연구실 발표 · 2026.04.28 · 30~40 MIN + Q&A

Modality Gap

Kanana 베타 preliminary + research proposal + 피드백 답변

SPEAKER

Yejin Yoon

DECK

v1.0

• 발표 구성

목차

연구 동기

컨텍스트

- 카카오 **Kanana-o** 베타테스터 선정 · 1일 `20` API call 한정
- 베타 컨텐츠로 **실험 사이클을 한 번 돌려본 것** · 결과보다 process 보고가 목적
- 베타 모델이므로 **결과 공개가 가능할지 미정**, 가능하면 워크샵 / 국내 short paper 시도
- 한국어 omni가 거의 없는 시점이라 직접 비교가 가능한 드문 시기

배경 연구 · 출발점

REFERENCE · CHOWERS ET AL. 2026

Is the Modality Gap a Bug or a Feature?

TL;DR. CLIP류 multi-modal 모델에서 image / text 임베딩이 공간상 분리돼 있는 **modality gap**은 robustness 관점에서 **bug**다, gap이 클수록 임베딩에 약한 perturbation을 줬을 때 모델 출력이 잘 흔들린다. 한 modality를 다른 modality 평균 쪽으로 옮기는 간단한 post-processing만으로 clean accuracy 손실 없이 robustness가 올라간다.

4/2 카카오 오프라인 밋업에서 개발자가 직접 modality gap 언급.

실험 설계

모델 매트릭스 · 한국어 ✓✗ × OMNI ✓

모델	크기	한국어	비고
Kanana-o	11.6B	KR	베타 · 1일 20 call
HyperCLOVA X	8B	KR	한국어 특화
Qwen2.5-Omni	7B	·	비특화 baseline
MiniCPM-o	8B	·	비특화 baseline

두 트랙

TRACK A · 지식

KMMLU × modality (input)

`text` · 텍스트 렌더 `PNG` · `TTS WAV` 동일 문항을 입력 모달만 바꿔 평가. 풀셋 n=1,100 (HCX/MiniCPM/Qwen) · Kanana는 30샘플 서브셋.

TRACK B · 감정 (설계 + 데이터 준비)

KoED 56샘플 × 4 variants (4종 입력 변형)

`text-bare` (텍스트만) · `image-bare` (텍스트 렌더 PNG) · `audio_neutral` (무도의 TTS) · `audio_emotion` (감정이 실린 원본 발화).

Track B는 다시 **감정 분류**와 **응답 생성** 두 sub-task로 나뉨.

데이터 준비 단계 · **본 배치는 발표 후** 진행 예정.

트랙 A · 지식 결과: 전 모델 공통 패턴

KMMLU 정답률 (%) · 풀셋 N=1,100 (HCX/MINICPM/QWEN) · 30샘플 서브셋 N=54-58 (KANANA)

모델	TEXT	IMAGE	AUDIO	불일치
HyperCLOVA X n=1,100	**49.4**	27.8	27.8	21.6
MiniCPM-o n=1,100	36.5	23.7	24.5	12.8
Qwen2.5-Omni n=1,100	33.0	29.1	27.5	3.9
HyperCLOVA X n=54	**51.9**	35.2	35.2	35.2
Kanana-o n=58	37.0	24.1	25.9	37.9
MiniCPM-o n=54	29.6	24.1	24.1	43.5
Qwen2.5-Omni n=54	22.2	27.8	24.1	38.9

불일치 = 같은 문항을 text / image / audio 세 modality로 풀었을 때, **한 modality라도 답이 달라진 경우**의 비율. 자료: log-12 · log-13.

