

# CLIP 유사도 기반 멀티모달 RAG 검증 기법

## CLIP similarity-based Multimodal RAG verification method

김은오, 이창근, 윤수연\*

(주)엔투솔루션, \*국민대학교

2025. 02. 07

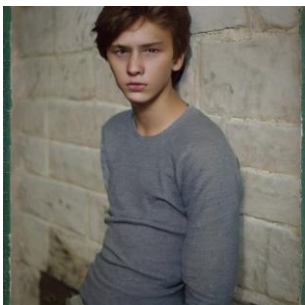


# A table of Contents

1. Instruction
2. Research Background
3. Research Questions
4. Improvement Measures
5. Conclusions

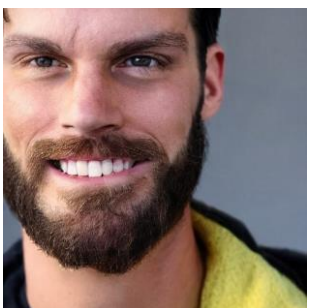


## AI의 발전으로 딥페이크 콘텐츠의 위협 증가 다양한 분야(이미지, 영상 등)에서 정교한 조작 콘텐츠가 빠르고 손쉽게 생성·유포



브래드 피트 딥페이크로 53세 프랑스 여성 12억 사기 당해

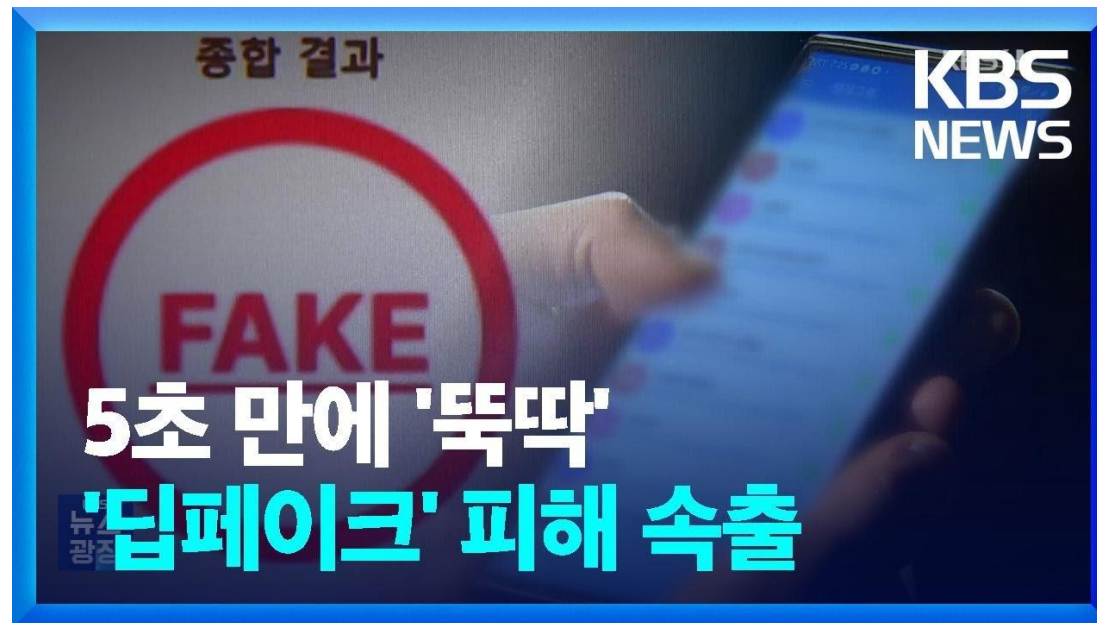
※ 저작권 금지 ※ 입력 2025.01.15 18:00 ※ 수정 2025.01.15 19:03 ※ 댓글 0 ※ 좋아요 1



