측정하는 지각 차이, 의미 적합성, 다차원 품질 예측은 서로 다른 정보를 담고 있으며, 각 지표는 광고 이미지의 다른 평가 차원에 민감하게 반응한다. 지각적 유사도 지표는 원본과 생성 이미지 사이의 시각적 근접성을 포착하는 데 유리하지만, 광고 메시지의 설득력이나 표현 전략의 적합성을 충분히 설명하지 못할 수 있다. 반대로 의미 기반 지표는 프롬프트와 이미지의 대응을 확인하는 데 유용하지만, 조명, 색감, 구도, 제품 물성 재현과 같은 광고사진의 조형적 완성도를 직접 판단하지는 않는다. 따라서 한 지표가 전문가 판단과 일치하지 않는다고 해서 곧바로 오류로 보기는 어렵다. 최근에는 생성 이미지 평가에서 인간 선호를 직접 수집하거나 보상모델로 학습하여 자동 지표와 인간 판단 사이의 간극을 줄이려는 시도도 이루어지고 있다[6],[7].

전문가 판단 역시 하나의 계산 기준으로 환원되기 어렵다. 광고사진 전문가는 시각적 완성도뿐 아니라 제품 재현, 표현 형식의 적합성, 메시지 전달력, 광고적 설득력 등을 종합적으로 고려한다. 특히 두 생성 모델의 결과물이 모두 일정 수준 이상의 품질을 보이거나 서로 다른 장단점을 가질 경우, 전문가 사이에서도 판단이 나뉠 수 있다. 따라서 자동 지표와 전문가 판단의 관계를 해석할 때에는 일치한 사례뿐 아니라 불일치한 사례가 어떤 조건에서 발생하는지도 함께 살펴야 한다. 이러한 관점은 인간 선호 기반 평가모델이 비교 상황에서 어떤 이미지가 더 적절하게 선택되는지를 학습한다는 점과도 연결된다[6],[7].

전문가 판단과의 일치도를 단순 일치율만으로 해석하면 오해가 생길 수 있다. 전문가 판단이 한 모델에 치우쳐 분포할 경우, 항상 그 모델을 선택하는 단순 기준선도 높은 일치율을 보일 수 있기 때문이다. 따라서 자료 분포에 좌우되지 않는 균형 정확도, 범주별 재현율, 우연 일치를 보정한 Cohen의 κ , 그리고 이진 분류의 균형적 대응 관계를 보여주는 MCC를 다수 범주 기준선과 함께 제시해야 자동 지표의 대응 관계를 적절히 해석할 수 있다[8].

이러한 관점에서 자동 지표는 전문가 판단을 대체하는 최종 판정자가 아니라, 전문가 검토가 필요한 지점을 좁히는 보조 도구로 이해되어야 한다. 자동 지표는 많은 생성 결과를 빠르게 비교하고, 지표 간 판정이 엇갈리는 사례를 선별하는 데 유용하다. 특히 불일치 사례는 광고 이미지 평가에서 어떤 차원이 충돌하고 있는지를 보여주는 분석 대상이 될 수 있다. 따라서 자동 지표의 가치는 전문가 판단을 그대로 재현하는 데만 있는 것이 아니라, 전문가가 집중적으로 검토해야 할 비교 지점을 구조화하는 데 있다.

2-3 프롬프트 개선과 반복 피드백

생성형 AI의 결과물을 사람의 판단이나 외부 신호로 반복적으로 개선하는 접근은 여러 영역에서 시도되어 왔다. 사람의 선호를 학습 신호로 사용하여 생성 결과를 정렬하거나[9], 모델이 산출한 결과를 평가 신호에 따라 반복적으로 수정하는 자기 개선 절차가 그 예이다[10]. 이러한 흐름은 이미지 생성 영역에서도 생성·평가·수정·재생성의 순환 구조 안에서 평가 신호가 다음 생성을 위한 피드백으로 활용될 수 있음을 보여준다. 특히 프롬프트 최적화 연구는 사용자의 입력을 모델이 더 잘 반응하는 프롬프트로 자동 조정함으로써, 수작업 중심의 프롬프트 엔지니어링을 보완할 수 있음을 보여준다[11].

생성형 AI 기반 광고 이미지 제작 과정도 이와 유사하다. 제작자는 하나의 프롬프트로 이미지를 생성한 뒤 결과를 검토하고, 표현 전략이나 광고 목적에 부합하지 않는 요소를 수정하여 다시 생성한다. 이 과정에서 자동 평가 지표가 단순한 사후 평가 도구를 넘어 프롬프트 수정 방향을 제안하는 피드백 신호로 활용된다면, 생성형 AI 광고 제작의 반복 비용을 줄이고 의사결정을 지원하는 도구로 기능할 수 있다.

다만 자동 지표를 활용한 개선 과정에는 방법론적 주의가 필요하다. 자동 지표로 생성 결과를 수정하고, 수정된 결과를 다시 동일한 지표로 평가하면 개선의 확인이 같은 기준 안에서 반복되는 순환 구조에 빠질 수 있다. 이 경우 지표 점수는 향상되더라도 실제 광고 품질이 향상되었다고 단정하기 어렵다. 따라서 개선 결과는 자동 지표 외부의 준거, 즉 광고사진 전문가의 판단을 통해 확인될 필요가 있다.

자동 지표의 피드백이 실제 개선에 기여했는지를 확인하려면, 같은 출발점에서 피드백 신호만 다르게 적용한 대조 조건과 비교해야 한다. 또한 자동 지표의 산출값 변화뿐 아니라, 최종 결과물이 전문가 판단에서 어떻게 평가되는지를 함께 검토해야 한다. 본 연구는 이러한 관점에서 광고 도메인 특화 지표를 사용하는 경우와 범용 지표를 사용하는 경우를 비교하여, 자동 평가 지표가 프롬프트 수정의 실질적 개선 신호로 활용될 수 있는 가능성을 예비적으로 검토한다.