관찰

- **한국어 특화(HCX) text 우위**가 가장 명확. 풀셋 49.4%, 다른 셋보다 13–16%p 높음
- 그러나 **image · audio로 가면 격차가 거의 사라진다**. HCX 풀셋 27.8/27.8
- **불일치율 12.8–43.5%**. 모달리티만 바꿔도 답이 갈리는 문항이 절반 가까이

발견 1 · 한국어 우위는 text에 갇혀 있다

가설 1 검증

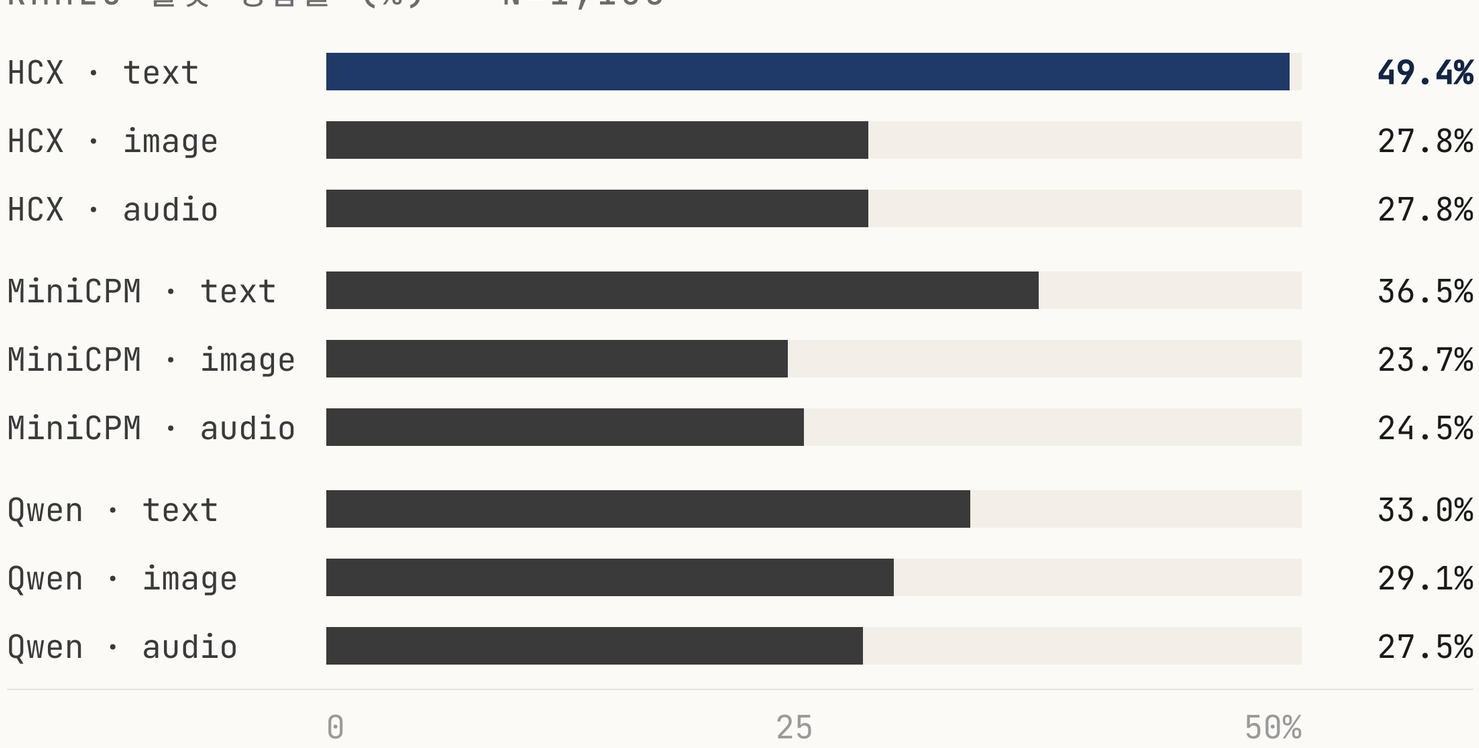
HCX text 우위 → image · audio에서 소멸

HCX는 text에서 49.4%로 다른 셋(33–37%)을 13–16%p 앞서지만, image · audio로 가면 27.8/27.8, 비특화 모델과 거의 같다.

"한국어 특화"라는 라벨이 modality 일반화에는 직접 전이되지 않는다는 첫 신호.