딥페이크로 생성된 얼굴 이미지



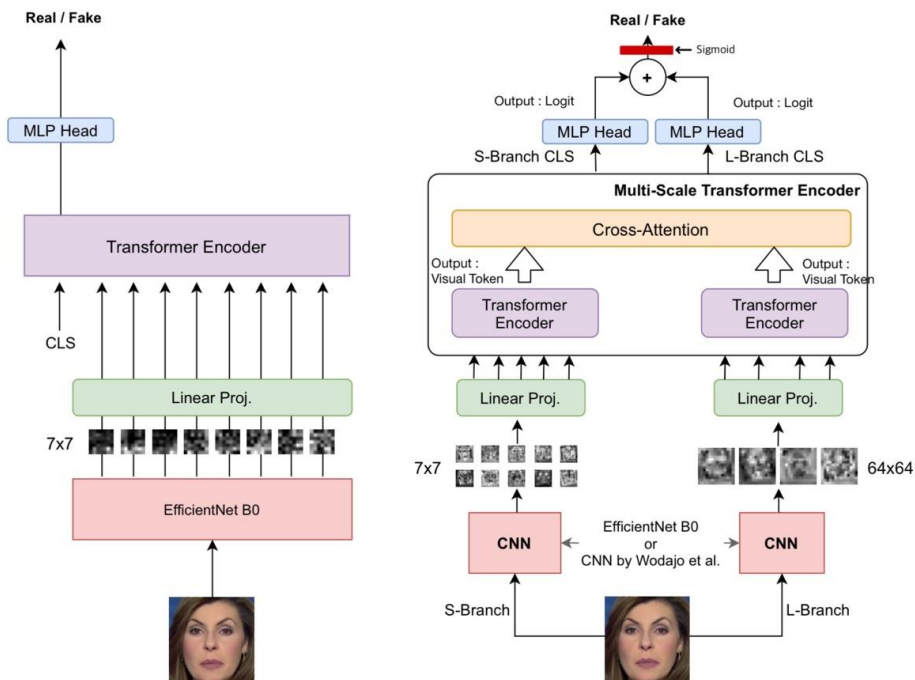
딥페이크 피해 사례(1)



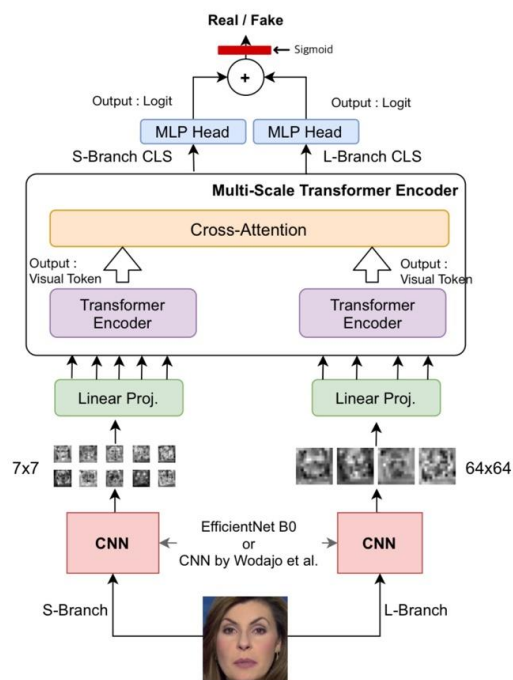
딥페이크 피해 사례(2)



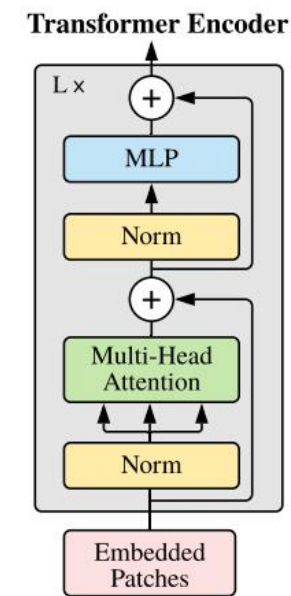
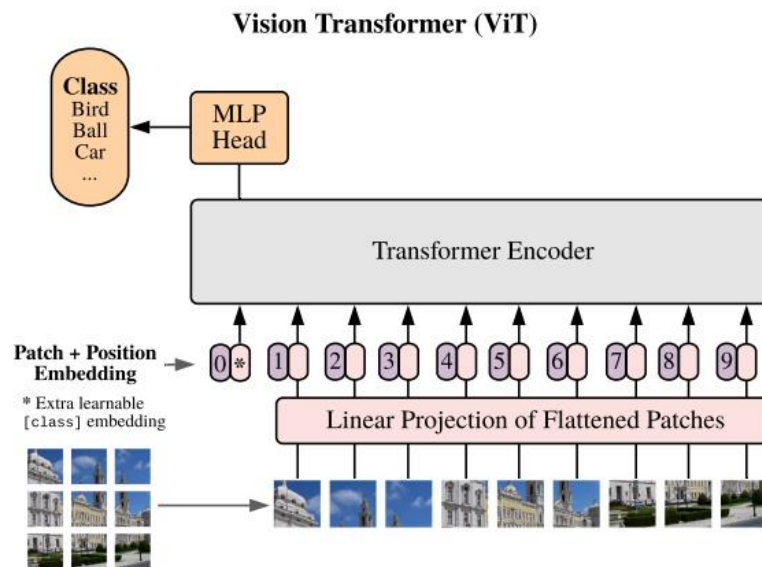
기존 머신러닝, 딥러닝 기반의 탐지는 복잡한 설정과 고도의 전문지식을 필요로 함.