III. 연구방법

3-1 독립 테스트셋의 구성

독립 테스트셋은 한국광고사진가협회가 주최하는 국제광고사진 공모전 수상작을 원본 광고사진으로 한다. 이 자료는 실제 매체 집행을 전제로 제작된 상업 광고 이미지와 달리 공모전 심사를 거쳐 선정된 이미지로, 출처와 선별 기준에서 학습 자료와 구분되는 조건을 제공한다. 표본 모집단은 2017년부터 2025년까지 출품된 1,640점 가운데 본상 또는 입선으로 수상한 334장이며, 팬데믹으로 공모전이 시행되지 않은 2020년과 2021년은 모집단에서 제외하였다.

표 1. 국제광고사진 공모전 연도별 출품작과 수상작 수

Table 1. Annual statistics of the international advertising photography competition

Year	Number of Entries	Number of Entrants	Number of Award-Winning Works
2017	273	144	34
2018	318	102	47
2019	322	140	54
2022	112	59	40
2023	227	101	59
2024	155	97	49
2025	233	106	51
Total	1,640	749	334

수상작 334장은 여섯 표현 형식으로 분류한 뒤 각 형식에서 30장씩 무작위 추출하여 총 180장으로 구성하였다. 표현 형식별 균등 구성을 채택한 것은 두 모델의 비교 결과가 특정 형식의 표본 비중에 좌우되지 않도록 하고, 표현 형식별 하위

집합 분석을 가능하게 하기 위해서이다. 분류에는 정감(AFF)·일상(DAY)·극적(DRM)·충족(FUL)·제품(PRD)·감각(SEN)의 여섯 범주를 적용하였다. 표현 형식 분류는 여섯 형식의 조작적 정의와 대표 사례, 경계 사례 판정 기준을 담은 코딩 매뉴얼을 기준으로 수행하였다. 광고사진 실무 경력을 가진 코더 2인이 사전 교육을 거쳐 독립적으로 분류하였고, 두 코더의 판단이 일치하지 않은 사례는 실무 경력을 가진 조정자 1인이 최종 판정하였다. 한 이미지가 둘 이상의 표현 형식 성격을 함께 지니는 경우에는 가장 지배적으로 작동하는 표현 형식 하나로 배정하였다.

3-2 역설계 프롬프트와 이미지 생성

평가 대상 상용 생성 모델은 Midjourney V7(2025년 4월 공개)과 Nano Banana Pro(모델명 Gemini 3 Pro Image, 2025년 11월 공개)이며, 두 모델은 자료 수집 시점의 안정화 버전으로 고정하였다. 180개 비교 세트의 이미지 생성과 표현 형식 분류는 2026년 3월에, 전문가 평가는 2026년 4월에 수행하였다. 프롬프트는 핵심 피사체, 재질과 텍스처, 조형적 구도, 광학적 조명, 카메라 기술 조건의 다섯 요소를 쉼표로 연결한 단일 문단으로 작성하였다. 작성 시 원본에서 관찰 가능한 시각적 사실만 포함하고, 영문 50~80단어 범위로 길이를 통일하며, 특정 모델의 고유 파라미터나 가중치 표기를 배제하여 두 모델에 동일한 입력 구조를 제공하는 원칙을 적용하였다. 초안은 대규모 언어 모델을 보조 도구로 작성한 뒤 연구자가 검토·수정하였다. 공모전 수상작은 이미지가 광고의 중심을 차지하므로, 커피와 슬로건은 재현 대상으로 삼지 않았다.

Midjourney V7은 동일 프롬프트에 대해 복수 이미지를 출력하므로, 생성의 확률적 우연성 편향을 줄이기 위해 후보 4장 가운데 주요 피사체 누락 여부, 원본의 표현 형식 반영 정도, 형태 왜곡 여부, 화면비와 구도 유지 여부를 기준으로 한 장을 선정하였다[12]. Nano Banana Pro는 화면비 출력의 일관성을 확보하기 위해 API 방식으로 이미지를 생성하였다. 최종 평가 자료는 원본 광고사진 180장, Midjourney V7 생성 이미지 180장, Nano Banana Pro 생성 이미지 180장으로, 총 540장이 180개의 비교 세트를 형성한다.

3-3 적용 자동 평가 지표

본 연구는 네 가지 자동 평가 지표를 적용한다. AD-LPIPS와 AD-Score는 광고사진 전문가의 판단을 반영하여 구성된 광고사진 도메인 특화 지표이고, 범용 LPIPS와 CLIPScore는 비교 기준으로 사용하는 범용 지표이다. 두 AD 지표는 범용 이미지 평가 지표가 광고사진의 표현 형식, 제품 설명력, 시각적 설득력, 전문가 판단을 충분히 반영하지 못한다는 문제의식에서 구성되었다.

두 지표는 본 연구의 독립 테스트셋과 분리된 별도의 학습

자료에서 구성되었다. 학습 자료는 국내 광고 포털에서 수집한 원본 광고사진과 두 상용 모델의 생성 이미지로 이루어진 비교 세트이며, 광고사진 제작·콘텐츠 제작·교육 경력을 가진 전문가 7인이 판단을 제공하였다. AD-LPIPS는 동일 원본에서 생성된 두 이미지 중 어느 쪽이 더 적합한지를 고르는 2AFC 쌍대 비교를 학습 신호로 삼아, 이미지 쌍을 입력받아 선택 확률을 출력하도록 선형 보정층을 학습한 변별 점수이다. AD-Score는 창의성·의미 전달력·제품 설명력·시각적 완성도의 5점 척도 다차원 평정을 목표값으로, 시각·의미·표현 형식 특징을 입력받아 네 차원 예측값을 산출하는 회귀 지표이다.

AD-LPIPS는 LPIPS 구조를 기반으로 하되, 동일 원본 광고사진에서 생성된 두 이미지 중 어느 쪽이 광고사진 전문가의 2AFC 쌍대 비교 판단에 더 부합하는지를 예측하도록 시각 특징의 선형 보정층을 재조정된 변별 점수이다. 따라서 AD-LPIPS는 비음수성이나 삼각부등식을 만족하는 엄밀한 거리 함수라기보다, 기준 이미지가 고정된 비교 세트에서 생성 이미지 간 상대적 적합성을 구분하는 점수로 해석한다[13].