21.6%p

HCX TEXT–IMAGE/AUDIO 갭 · n=1,100



발견 2 · 정답률 평균은 비슷해도, 맞히는 문항은 다르다

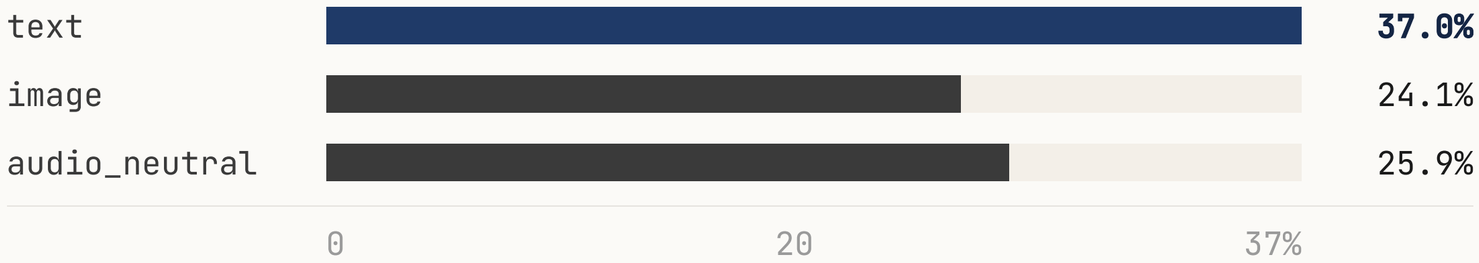
발견 2

같은 KMMLU 30문항을 3 variants로 풀게 함

`text · image · audio_neutral`, 입력 modality만 바꾼 3개 condition. modality별 정답률 평균은 text 37% · image 24% · audio 26%로 비슷하게 나오지만, 같은 문항이 modality마다 답이 달라지는 비율은 **37.9%**.

즉, 평균은 비슷해도 modality별로 맞히는 문항이 겹치지 않음, "평균 한 줄"로는 보이지 않던 변동.

KANANA-O · 30샘플 서브셋 정답률 (%)



"평균 한 줄"이 가린 변동을 진단할 수 있음 → Appendix · F3 답변의 핵심 도구.

37.9%

KANANA 모달리티별 답 불일치율 · n=58

트랙 B · 감정 설계 와 준비 현황

실험 설계

KOED 6축 · 56샘플 (5턴 이상 필터, SEED=42)

축	N	대표 라벨
기쁨	10	proud · joyful · trusting · hopeful · caring
슬픔	10	sad · lonely · sentimental · guilty · ashamed
분노	10	angry · furious · annoyed · disgusted · jealous
중립/불안	10	anticipating · apprehensive · afraid · prepared
정 (情)	8	한국 고유 · 따뜻하고 그리운 톤
한 (恨)	8	한국 고유 · 깊고 먹먹한 톤

정답·한국어 인식 여부를 별도 축으로 분리해 집계했다.

트랙 진행 현황 (2026-04-25 기준)

데이터 준비

KoED 재샘플링 + 톤-프롬프트 매핑 + TTS

Kanana TTS로 각 샘플당 2종(neutral / emotion-tone) 오디오 생성.
56샘플 × 2 = 112개 WAV 완료.

본 배치 · 진행 현황

4모델 × 4 variants × 56샘플 = 896 호출 완료

감정 분류는 Kanana-o 제외 3모델(HCX · Qwen · MiniCPM) 스코어 도출 완료.
응답 생성은 LaaJ judge 설계까지 마침, 채점 콜은 아직 미실행.

감정 트랙은 H3(음성 단서가 분류에 기여) 검증을 위해 audio_neutral과 audio_emotion을 같은 발화로 묶음 → 다른 변수 통제.

트랙 B · 감정 분류: 모델마다 다른 패턴

KOED 56샘플 분류 정확도 (%) · 4 VARIANTS

모델	TEXT	IMAGE	A_NEUTRAL	A_EMOTION
HyperCLOVA X 한국어 특화	**33.9**	17.9	17.9	17.9
Qwen2.5-Omni 비특화	**37.5**	30.4	32.1	33.9
MiniCPM-o 비특화	37.5	21.4	37.5	**46.4**
Kanana-o 한국어 특화 · 미진행	,	,	,	,

audio variants는 60초 제한 적용 · 응답 생성 task는 LaaJ 콜 미실행으로 본 발표에서 제외.

세 모델, 세 패턴

MINICPM

음성 단서가 분류에 +9%p

a_emotion 46.4% > text 37.5%. 슬픔 text 2/10 → a_emotion 6/10.