(a) Efficient ViT architecture.



(b) Convolutional Cross ViT architecture.





LMM은 전문 지식이 없는 사용자도 손쉽게 접근할 수 있을 뿐 아니라, 이해 가능한 자연어 형태의 판단 근거 제공으로 투명성 및 신뢰성을 확보할 수 있다.



OpenAI  
GPT-4o



이 이미지는 딥페이크로 조작된 이미지야? 예 아니오로 답하고, 딥페이크 이미지라면 근거를 함께 설명해줘



예, 이 이미지는 딥페이크로 조작된 이미지로 보입니다. 근거는 다음과 같습니다:

근거:

1. 피부 텍스처와 디테일의 부자연스러움:

- 얼굴의 피부가 비정상적으로 매끄럽고, 자연스러운 모공이나 주름이 부족합니다. 특히, 눈, 코, 입 주변의 디테일이 흐릿하거나 인위적입니다.

2. 머리카락과 헤어라인의 부조화:

- 머리카락과 두피의 경계가 흐릿하고, 자연스러운 텍스처를 제대로 표현하지 못한 부분이 있습니다. 딥페이크는 머리카락 세부 디테일을 재현하는 데 약점이 있습니다.

3. 조명과 그림자의 불일치:

- 얼굴에 비치는 조명과 주변 배경 조명의 방향이 일치하지 않는 모습입니다. 얼굴과 목에 조명 효과가 부자연스럽습니다.

4. 표정의 부자연스러움:

- 입 모양은 웃는 듯하지만 눈이나 다른 표정 요소가 감정을 따르지 않는 부조화를 보입니다. 특히, 입술 주변의 경계선이 뚜렷하지 않고 흐릿합니다.

5. 배경과 얼굴의 이질감:

- 얼굴과 배경이 부자연스럽게 분리된 느낌을 줍니다. 얼굴 주위 배경의 흐릿한 부분이 딥페이크 합성 과정에서 발생한 흔적일 가능성이 높습니다.

이러한 특징들은 딥페이크 알고리즘으로 생성된 이미지에서 흔히 나타나는 징후들로, 해당 이미지가 조작된 것임을 강하게 시사합니다.