AD-Score는 광고사진 전문가의 다차원 품질 평정값을 예측하도록 구성한 보조적 회귀 지표이다. 이 지표는 생성 이미지에 대해 창의성, 의미 전달력, 제품 설명력, 시각적 완성도의 예측값을 산출하고, 네 차원의 평균을 종합 예측값으로 사용한다. 다만 AD-Score는 생성 이미지의 최종 품질을 직접 판정하는 절대 점수가 아니라, 전문가 평정값의 경향을 계산적으로 추정하는 보조 지표로 해석한다.

본 연구에서는 두 AD 지표를 학습에 사용하지 않은 독립 테스트셋에 적용하여 외부 자료에서의 산출 양상, 범용 지표와의 차이, 전문가 다수결 판단과의 관계를 검토한다. 범용 LPIPS는 광고 도메인으로 재보정하지 않은 원래의 시각 차이 지표이고[13], CLIPScore는 역설계 프롬프트와 생성 이미지 사이의 의미 적합성을 측정하는 의미 기반 지표이다[14],[15]. AD-LPIPS와 범용 LPIPS는 값이 작을수록 원본과의 시각 차이가 작은 것으로, CLIPScore와 AD-Score 종합 예측값은 값이 클수록 각각 의미 적합성과 다차원 품질 예측값이 높은 것으로 해석한다.

3-4 통계 분석과 전문가 다수결 비교

네 지표 산출값에 대한 주 분석으로 2×6 혼합설계 분산분석을 적용하였다. 반복측정 요인은 생성 모델 두 수준, 집단 간 요인은 표현 형식 여섯 수준으로 구성하고, 생성 모델과 표현 형식의 주효과 및 두 요인의 상호작용을 검정하였다. 분산분석에 앞서 등분산성은 Levene 검정으로, 정규성은 표현 형식별 Shapiro-Wilk 검정으로 점검하였으며, 형식별 표본은 30개로 균형을 맞추었다. 네 지표를 각각 종속변수로 네 차례 검정하므로 제1종 오류 누적을 통제하기 위해 Bonferroni 보정 후 유의수준을 .0125로 설정하였다[16].

생성 모델 주효과의 크기와 방향은 대응 표본 t-검정과 Wilcoxon 부호 순위 검정, 그리고 짝지어진 차이에 대한 Cohen dz로 구체화하였다.

자동 지표 판정과 전문가 판단의 관계를 검토하기 위해, 연구에 참여하지 않은 광고사진 전문가 3인이 180개 비교 세트를 2AFC 방식으로 평가하였다. 평가 질문과 방식은 학습 자료 수집 시의 2AFC 평가와 동일하게 유지하였다. 각 비교 세트에서 3인 중 2인 이상이 선택한 모델을 전문가 다수결 판단으로 정의하였다. 전문가 다수결 판단이 한 모델에 치우칠 수 있으므로, 단순 일치율만이 아니라 다수 범주 기준선과 균형 정확도, 범주별 재현율, Cohen의 κ , MCC를 함께 산출하였다[8],[17]. 전문가 3인의 평가 자료 자체에 대한 평가자 간 신뢰도는 Fleiss의 κ 와 Krippendorff의 α 로 확인하였다[18]-[20].

3-5 프롬프트 개선 루프 설계

프롬프트 개선 루프는 생성·평가·수정·재생성이 정해진 절차에 따라 반복되는 구조로 구성하였다. 자동 지표의 피드백이 개선에 기여했는지를 분리해 확인하기 위해, 같은 원본과 같은 초기 프롬프트에서 출발하되 피드백 신호만 달린 두 루프를 비교하였다. AD 지표 루프는 AD-LPIPS와 AD-Score를, 대조 루프는 범용 LPIPS와 CLIPScore를 피드백 신호로 사용하며, 프롬프트를 수정하는 언어 모델과 이미지 생성 모델, 반복 횟수는 두 루프에서 동일하게 두었다.

지표 산출값은 그 자체로 전달하지 않고 사전에 고정된 규칙에 따라 수정 지시로 변환하여 언어 모델에 전달하였다. AD-Score는 네 차원 예측값 중 가장 낮은 차원을 취약 차원으로 지정하여 해당 차원의 수정 지시로, AD-LPIPS와 범용 LPIPS는 원본의 구도·조명·색조·분위기에 근접하라는 지시로, CLIPScore는 프롬프트와 이미지의 의미 적합성을 유지하라는 지시로 변환하였다. 수정 지시가 상충할 경우 원본 광고사진의 표현 형식 유지를 최상위 제약으로 두었다. 실험 표본은 표현 형식별로 시각적 특성이 뚜렷한 원본 3장씩 모두 18장으로 구성하고, 이미지 생성 모델은 Nano Banana Pro로 고정하여 프롬프트 개선을 두 차례 반복하였다.

두 루프의 최종 결과는 자동 지표 외부의 준거로 확인하였다. 연구에 참여하지 않은 광고사진 전문가 3인이 두 루프의 최종 이미지를 2AFC 방식으로 비교하여, 각 사례에서 더 적절한 광고사진을 선택하였다. 표본이 18장으로 작으므로 해석은 통계적 유의성보다 변화의 방향과 사례별 양상, 전문가의 선택 비율을 중심으로 하며, 정확 이항 검정과 Fleiss의 κ 는 보조적 참고값으로 제시한다.

IV. 연구결과

4-1 네 자동 평가 지표의 기술 통계

180개 비교 세트에 대해 네 지표를 산출한 결과, 전체 평균 수준에서 네 지표 모두 Nano Banana Pro 우위 방향의 값을 보였다. AD-LPIPS와 범용 LPIPS는 Nano Banana Pro에서 더 낮은 값을, CLIPScore와 AD-Score 종합 예측값은 Nano Banana Pro에서 더 높은 값을 산출하였다.

표 2. 네 자동 평가 지표의 기술 통계

Table 2. Descriptive statistics of the four automatic evaluation metrics

Metric	Midjourney V7 Mean (SD)	Nano Banana Pro Mean (SD)	Mean Difference
AD-LPIPS Score	0.1089 (0.0153)	0.1010 (0.0161)	-0.0078
General-Purpose LPIPS Score	0.6066 (0.0976)	0.5757 (0.1004)	-0.0309
CLIPScore	0.7204 (0.0907)	0.7365 (0.0865)	+0.0161
AD-Score Mean Prediction	3.5066 (0.8615)	3.7222 (0.7915)	+0.2155

4-2 2×6 혼합설계 분산분석 결과

분산분석에 앞서 등분산성은 네 지표 모두에서 충족되었고, 정규성은 일부 형식에서 경미한 위반이 나타났으나 형식별 표본이 30개로 균형을 이루어 분석에 무리가 없는 수준이었다.