HCX

한국어 강점이 text에만 갇힘

image·audio는 56/56 모두 "중립"으로 디폴트-답 → modality 종속.

QWEN

평탄한 baseline

갭 4%p, 운율 효과 미세. 단 a_emotion 중립 0/10, trade-off 존재.

한계와 다음 단계

한계

- **표본 규모**: 지식 1,100 확보했으나 Kanana 비교는 30샘플 서브셋 (n=54–58), KoED는 단 56샘플로 분류만 채점
- **Kanana-o 미진행**: 베타 1일 20 call 제약 + 4 variants × 56 = 224 call 필요 → 한국어 특화 모델 한 축이 비어 있음
- **Audio variants 60초 제한**: 일부 발화 잘림. 중간 운율이 손실된 채로 채점된 케이스 존재 가능성
- **응답 생성 task 미보고**: LaaJ judge 설계 완료, 채점 콜 미실행. 본 발표는 분류 정확도만 다룸
- image·audio는 평균 한 줄 보고, 문항별 변동 진단 도구 미흡

PART 2로의 TRANSITION

PRELIMINARY → PROPOSAL

"4 variants 설계가 이미 modality × 음성 단서 요인을 분해하고 있었다"

한국어 우위가 text 안에만 갇히는 HCX, 음성 단서로 +9%p 회수하는 MiniCPM, 같은 문항이 모달리티마다 38% 비율로 답이 갈리는 현상, 셋 다 modality gap이 task 정보 구조에 따라 **방향이 뒤집히는** 현상이라는 새 연구의 첫 신호.

이를 제대로 측정하려면 input × output × task type 통합 프레임워크가 필요 → **REMODE**.

핵심 주장 · gap의 방향은 task가 결정한다

CORE CLAIM

Modality gap은 bug가 아니라 체계적 현상

"task가 요구하는 정보의 modality-specificity에 따라 gap의 방향과 크기가 결정된다."

- **Reversal** · modality-neutral 과제와 음성 단서 의존 과제에서 gap 방향이 반대로 갈 수 있음
- **Routing** · 그렇다면 task별 modality 선택으로 성능 회수 가능

Part 1과 인과관계로 묶지 않음. 본 주장은 자체 출발 가능.

한 줄 요약

CORE CLAIM

Modality gap은 task 정보 구조에 따라 방향과 크기가 달라지는 체계적 현상

자료: plan §핵심 주장

선행연구 공백

#	구분	현재 상태
1	**Output modality 변수**	모든 선행연구가 output을 text로 고정
2	**Input × Output × Task type 통합**	각 축은 별개 연구로만 다뤄짐, 통합 평가 부재
3	**과제 특성별 gap 역전**	풍자 audio 우위 등 개별 보고만 있음, 같은 모델 셋으로 직접 비교 없음
4	**Mismatch cost 정량화**	"있다"는 정성 보고만, 비용을 수치로 잡은 연구 없음
5	**Audio LLM이 청취 단서를 쓰는 조건**	LISTEN 등은 "못 쓴다"는 실패 진단까지, 어떤 조건에서 쓰게 되는지는 미답

한국어 axis는 본 표에서 제외. 메인 contribution은 영어 도메인, 한국어는 cross-lingual 보조축. 자료: plan §선행연구 공백.

선행연구

묶음 A · INPUT MODALITY GAP 측정

논문	연도	한 줄 요약	커버	미커버
OmnixR	ICLR 2025	실제·합성 omni 입력으로 LMM 추론 일관성 평가, gap 존재 입증	#3 부분	#1·#2·#4·#5
XModBench	arXiv 2025.10	input modality 5종 × 과제 그리드, modality별 정확도 비대칭 측정	#2 부분	#1·#3·#4·#5
REST	arXiv 2025.12	같은 의미를 modality만 바꿔 반복 질의, 응답 일관성으로 gap 정량화	#3	#1·#2·#4·#5
CMC	arXiv 2024.11	Cross-Modal Consistency · 동일 질문 modality 변환 시 응답 변화율 측정	#3	#1·#2·#4·#5
Beyond Text-Dominance	arXiv 2026.04	text 우위 가정 비판, 과제별로 modality 선호가 갈리는 사례 보고	#3	#1·#2·#4·#5