## LMM의 환각 현상 방지 및 정확도 향상을 위한 멀티모달 RAG

LLM의 고질적 환각 문제

RAG 기법을 통한 환각 완화

멀티모달(LMM) 환경에서의  
RAG 연구 부족

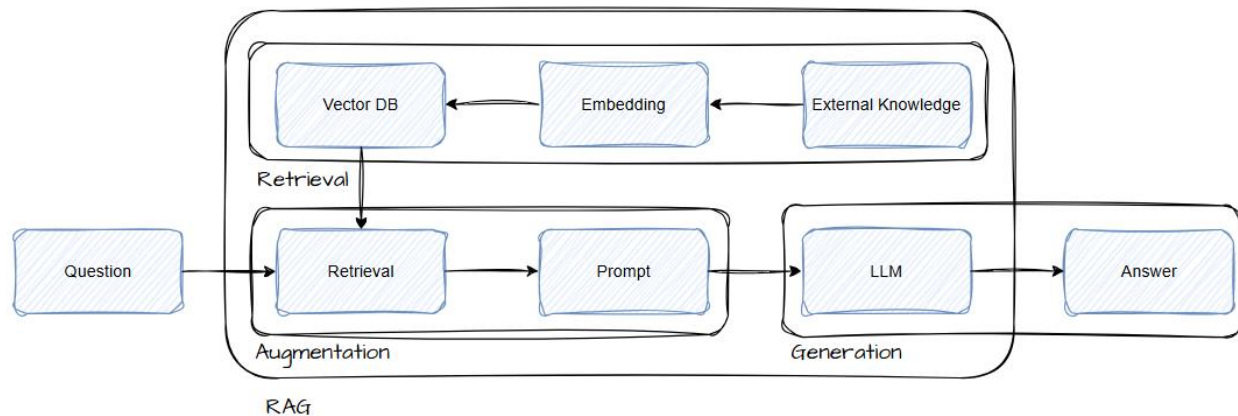


CLIP 기반 멀티모달 RAG 기법





## RAG에서의 핵심은 Retrieve1이며 이를 가능케 하는것이 CLIP 모델의 이미지-텍스트 유사도 계산



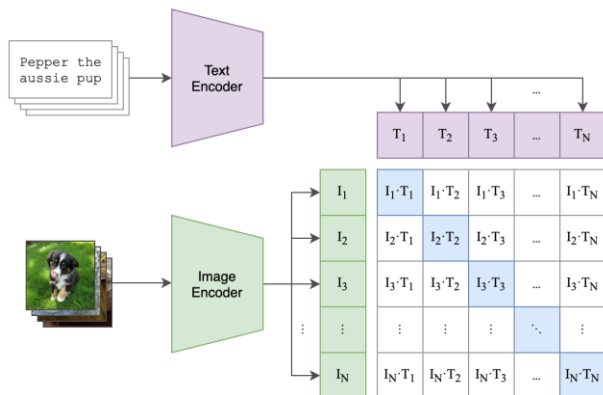
### 1) RAG(검색 증강 생성)

- ✓ LLM의 성능을 강화하기 위해 외부 지식 소스(External Knowledge Sources)를 통합하는 기술.
- ✓ 검색, 증강, 생성 세 단계로 작동한다.
- ✓ 이를 통해 정확하고 신뢰할 수 있는 응답 제공

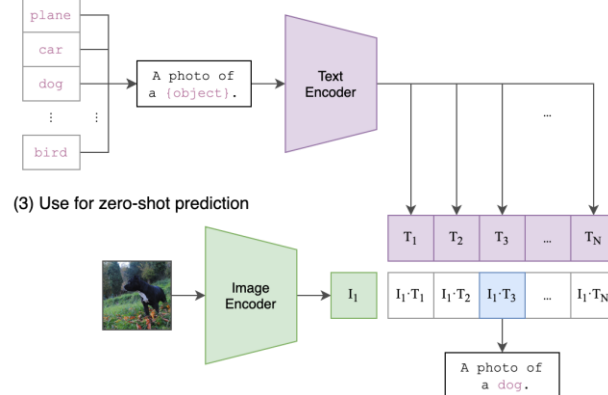
### 2) CLIP Model

- ✓ 약 4억 개의 대규모 이미지-텍스트 쌍을 대조 학습 (Contrastive Learning)
- ✓ 이미지와 텍스트 간 의미를 코사인 유사도로 계산.
- ✓ 이미지 내 시각적 특징을 텍스트 형태의 키워드나 설명문으로 효율적으로 매핑.

(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction



## 데이터셋 구성 및 활용

- DeepFakeFace[6]
  - IMDB-WIKI를 기반으로 Stable Diffusion v1.5, Stable Diffusion Inpainting, InsightFace 등의 기법을 적용해 구축된 데이터셋
  - 90,000개의 딥페이크 이미지와 30,000개의 실제 이미지로 구성
  - 다양한 합성 기법을 반영해 폭넓은 딥페이크 사례 포괄
  - GPT-4o 모델을 활용하여, 딥페이크 및 실제 이미지에 대한 레이블링 및 키워드 추출
  - 이후 CLIP를 활용하여 512차원의 임베딩 형태로 DB에 저장
  - 10% 데이터는 테스트 데이터 셋으로 활용
- Seq-Deepfake[7]
  - 여러 단계의 연속적 얼굴 조작을 수행한 이미지
  - 보다 복합적인 위조 양상을 반영함으로써, 모델의 일반화 능력 검증 가능
  - 해당 데이터셋은 테스트 용으로만 사용