생성 모델의 주효과는 네 종속변수 모두에서 유의하였다. AD-LPIPS 변별 점수는 $F(1, 174) = 72.04, p < .001, \eta_p^2 = .293$ 의 큰 효과를, 범용 LPIPS는 $F(1, 174) = 47.33, p < .001, \eta_p^2 = .214$ 의 큰 효과를 보였다. AD-Score 종합 예측값은 $F(1, 174) = 15.00, p < .001, \eta_p^2 = .079$ 의 중간 효과를, CLIPScore는 $F(1, 174) = 7.59, p = .007, \eta_p^2 = .042$ 의 작은 효과를 나타냈다. 두 상용 생성 모델의 산출값이 AD-LPIPS와 AD-Score에서 유의한 차이를 보일 것이라는 연구가설은 지지되었다.

모델과 표현 형식의 상호작용은 네 종속변수 어디에서도 유의수준에 도달하지 않았다. AD-Score 종합 예측값의 상호작용이 $F(5, 174) = 2.14, p = .063, \eta_p^2 = .058$ 로 보정 전 유의수준에 근접하였으나 유의하지는 않았으며, 네 상호작용의 η_p^2 는 .036에서 .058 사이로 효과 크기 자체도 작았다. 두 모델 간 산출값의 차이가 표현 형식에 따라 달라질 것이라는 연구가설은 지지되지 않았다. 이 결과는 표현 형식에 따른 조절효과가 존재하지 않음을 확정하는 것이 아니라, 형식별 표본 30개 규모에서 조절효과를 탐지할 통계적 증거가 확보되지 않았음을 의미한다. 동시에 이 비유의는 그 자체로 하나의 발견에 해당한다. 두 모델 간 산출값의 차이가 특정 표현 형

식에 쏠리지 않고 여섯 형식 전반에 걸쳐 비교적 일관된 크기로 나타났다는 것이며, 모델 사이의 차이가 표현 형식에 국한된 것이 아니라 자동 지표가 포착하는 범위 안에서 두 모델의 전반적 경향 차이에 해당함을 보여준다. AD-Score 종합 예측값의 경계 근접 추세는 표본 규모를 확대한 후속 검토에서 다시 다룰 여지를 남긴다.

표 3. 독립 테스트셋의 2×6 혼합설계 분산분석 결과

Table 3. Results of the 2 × 6 mixed-design ANOVA on the independent test set

Metric	Effect	F	df	p	η_p^2
AD-LPIPS	Expression Form	0.66	(5, 174)	.658	.019
	Generative Model	72.04	(1, 174)	< .001	.293
	Model × Expression Form	1.31	(5, 174)	.262	.036
AD-Score	Expression Form	5.33	(5, 174)	< .001	.133
	Generative Model	15.00	(1, 174)	< .001	.079
	Model × Expression Form	2.14	(5, 174)	.063	.058
LPIPS	Expression Form	5.04	(5, 174)	< .001	.127
	Generative Model	47.33	(1, 174)	< .001	.214
	Model × Expression Form	1.89	(5, 174)	.098	.052
CLIP Score	Expression Form	3.79	(5, 174)	.003	.098
	Generative Model	7.59	(1, 174)	.007	.042
	Model × Expression Form	2.05	(5, 174)	.074	.056

표현 형식의 주효과는 연구가설로 설정하지 않았으나 분산 분석 과정에서 함께 산출되었다. AD-Score 종합 예측값 ($F(5, 174) = 5.33, p < .001, \eta_p^2 = .133$), 범용 LPIPS ($F(5, 174) = 5.04, p < .001, \eta_p^2 = .127$), CLIPScore ($F(5, 174) = 3.79, p = .003, \eta_p^2 = .098$)의 세 지표에서 표현 형식 주효과가 유의하였으며, 효과 크기는 중간 수준이었다. AD-LPIPS에서는 표현 형식 주효과가 유의하지 않았는데 ($F(5, 174) = 0.66, p = .658$), 이는 AD-LPIPS가 표현 형식별로 대응된 원본과의 상대적 변별을 산출하는 지표여서 형식 간 수준 차이가 상쇄된 결과로 해석된다. 세 지표에서 확인된 표현 형식 주효과는 표현 형식이 자동 평가 지표 산출값을 체계적으로 가르는 분석 단위임을 보여준다. 모델 격차의 조절변수로는 기능하지 않았으나, 표현 형식 자체가 지표 산출값을 구분하는 유의한 변수라는 점에서 표현 형식을 분석 단위로 삼은 설계가 독립 테스트셋에서도 뒷받침된다.

대응 표본 비교에서도 네 지표 모두 두 모델 사이의 평균 차이가 유의하였다. AD-LPIPS는 $t(179) = -8.451$, 범용 LPIPS는 $t(179) = -6.796$ 으로 큰 폭의 차이를 보였고, AD-Score는 $t(179) = 3.426(p = .0008)$, CLIPScore는 $t(179) = 2.716(p = .0073)$ 이었다. Cohen dz는 AD-LPIPS -0.630, 범용 LPIPS -0.507, AD-Score +0.255, CLIPScore +0.202로, 네 지표의 95% 신뢰구간이 모두 0을

표 4. 네 자동 평가 지표의 모델 간 비교 결과

Table 4. Model comparison results for the four automatic evaluation metrics

Metric	t(179)	Cohen's dz	dz 95% CI	NB Preference Rate	Exact Binomial p
AD-LPIPS	-8.451	-0.630	[-0.800, -0.486]	81.1%	9.8×10^{-18}
LPIPS	-6.796	-0.507	[-0.655, -0.368]	71.1%	1.4×10^{-8}
CLIPScore	+2.716	+0.202	[+0.056, +0.341]	57.2%	.062
AD-Score	+3.426	+0.255	[+0.115, +0.417]	60.0%	.0089

포함하지 않았다. 다만 CLIPScore의 Wilcoxon 검정 p값은 .0143으로 Bonferroni 보정 기준선 안쪽까지는 들어서지 못하여, 순위 기반 검정에서는 상대적으로 약한 차이로 해석된다.