*묶음 A는 **input modality gap의 존재·일관성**까지 다룸. output 변수(공백 #1)와 mismatch cost(공백 #4)는 미답.*

선행연구

묶음 B · SPEECH-TEXT 내부 + AUDIO-SPECIFIC

논문	연도	한 줄 요약	커버	미커버
Alignment Path	EMNLP 2025	speech-text 내부 정렬 경로 분석, audio encoder 표현이 어디서 텍스트와 정합되는지 추적	#5 부분	#1·#2·#3·#4
Anatomy	arXiv 2026.03	audio LM 내부 표현 해부, 음성 단서가 어떤 layer에서 보존/손실되는지 분석	#5 부분	#1·#2·#3·#4
TARS	arXiv 2026.01	speech-text 정렬 학습 기법, audio-only 단서를 위한 표현 강화	#5	#1·#2·#3·#4
LISTEN	EACL 2026	현 omni 모델의 audio 청취 실패 패턴 진단, 음성 단서 무시 사례 정리	#5 진단	#1·#2·#3·#4
Sarcasm	arXiv 2025.09	비꼬임 인식에서 음성 단서가 결정적인 사례, modality 종속 과제 예시	#3 사례	#1·#2·#4·#5

5개 공백 중 단일 논문이 3개 이상 메운 사례 부재. *본 연구는 5개 모두 다룸.*

연구 질문 · RQ1~4

RQ 1

정보 대칭 과제에서 gap 존재?

input × output 조합별 정확도 차이를 통제된 비교로 측정.

RQ 2

I-O modality 일치가 유리한가

audio in → speech out이 mismatch 조합보다 안정적인가. open-source 모델 CKA로 내부 처리 효율 직접 검증 가능.

RQ 3

모달 고유 정보 과제에서 gap 패턴 역전?

감정·풍자 등 음성 단서 의존 과제에서 modality-neutral과 정반대 우열이 나타나는가.

RQ 4

Mismatch cost와 routing 회복

비최적 조합 사용 시 비용은 얼마이며, task-aware routing이 그 비용을 얼마나 회복하는가.

RQ1·2가 gap 존재 + 내부 메커니즘, RQ3가 reversal, RQ4가 routing 가치.

가설 · H1과 H3의 정반대 예측

H1 · 정보 대칭 과제

text → text가 가장 잘 맞는다

주장, 정답에 추가 단서(음·이미지)가 필요 없는 지식형 QA는, text 입력·text 출력 조합이 가장 정확.

근거, 텍스트는 모델이 가장 많이 학습한 통로이자, 다른 모달리티는 결국 텍스트 표현으로 정렬해서 풀어야 함.

측정, KMMLU 한·영을 4 variants × 모델로 돌려 정확도를 직접 비교.

prelim → 본 측정 · Part 1은 한국어 (KMMLU·HRM8K)로만 봤음. 이미 text 우위 확인됨 → 본 연구는 영어 지식셋(MMLU·MMLU-Pro 등)으로 확장해 modality-neutral 가정의 일반성을 본다.

H2 · 내부 처리 효율

입력·출력이 같은 모달일 때 내부가 깔끔

주장, input과 output이 같은 modality(예: text→text, audio→audio)일 때, 내부 표현이 더 정렬되어 있고 추가 변환 비용이 적음.

근거, 모달이 다르면 모델이 내부에서 한 번 더 변환·정렬하는 cost를 지불, 그 흔적이 layer 표현에 남을 것.

측정, open-source 모델(Qwen·MiniCPM)의 layer-wise CKA로 일치 / 불일치 조합 간 표현 유사도를 비교.

prelim → 본 측정 · Part 1은 입력 정확도까지만 보고 내부 표현은 못 봤음. 본 연구에서 처음으로 layer 단위 처리 cost 자체를 측정해, 입력 정확도 뒤의 메커니즘을 검증.

H3 · 음성 단서 의존 과제

audio in이 text in을 이긴다

주장, 감정·풍자처럼 **운율(말투·억양)이 의미를 좌우**하는 과제에서는, audio 입력이 text 입력 (=ASR transcribe된 글)보다 정확.