## 실험 모델 구성

### • 모델 목록

표 1. 실험 모델 구성 ↓

Table 1. Experimental model configuration ↓

Model ↓	Detail ↓
Gemini-1.5-flash ↓ (Zero-shot) ↓	Google [8] ↓
Gemini-1.5-flash ↓ (fine-tuned) ↓	2000 개의 <u>딥페이크</u> 이미지로 파인 튜닝 ↓
Llama-3.2-90B- Vision-Instruct ↓	Meta [9] ↓
Llama-3.2-11B- Vision-Instruct ↓	Meta ↓



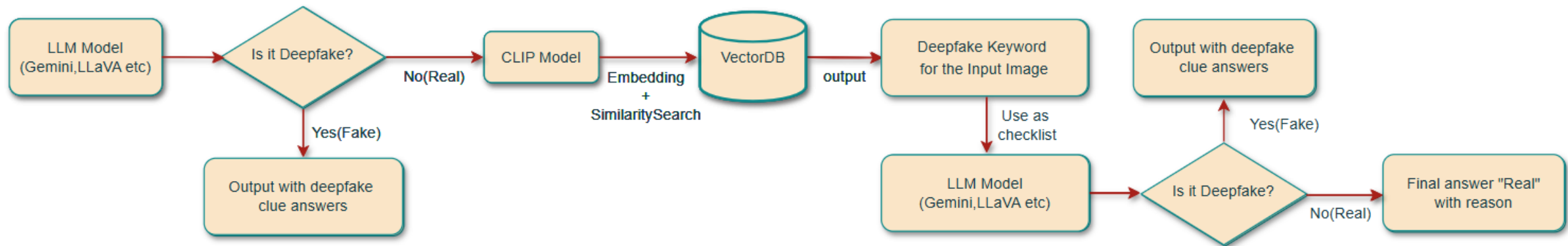
- 2024.05.14 출시
- 파라미터 약 32B로 추정

- 2024.09.25 출시
- 파라미터는 총 4개의 버전 (1B/3B/11B/90B)
- 멀티모달은 11B/90B에서만 지원





## 딥페이크 사례를 놓치는 현상을 최소화하기 위해, RAG를 통해 한번 더 검증하는 파이프라인 구성



- RAG 파이프라인은 다음과 같다. 모델로 하여금 1차 탐지를 수행하게 한 뒤, 실제 사진으로 분류한 경우 CLIP 모델을 활용한 임베딩 및 유사도 계산을 수행해 관련성이 높은 단서들을 매칭하게 된다.
- 이후 매칭된 단서들을 체크리스트로 활용하여 모델로 하여금 한번 더 탐지하여 최종 결과를 도출한다.



## CLIP의 이미지-텍스트 유사도 계산을 기반으로 한 RAG기법



유사 텍스트 내용 및 유사도:  
Rank 1: Similarity 0.2544  
Text: **\*\*Analysis of the Image for Deepfake Indicators\*\***

1. **\*\*Key Facial Features Analysis\*\***

- **\*\*Alignment, Proportions, and Symmetry\*\***: The eyes appear to be slightly misaligned with the rest of the face.
- **\*\*Pupil Reflection and Shape\*\***: Reflections in the eyes seem unnatural.
- **\*\*Unnatural Aspects\*\***: The jawline appears somewhat exaggerated.

2. **\*\*Skin Texture and Tone Analysis\*\***

- **\*\*Texture, Smoothness, and Tone Transitions\*\***: The skin appears slightly grainy.
- **\*\*Differences in Skin Color\*\***: There appears to be a mismatch in skin tone across different areas.

3. **\*\*Low-Level Detail Inspection\*\***

- **\*\*Edges and Borders\*\***: Some blending at the edges where the face meets the background.
- **\*\*Sharpness and Distortion\*\***: Facial features like eyes and mouth appear slightly distorted.
- **\*\*Reflections and Shadows\*\***: Shadows on the face do not fully match the lighting source.

4. **\*\*Lighting, Theme, and Environment Consistency\*\***

- **\*\*Lighting and Shadow Consistency\*\***: The lighting on the face is inconsistent with the background.
- **\*\*Theme Consistency\*\***: The facial expression and posture seem slightly off.