비교 세트별 우세 모델 판정에서 AD-LPIPS와 범용 LPIPS, AD-Score는 Nano Banana Pro 우세 비율이 무작위 기대값과 유의하게 달랐으나, CLIPScore는 우세 비율이 57.2%로 보정 기준에 도달하지 않았다. 지각 계열인 AD-LPIPS와 범용 LPIPS의 우세 모델 판정 일치율은 84.4%로 높았던 반면, CLIPScore와 AD-Score의 일치율은 56.1%, AD-LPIPS와 CLIPScore의 일치율은 57.2%에 머물렀다. 네 지표가 모두 같은 모델을 우세로 판정한 비교 세트는 32.8%에 그쳐, 전체의 약 3분의 2에서 적어도 한 지표가 다른 우세 모델을 판정하였다. 자동 지표 사이의 이러한 불일치는 지각 차이·의미 적합성·다차원 품질 예측이 서로 다른 정보를 제공하는 데서 비롯되며, 단일 지표가 광고 이미지 평가의 관련 차원을 포괄적으로 대표하기 어렵다는 점을 보여준다.

4-3 표현 형식별 하위 집합 분석

상호작용이 유의하지 않았으므로 표현 형식별 하위 집합 분석은 조절효과를 검정하는 확증적 분석이 아니라, 여섯 형식 각각에서 두 모델의 차이가 어떤 크기와 방향으로 나타나는지를 살피는 탐색적 분석이다. AD-LPIPS는 감각 표현형을 제외한 다섯 형식에서 Nano Banana Pro 우위가 보정 후에도 유의하였고, 효과 크기는 정감(dz = -0.928)과 제품(dz = -0.858)에서 컸으며 충족·극적·일상에서 중간 수준이었다. 감각 표현형은 dz = -0.304(p = .106)로 유의 영역에 이르지 못하였다.

표 5. 표현 형식별 두 모델 비교의 효과 크기(Cohen dz)

Table 5. Effect sizes of model comparisons by expression form (Cohen's dz)

Expression Form	AD-LPIPS	LPIPS	CLIPScore	AD-Score
AFF	-0.928 ✓	-0.700 ✓	+0.273	+0.336
DAY	-0.513 ✓	-0.426	-0.091	+0.514 ✓
DRM	-0.619 ✓	-0.606 ✓	+0.125	+0.117
FUL	-0.682 ✓	-0.519 ✓	+0.156	-0.252
PRD	-0.858 ✓	-0.767 ✓	+0.709 ✓	+0.529 ✓
SEN	-0.304	-0.050	+0.045	+0.335

지각 계열 지표에서 음의 dz는 Nano Banana Pro의 산출값 감소를, 의미·품질 계열 지표에서 양의 dz는 Nano Banana Pro의 값 증가를 의미하며, ✓는 Bonferroni 보정 후 유의를 가리킨다. 제품 표현형은 네 지표 모두가 Nano Banana Pro 우위를 가장 일관되게 판정한 형식이었고, CLIPScore가 유의한 차이를 보인 유일한 형식이기도 하였다. 반대로 감각 표현형은 네 지표 어디에서도 보정 후 유의한 차이가 확인되지 않았는데, 색감·질감·분위기처럼 시각적으로 미묘한 속성을 단서로 삼는 형식에서는 모델 간 차이를 안정적으로 측정하기 어려울 가능성을 보여준다. 충족 표현형에서는 지각 계열 지표가 Nano Banana Pro 우위를 판정한 반면 AD-Score는 Midjourney V7 방향(dz = -0.252)으로 산출되어, 평가 관점에 따라 우위 방향이 달라질 수 있음이 탐색적으로 관찰되었다. 이러한 결과는 어떤 모델이 전반적으로 우수한가라는 평균적 질문보다, 어떤 표현 형식에서 어떤 평가 관점에 따라 우위가 나타나는가라는 조건부 질문이 광고 이미지 평가에 더 적절함을 시사한다.

4-4 전문가 다수결 판단과의 일치 및 불일치

전문가 3인의 2AFC 평가 자료는 비교적 높은 합의 구조 위에서 형성되었다. Fleiss의 κ 는 0.7665, Krippendorff의 α 는 0.7669, 평균 Cohen의 κ 는 0.7676으로 상당한 수준의 일치에 해당하였고, 세 평가자가 모두 같은 모델을 선택한 강한 합의 비교 세트가 154개(85.6%)였다. 전문가 다수결 판단의 우세 모델 분포는 Nano Banana Pro 129개(71.7%), Midjourney V7 51개(28.3%)로 Nano Banana Pro 쪽에 치우쳤다.

자동 지표와 전문가 다수결 판단의 단순 일치율은 AD-LPIPS 70.6%, 범용 LPIPS 67.2%, CLIPScore 62.2%, AD-Score 52.8% 순이었다. 그러나 전문가 다수결에서 Nano Banana Pro가 71.7% 선택되었으므로, 항상 Nano Banana Pro를 선택하는 다수 범주 기준선의 일치율도 71.7%에 이른다. AD-LPIPS의 단순 일치율 70.6%는 이 기준선을 넘지 못하므로, 단순 일치율만으로 자동 지표의 대응 관계를 해석할 수 없다.

자료 분포를 보정한 지표를 함께 보면 양상이 달라진다. 균형 정확도는 CLIPScore가 61.2%로 가장 높았고 범용 LPIPS 59.9%, AD-LPIPS 58.7% 순이었으며, Cohen의 κ 는 세 지표 모두 0.19대의 약한 양의 일치 수준이었다.