근거, text로 변환하는 순간 화자의 톤·길이·강세가 사라짐. 이 손실은 단순 정성 보고로만 있었음.

측정, KoED(한국어) + 영어 감정·풍자 셋 (MELD·MUStARD 등)에서 a_emotion vs text 정확도 차로 손실량을 직접 정량화.

prelim → 본 측정 · Part 1 KoED 56샘플에서 MiniCPM이 a_emotion +9%p 회수를 보여줬음 → 첫 양성 신호. 본 연구는 영어로 확장하고 표본을 키워 reversal이 언어와 무관한 현상인지 확인.

실험 매트릭스

INPUT × OUTPUT 그리드

OUT \ IN	TEXT IN	IMAGE IN	AUDIO IN
text out	①	②	③
speech out	④	⑤	**⑥ (I-O match)**

⑥은 RQ2의 I-O 일치 후보. 매 task × 매 모델마다 6칸 모두 측정.

도메인

- **영어 메인**, MMLU-Pro · IEMOCAP · MELD
- **한국어 cross-lingual 보조**, 한 줄로만 표기 · 메인 contribution 아님

IMAGE OUTPUT 제외 · 세 이유

- **형식** · 지식 정답·감정 라벨을 이미지로 출력하라는 task 자체가 부자연스러움
- **평가** · 이미지-텍스트 매칭은 평가 모델 bias confound가 커서 fair comparison 어려움
- **시나리오** · QA·SER 사용 맥락에서 이미지 출력의 실용 가치 약함

Any-to-any 모델(Ming-flash-omni 2.0 / CoDi-2 / AnyGPT / NExT-GPT)은 비교 모델 풀 후보로 검토하되, 출력 변수로는 비채택. 자세한 답변은 Appendix · F1.

과제 분류 · 라벨에서 요인 점수로

원안 · 과제를 세 묶음으로 라벨링

각 라벨 = "이 묶음에선 이 modality가 유리할 것"이라는 사전 가정.

라벨	예시 데이터셋	의미
modality-neutral	MMLU-Pro 지식 QA	입력 modality와 무관하게 정답 동일
audio-advantage	IEMOCAP / MELD 감정	말투·억양이 정답을 좌우
image-advantage	차트 QA 류 (보류)	그림이 정답을 좌우 · 본 발표 미포함

한계 · 결과를 가정한 분류

결론을 라벨에 미리 박아두는 셈

"audio가 유리할 것이라 미리 분류한 묶음에서 audio가 이겼다"는 결론은 라벨 정의의 반복일 뿐. 측정된 gap이 modality 차이인지, 분류 기준에서 나온 결과인지 구분 불가.

대응 · 요인 점수

라벨을 떼고 좌표로 다시 본다

각 task에 **지각 단서 · 운율 의존 · 기호 추론 · 외부 지식** 등 요인별 점수를 부여. 같은 데이터셋도 "운율 0.8 / 기호 0.2"처럼 좌표로 표현.

→ modality gap을 요인의 함수로 회귀해, gap이 어떤 요인과 정렬되는지 직접 확인. **요인 후보·점수 기준은 미정, Appendix · F3.**

데이터셋

데이터셋

- **정보 대칭**, MMLU-Pro (영어, 12K, 10지) · OmnixR과 동일 소스 · 500~1000 샘플 사용
- **Audio-advantage**, IEMOCAP / MELD (영어)
- **한국어 보조**, KMMLU · KEMDy20 / K-EmoCon · cross-lingual 분석에만 활용

MODALITY 변환

변환	도구	측정 포인트
text → image	Pillow 렌더 PNG	OCR 한계를 통제하기 위해 동일 폰트·크기로 렌더링, modality 차이만 격리
text → audio	OpenAI TTS / `say -v Yuna`	운율 평탄(중립 TTS) vs 감정 합성을 분리, 말투 단서가 정답에 미치는 영향 측정
audio → text	ASR (Whisper 등)	**transcribe 손실량 직접 측정**, audio→text 우회 경로의 정보 손실을 정량화