유사 텍스트 내용 및 유사도:  
Rank 1: 유사도 0.2544  
Text: 딥페이크 단서 분석

- 주요 얼굴 특징 분석
    - 정렬, 비율 및 대칭: **시선이 한 방향으로 뚜렷하게 집중되지 않는 등 눈의 정렬이 다소 어긋나** 보입니다. 얼굴의 나머지 부분에 비해 코와 입의 정렬 등 얼굴 특징에 **약간의 비대칭**이 있을 수 있습니다.
    - 동공 반사 및 모양: **눈동자의 반사가 일관성이 없어** 보이며 조작되었을 가능성이 있습니다.
    - 부자연스러운 측면: **턱선이 다소 과장**되어 보이고 나머지 얼굴 구조와 일치하지 않습니다. **입술 모양이 어색**해 보이며 **치아 표현이 불분명**하여 편집된 것으로 보입니다.
  - 피부 질감 및 톤 분석
- ...후략...

- Base64 형태로 변환되는 이미지를 텍스트와의 의미적 유사도를 계산하기 위해 CLIP 모델 활용
- 실제로 입력된 이미지와 연관성이 높은 텍스트를 매칭하는 것을 확인



## 효과적인 Context 사용을 위한 프롬프트 엔지니어링

표 3. 프롬프트에 따른 탐지 편향 (LLaMA-3.2-11B-Vision-Instruct-Turbo)

Prompt	Real	Fake
기본 프롬프트(이진 분류)	68.0%	32.0%
기본 프롬프트+판단 근거 요구+탐지 방법 제시	32.0%	68.0%

### Instructions:

1. For each keyword, carefully examine the image and objectively verify whether the described clue is present. Avoid making assumptions and clearly state if no issues are observed.
2. Maintain neutrality in your evaluation and do not assume the presence of manipulation unless there is strong and consistent evidence across multiple clues.
3. If anomalies are observed, provide a factual description without inferring intent or confirming manipulation prematurely.
4. If no clear anomalies are found, highlight natural variations or explain why observed features may align with expected real-world variations.
5. There is a high probability that the provided list of image URLs are REAL; only categorize them as FAKE if there is a high probability.

프롬프트 작성 예시(제약 조건, 데이터 특성 제시, 출력 형식 등)

- 프롬프트에 딥페이크 단서 및 방법을 제시할수록, 탐지 결과가 편향되는 현상이 발생하여 프롬프트 엔지니어링을 통해 이를 억제하였음.



## RAG 기법 적용 전후 성능 변화 관찰

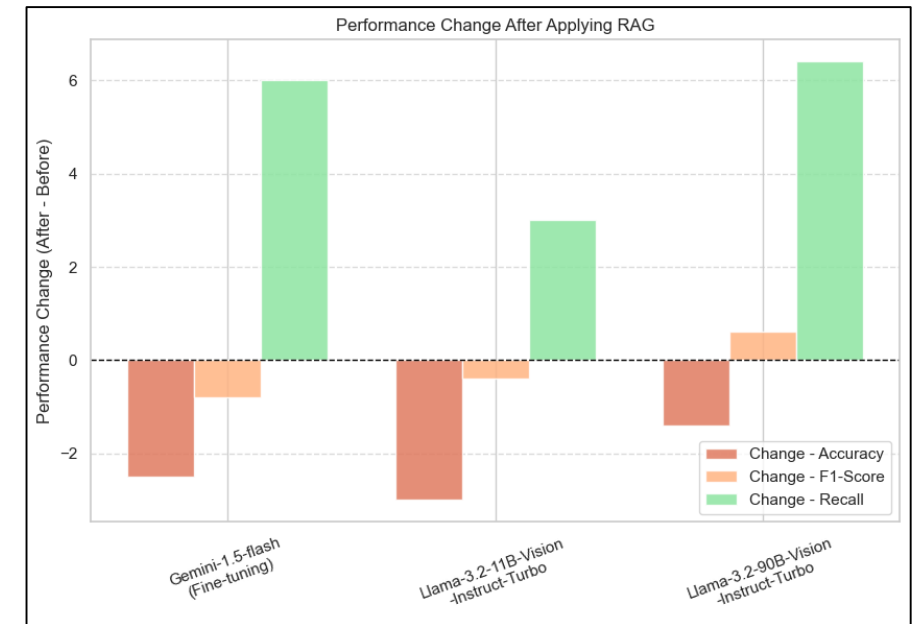
표 4. 기존 LMM 모델 답페이크 탐지 성능

Model	Accuracy	F1-score	Recall
Gemini-1.5-flash(zero-shot)	55.6	20.0	11.0
Gemini-1.5-flash(Fine-tuning)	82.5	81.4	77.0
Llama-3.2-11B-Vision-Instruct-Turbo	59.0	57.29	55.0
Llama-3.2-90B-Vision-Instruct-Turbo	74.7	70.9	71.4