표 6. 자동 지표와 전문가 다수결 판단의 일치도 및 보정 지표

Table 6. Agreement and adjusted indices between automatic metrics and expert majority judgments

Metric	Simple Agreement	NB Recall	MJ Recall	Balanced Accuracy	Cohen's κ	MCC
Always-NB Baseline	71.7%	100.0%	0.0%	50.0%	0.000	—
AD-LPIPS	70.6%	86.0%	31.4%	58.7%	0.193	0.200
LPIPS	67.2%	76.7%	43.1%	59.9%	0.197	0.197
CLIPScore	62.2%	63.6%	58.8%	61.2%	0.194	0.204
AD-Score	52.8%	58.9%	37.3%	48.1%	-0.034	-0.035

AD-Score는 균형 정확도 48.1%, Cohen의 κ -0.034, MCC -0.035로 전문가 다수결 판단과의 일치가 무작위 기대값을 넘었다고 보기 어려웠다. 이는 원본 광고사진의 다차원 품질 평정값을 예측하도록 학습된 회귀 지표가 AI 생성 이미지 비교 상황에서는 동일한 수준의 대응을 보이지 못함을 의미한다. 보정 지표를 종합하면, AD-LPIPS가 범용 LPIPS나 CLIPScore에 견주어 명확한 우위를 점한다는 결론에는 이르기 어렵다. 따라서 독립 테스트셋에서 자동 지표가 산출한 결과는 전문가 판단과의 부분적 대응 가능성을 가리키는 자료에 한정되며, 비교 조건 전반에서 안정적 우월성을 갖는다는 일반화로 확장되지 않는다.

전문가 간 합의 강도에 따라 자동 지표의 일치율이 달라지는 양상도 관찰되었다. 세 평가자가 모두 같은 모델을 선택한 강한 합의 비교 세트에서는 네 지표 모두 전체 평균과 비슷하거나 높은 일치율을 보인 반면, 2:1로 다수결이 형성된 약한 합의 비교 세트 26개에서는 일치율이 전반적으로 낮아졌다. 전문가들 사이에서도 판단이 갈리는 비교 세트에서는 자동 지표 역시 전문가 다수결 판단과 안정적으로 일치하기 어려웠으며, 이 분석은 약한 합의 표본 수가 작아 탐색적 결과로 해석된다.

4-5 프롬프트 개선 루프 결과

표현 형식별 3장씩 18장을 대상으로 AD 지표 루프와 대조 루프를 수행하였다. 두 루프의 최종 이미지에 대한 자동 지표 산출값은 뚜렷한 차이를 보이지 않았다. AD-Score 종합 예측값은 초기 이미지 3.29에서 AD 지표 루프 최종 3.44, 대조 루프 최종 3.17이었고, 18장 가운데 AD 지표 루프가 더 높은 사례가 10장, 대조 루프가 더 높은 사례가 8장이었다. AD-LPIPS와 CLIPScore의 단계별 평균에서도 두 루프 사이의 차이는 작았다. 자동 지표만 놓고 보면 어느 루프가 일관되게 우세하다고 보기 어려웠다.

전문가 판단에서는 다른 양상이 나타났다. 두 루프의 최종 이미지를 비교한 전문가 3인의 2AFC 다수결에서 AD 지표 루프의 최종 이미지가 선택된 사례가 14개, 대조 루프가 선택된 사례가 4개였고, 54개의 개별 판단에서는 AD 지표 루프가 41회(75.9%) 선택되었다. 18개 사례의 다수결에 대한 정확도 이항 검정에서 양측 p 는 약 .031로, 무작위 기대 수준에서 벗어난 유효성이 확인되었다.

표 7. 프롬프트 개선 루프의 두 루프 비교 결과

Table 7. Comparison results between the two prompt-improvement loops

Comparison Item	Initial Image	AD-Metric Loop	Control Loop
AD-Score Mean Prediction	3.29	3.44	3.17
Mean AD-LPIPS Discrimination Score	0.108	0.102	0.101
Mean CLIPScore	0.748	0.794	0.773
Number of Final-Image Wins (n = 18)	—	10	8
Expert 2AFC Majority Judgment (n = 18)	—	14	4
Individual Expert Judgments (n = 54)	—	41 (75.9%)	13 (24.1%)

다만 전문가 사이의 판단 일치도는 높지 않았다. 세 평정자의 판단에 대한 Fleiss의 κ 는 약 -0.07로 우연 수준을 넘는 일치를 보이지 않았다. 다수결의 방향성과 평정자 간 일치도는 서로 다른 것을 측정하므로, 다수결이 한쪽으로 기울면서도 개별 평정자의 판단은 갈릴 수 있다. 표현 형식별 다수결에서는 정감·극적·충족 표현형의 세 사례 전부와 일상 표현형의 두 사례에서 AD 지표 루프가 선택되었고, 제품과 감각 표현형에서는 각각 두 사례에서 대조 루프가 선택되었다. 원본의 시각적 속성을 충실히 재현하는 것이 평가에서 큰 비중을 차지하는 제품·감각 표현형에서는 원본 근접을 직접 겨냥하는 범용 LPIPS의 피드백이 상대적으로 유리하게 작용하였을 가능성이 있으나, 형식별 사례 수가 3개에 그쳐 사례 수준의 관찰로 제한된다.

자동 지표 층위에서 두 루프가 대등하게 나타난 반면 전문가 다수결에서 AD 지표 루프가 우세하였다는 결과는, AD 지표를 피드백으로 사용한 개선이 AD-Score 예측값의 변화만으로는 드러나지 않는 질적 차이를 만들어낼 가능성을 시사한다. 다만 표본이 18장에 그치고 전문가 사이의 일치도가 낮으므로, 이 결과는 AD 지표의 우월성을 확정하는 근거가 아니라 그 우위 가능성을 보여주는 예비적 관찰에 해당한다.

4-6 종합 논의

1) 독립 테스트셋에서 자동 지표의 적용 가능성

두 모델의 전반적 차이는 네 지표 모두에서 유의하게 포착

되었고, 지각 계열 지표에서 큰 효과 크기로 나타났다. 학습에 사용하지 않은 외부 자료에서도 자동 지표가 모델 간 경향 차이를 안정적으로 잡아낸다는 점은 적용 가능성의 한 근거가 된다. 그러나 같은 결과가 전문가 판단과의 일치로 직결되지는 않았다. 보정 지표 기준에서 자동 지표 사이의 우열이 분명하지 않았고, 가장 높은 단순 일치율을 보인 AD-LPIPS도 다수 범주 기준선을 넘지 못하였다. 모델 간 차이를 포착하는 능력과 전문가 판단에 대응하는 능력은 별개의 문제이며, 외부 적용 가능성의 해석은 후자에서 제한적이다.

2) 표현 형식별 해석의 필요성

전체 평균 수준의 비교만으로는 자동 지표의 양상이 충분히 드러나지 않았다. 모델과 표현 형식의 상호작용이 유의하지 않아 두 모델의 격차는 여섯 형식 전반에 일관되게 나타났지만, 표현 형식의 주효과가 세 지표에서 유의하여 표현 형식 자체는 지표 산출값을 체계적으로 가르는 변수였다. 표현 형식별 하위 집합에서도 제품 표현형은 네 지표가 가장 일관되게 같은 방향을 가리킨 반면 감각 표현형은 어느 지표도 유의한 차이를 잡지 못하였고, 충족 표현형에서는 지각 계열 지표와 품질 예측 지표의 방향이 엇갈렸다. 광고 이미지 평가에서 자동 지표를 적용할 때 전체 평균과 함께 표현 형식 단위의 해석이 병행되어야 하는 이유가 여기에 있다.

3) 자동 지표와 전문가 판단의 불일치

자동 지표 사이의 불일치와 자동 지표·전문가 판단 사이의 불일치는 오류라기보다 측정 대상의 차이에서 비롯된다. 지각 차이, 의미 적합성, 다차원 품질 예측은 서로 다른 정보를 담으므로, 네 지표가 같은 비교 세트에서 다른 우세 모델을 가리키는 일이 빈번하였다. AD-Score가 내부 검증에서는 전문가 평정값과 대응하였으나 생성 이미지 비교 상황에서 전문가 다수결과의 일치가 무작위 수준에 머문 결과는, 원본 광고사진의 품질 예측과 생성 이미지의 상대적 비교가 다른 과제임을 보여준다. 이러한 불일치는 단일 지표로 광고 이미지 품질을 확정하기 어렵다는 점을, 그리고 자동 지표가 전문가 판단을 대체할 수 없다는 점을 함께 드러낸다.

4) 프롬프트 개선 루프의 실무적 가능성

자동 지표 층위에서 두 루프가 대등하였음에도 전문가 다수결에서 AD 지표 루프가 우세하였다는 결과는, 자동 지표의 피드백이 수치로 환산되기 이전의 시각적 완성도나 장면의 설득력을 함께 조정하였을 가능성을 시사한다. 전문가의 2AFC 판단은 수치에 그대로 담기지 않는 차이를 포착하는 준거로 기능하였다. 이 결과는 자동 지표가 단순 사후 평가를 넘어 프롬프트 개선의 방향을 좁히는 보조 신호로 쓰일 수 있음을 보여주며, 자동화된 품질 개선 절차로 확대 해석되어서는 안 된다. 자동 지표는 개선의 방향을 제안하고, 그 성패의 최종 판정은 전문가 검토에 남는다.

V. 결 론

본 연구는 광고사진 전문가의 판단을 반영하여 구성된 도메인 특화 자동 평가 지표와 범용 지표를 학습에 사용하지 않은 독립 테스트셋에 적용하여, 두 상용 생성 모델의 비교와 표현 형식별 산출 양상, 전문가 판단과의 관계, 그리고 프롬프트 개선의 가능성을 검토하였다.

분석 결과, 생성 모델의 주효과는 네 지표 모두에서 유의하였고, 모델과 표현 형식의 상호작용은 어느 지표에서도 유의하지 않았다. 이는 두 모델 간 산출값의 차이가 특정 표현 형식에 집중되기보다 여섯 표현 형식 전반에서 비교적 일관되게 나타났음을 의미한다. 반면 표현 형식의 주효과는 세 지표에서 유의하게 나타나, 표현 형식이 자동 평가 지표 산출값을 체계적으로 구분하는 분석 단위로 작동할 수 있음을 확인하였다.

전문가 다수결 판단과의 비교에서는 자동 지표 사이의 명확한 우열이 보정 지표 기준에서 확인되지 않았다. 가장 높은 단순 일치율을 보인 AD-LPIPS도 다수 범주 기준선을 넘지 못하였으며, AD-Score는 생성 이미지 비교 상황에서 전문가 다수결 판단과 충분히 대응하지 못하였다. 이는 자동 지표가 모델 간 경향 차이를 포착하는 능력과 전문가 판단을 재현하는 능력이 서로 다른 문제임을 보여준다. 따라서 자동 평가 지표는 전문가 판단을 대체하는 최종 판정자가 아니라, 전문가 검토가 필요한 비교 지점을 구조화하는 보조 도구로 해석되어야 한다.

프롬프트 개선 루프에서는 자동 지표 층위에서 AD 지표 루프와 범용 지표 기반 대조 루프가 대등하게 나타났으나, 전문가 2AFC 다수결에서는 AD 지표 루프의 최종 이미지가 더 자주 선택되었다. 이는 도메인 특화 자동 지표가 프롬프트 수정 방향을 제안하는 피드백 신호로 활용될 가능성을 보여준다. 다만 이 결과는 표본 규모가 작고 전문가 간 일치도가 낮은 조건에서 얻어진 예비적 관찰이므로, 자동화된 품질 개선 절차의 우월성을 입증하는 결과로 확대 해석하기는 어렵다.