표 5. RAG 적용 후 답페이크 탐지 성능

Model	Accuracy	F1-score	Recall
Gemini-1.5-flash(zero-shot)	X	X	X
Gemini-1.5-flash(Fine-tuning)	80.0	80.6	83.0
Llama-3.2-11B-Vision-Instruct-Turbo	56.0	56.9	58.0
Llama-3.2-90B-Vision-Instruct-Turbo	73.3	71.5	77.8

\* Gemini-1.5-flash(zero-shot) 모델의 경우 낮은 탐지 성능으로 fine-tuning하여 사용



RAG 적용 후 성능 지표 변화



## LMM 환경에서의 CLIP 유사도 기반 RAG 기법 적용을 통해, 딥페이크 탐지 모델의 성능을 향상시키는 방법을 제안

- 본 연구에서는 CLIP 유사도 기반의 RAG 기법을 통해, 다양한 모델의 정확도 소폭 감소를 감수하고 재현율을 증가시키는 방법을 제안하였음.
- 이는 재현율이 중요한 분야 ( ex : 딥페이크 탐지, 의료영상 분석, 이상 탐지 등)에서의 활용 가능성이 높음
- 추후 더욱 광범위한 테스트 환경(데이터셋, 모델의 종류, 분야 등)을 통해 연구를 발전 시킬 수 있을 것이며, 음성 및 영상 데이터로의 확장, 처리 속도 개선 등의 과제가 남아있음

- [1] S. Jia, R. Lyu, K. Zhao, Y. Chen, Z. Yan, Y. Ju, C. Hu, X. Li, B. Wu, 그리고 S. Lyu, “Can ChatGPT Detect DeepFakes? A Study of Using Multimodal Large Language Models for Media Forensics,” Proc. 2024 IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 4324–4333, 2024.
- [2] Y. Li, X. Liu, X. Wang, B. S. Lee, S. Wang, A. Rocha, 그리고 W. Lin, “FakeBench: Probing Explainable Fake Image Detection via Large Multimodal Models,” unpublished, 2024.
- [3] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, 그리고 D. Kiela, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.
- [4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, 그리고 I. Sutskever, “Learning transferable visual models from natural language supervision,” Proc. Int. Conf. Machine Learning (ICML), vol. 139, pp. 8748–8763, Jul. 2021.
- [5] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, 그리고 L. Schmidt, “OpenCLIP (Version v0.1),” Zenodo Software Repository, DOI: 10.5281/zenodo.5143773, 2021.
- [6] H. Song, S. Huang, Y. Dong, 그리고 W.-W. Tu, “Robustness and generalizability of Deepfake Detection: A study with Diffusion Models,” unpublished, 2023.
- [7] R. Shao, T. Wu, 그리고 Z. Liu, “Detecting and recovering sequential deepfake manipulation,” Proc. European Conf. Computer Vision (ECCV), Cham: Springer Nature Switzerland, pp. 712–728, Oct. 2022.
- [8] G. G. Team, “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,” 2024. [Online]. Available: <https://goo.gl/GeminiV1-5>
- [9] J. Chi, U. Karn, H. Zhan, E. Smith, J. Rando, Y. Zhang, 그리고 M. Pasupuleti, “Llama Guard 3 Vision: Safeguarding Human-AI Image Understanding Conversations,” arXiv preprint, arXiv:2411.10414, 2024.
- [10] <https://news.kbs.co.kr/news/pc/view/view.do?ncd=8048463>
- [11] <https://www.aitimes.com/news/articleView.html?idxno=167195>
- [12] D. Coçcomini, N. Messina, C. Gennaro, 그리고 F. Falchi, “Combining EfficientNet and Vision Transformers for Video Deepfake Detection,” unpublished, 2021.



# Thank You

eunoh.kim@ntoday.kr

changgeun.lee@ntoday.kr, \*1104py@kookmin.ac.kr