본 연구의 한계도 분명하다. 독립 테스트셋은 국제광고사진 공모전 수상작 180장으로 구성되어 모든 광고사진 장르와 문화권, 제품군을 포괄하지 못하며, 분석 대상 생성 모델도 자료 수집 시점의 두 상용 모델로 한정되었다. 또한 전문가 다수결 판단은 3인의 2AFC 평가에 근거하므로 평가자 수가 제한적이고, 카피·슬로건·타이포그래피와 소비자 반응은 분석 범위에 포함하지 않았다. 또한 표현 형식 분류는 여러 표현 전략을 동시에 지닌 이미지를 지배적 형식 하나로 배정하였고, 두 생성 모델의 후보 표집 방식이 서로 달라, 분류의 단일성과 생성 조건의 동일성 측면에서 제약이 있다. 이러한 조건은 후속 연구에서 다중 라벨 분류와 통일된 표집 절차로 보완될 수 있다. 프롬프트 개선 루프 역시 표현 형식별 3장씩 총 18장을 대상으로 한 예비 실험이므로, 그 결과는 가능성 탐색의 범위 안에서 해석되어야 한다. 향후 연구에서는 독립 테스

트셋의 규모와 표현 형식별 표본을 확대하고, 외부 평가에 참여하는 전문가 풀을 현재의 3인에서 더 넓은 규모로 확장하여 다수결의 안정성과 평가자 간 일치도를 함께 높일 필요가 있다. 프롬프트 개선 루프 또한 현재의 표현 형식별 3장 규모를 넘어 형식별 표본과 반복 횟수를 확대하고, 최신 세대의 생성 모델을 포함함으로써 자동 지표와 전문가 판단의 관계를 재검증하고자 한다. 이러한 후속 작업을 통해 자동 평가 지표는 전문가 판단을 대체하는 도구가 아니라, 생성형 AI 광고 제작 과정에서 전문가 검토를 보조하고 그 초점을 좁히는 실무적 장치로 발전할 수 있을 것이다.

감사의 글

본 논문은 저자의 박사학위논문 중 일부 내용을 발췌하여 수정·보완한 것임.

참고문헌

- [1] H. Sim and J. I. Choi, "The Effect of the Expression Form of Advertising Photos on Advertising Effectiveness - Focusing on Print Advertisements for Domestic Home Appliances," *The Society of Modern Photography & Video*, Vol. 24, No. 2, pp. 55-85, 2021.
- [2] S. Hartwig, D. Engel, L. Sick, H. Kniesel, T. Payer, P. Poonam, ... and T. Ropinski, "A Survey on Quality Metrics for Text-to-Image Generation," *IEEE Transactions on Visualization and Computer Graphics*, Vol. 31, No. 10, pp. 9464-9483, October 2025. <https://doi.org/10.1109/TVCG.2025.3585077>
- [3] D. Ghosh, H. Hajishirzi, and L. Schmidt, "GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment," in *Advances in Neural Information Processing Systems*, Vol. 36, pp. 52132-52152, 2023. <https://doi.org/10.52202/075280-2270>
- [4] Y. Hu, B. Liu, J. Kasai, Y. Wang, M. Ostendorf, R. Krishna, and N. A. Smith, "TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris, France, pp. 20406-20417, October 2023.
- [5] A. Aghazadeh and A. Kovashka, "CAP: Evaluation of Persuasive and Creative Image Generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Honolulu, HI, pp. 16970-16980, October 2025.
- [6] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, ... and Y. Dong, "ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation," in *Advances in Neural Information Processing Systems*, Vol. 36, pp. 15903-15935, 2023. <https://doi.org/10.52202/075280-0700>
- [7] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy, "Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation," in *Advances in Neural Information Processing Systems*, Vol. 36, pp. 36652-36663, 2023. <https://doi.org/10.52202/075280-1594>
- [8] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 37-46, April 1960. <https://doi.org/10.1177/001316446002000104>
- [9] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, ... and P. F. Christiano, "Learning to Summarize with Human Feedback," in *Advances in Neural Information Processing Systems*, Vol. 33, pp. 3008-3021, 2020.
- [10] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegrefe, ... and P. Clark, "Self-Refine: Iterative Refinement with Self-Feedback," in *Advances in Neural Information Processing Systems*, Vol. 36, pp. 46534-46594, 2023. <https://doi.org/10.5555/3666122.3668141>
- [11] Y. Hao, Z. Chi, L. Dong, and F. Wei, "Optimizing Prompts for Text-to-Image Generation," in *Advances in Neural Information Processing Systems*, Vol. 36, pp. 66923-66939, 2023. <https://doi.org/10.52202/075280-2923>
- [12] A. Beirami, A. Agarwal, J. Berant, A. D'Amour, J. Eisenstein, C. Nagpal, and A. T. Suresh, "Theoretical Guarantees on the Best-of-N Alignment Policy," in *Proceedings of the 42nd International Conference on Machine Learning*, Vancouver, Canada, PMLR 267, pp. 3580-3602, 2025.
- [13] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City: UT, pp. 586-595, June 2018. <https://doi.org/10.1109/CVPR.2018.00068>
- [14] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A Reference-Free Evaluation Metric for Image Captioning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic, pp. 7514-7528, November 2021. <https://doi.org/10.18653/v1/2021.emnlp-main.595>
- [15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, ... and I. Sutskever, "Learning Transferable

Visual Models from Natural Language Supervision,” in *Proceedings of the 38th International Conference on Machine Learning*, PMLR 139, pp. 8748-8763, 2021.

- [16] C. E. Bonferroni, *Teoria Statistica delle Classi e Calcolo delle Probabilità*, Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze, No. 8. Florence, Italy: Seeber, 1936.
- [17] B. W. Matthews, “Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme,” *Biochimica et Biophysica Acta*, Vol. 405, No. 2, pp. 442-451, October 1975. [https://doi.org/10.1016/0005-2795\(75\)90109-9](https://doi.org/10.1016/0005-2795(75)90109-9)
- [18] J. L. Fleiss, “Measuring Nominal Scale Agreement Among Many Raters,” *Psychological Bulletin*, Vol. 76, No. 5, pp. 378-382, November 1971. <https://doi.org/10.1037/h0031619>
- [19] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 4th ed. Thousand Oaks, CA: SAGE, 2018.
- [20] J. R. Landis and G. G. Koch, “The Measurement of Observer Agreement for Categorical Data,” *Biometrics*, Vol. 33, No. 1, pp. 159-174, March 1977. <https://doi.org/10.2307/2529310>



심현준(Hyun-Jun Sim)

2022년 : 상명대학교 대학원
(예술학박사)

2026년 : 홍익대학교 대학원
(미술학박사)

2006년~현 재: (주)트리니티팩토리 대표

2022년~현 재: (사)한국광고사진가협회 회장

2024년~현 재: 배재대학교 광고사진영상학과 초빙교수

※관심분야: 광고사진, 생성형 AI, 실감미디어콘텐츠