비교 모델

모델	분류	출력 MODALITY	비고
GPT-4o	메인	text · speech	API 안정
Gemini 2.5 Pro	메인	text · speech	API 안정
Qwen2.5-Omni	메인	text · speech	open-source · CKA 가능
Qwen3-Omni	메인	text · speech	open-source
Qwen3.5-Omni	메인	text · speech	open-source
Kanana-o	preliminary	text · speech	API 제약(20/일) + 한국어 특화 → 메인 제외, Part 1 자료로
HCX-SEED-Omni	drop 후보	text · speech	한국어 특화, 영어 성능 저하 우려
추가 분석 옵션 · Ming-flash-omni 2.0 (MoE) / NExT-GPT (diffusion hybrid) / AnyGPT (discrete token) → 아키텍처 유형이 gap 패턴에 영향?			

측정 지표

지표

- **정확도**, modality 조건별 정답률(QA) · F1(SER)
- **일관성**, cross-modal disagreement rate · Kendall's tau
- **Modality mismatch cost**, 최적 대비 비최적 사용 시 하락폭
- **Task-aware routing gain**, informed vs naive routing 차이

OPEN ISSUE

SPEECH OUTPUT 평가

ASR transcribe 후 비교 vs 직접 semantic similarity

fair comparison 방법론은 plan 단계에서 미해결. **open issue로 정직하게 공시**, Q&A 시 토론 가능.

예상 기여

CONTRIBUTION 1

Input × Output × Task type 통합 평가 프레임워크

Output modality 변수화가 unique. 선행연구는 모두 text 고정.

CONTRIBUTION 2

Modality gap reversal 실증

modality-neutral text 우위 → 음성 단서 의존 audio 우위 패턴을 같은 모델 셋에서 직접 측정.

CONTRIBUTION 3

Modality mismatch cost 정량화

"있다"는 정성 보고를 비용 수치로.

CONTRIBUTION 4

Task-aware modality routing 효과 실증

routing이 mismatch cost를 얼마나 회복하는가, 실증.

*XModBench·CMC와의 차별점은 **output 변수 + task type 분화 + mismatch cost + routing** 모두 본 연구만. 자세한 답변은 Appendix · F2. 한국어는 cross-lingual 보조 분석.*

논문 구조와 venue

논문 구조 · 7 섹션

- **1.** Introduction · 핵심 주장과 reversal 동기
- **2.** Related Work · 5개 공백 매핑
- **3.** Framework · input × output × task type 정의
- **4.** Experimental Setup · 5모델 × 6셀 × 3 task
- **5.** Results & Analysis · RQ1~4
- **6.** Discussion · routing 함의
- **7.** Conclusion

VENUE 우선순위

슬롯	근거
ACL evaluation (1순위)	OmnixR ICLR · LISTEN EACL · REST arXiv 후속 · long paper 분량
Interspeech / ICASSP	audio 축 강조 시 가능, 우선순위 차순위

자료: plan §예상 논문 구조 · §venue.

정리

오늘의 핵심 메시지

- **Part 1** · Kanana-o 베타 preliminary에서 같은 task가 모달리티마다 정답을 다르게 만든다는 첫 신호를 확인했다
- **Part 2** · 이 신호를 input × output × task 세 축의 reversal 현상으로 정식화하는 research proposal, REMODE
- 본 발표는 Part 2의 plan을 함께 들여다보고 다듬기 위한 자리

남은 질문

- 요인 점수 표의 차원 후보를 어디까지 늘리는 것이 적절한가
- 비교 모델 풀에서 한국어 특화 계열을 어느 비중까지 유지할 것인가
- mismatch cost를 단일 지표로 합칠지, 차원별로 따로 보고할지

사전 받은 피드백 4건에 대한 답변은 **Appendix**에서 다룬다. Q&A 중 필요 시 직접 이동.

END OF MAIN TALK · Q&A

2026.04.28

받은 피드백 4건

#	피드백 요지	답변 전략	상태
F1	any-to-any 모델까지 범위에 넣어야 하는가	이미지 output 제외 justify · 현 omni 동일 파이프라인 미지원	정리됨
F2	XModBench / Cross-Modal Consistency와 어떻게 다른가	output 변수 + 과제별 역전 + mismatch cost가 unique	정리됨
F3	**결과를 가정한 분류 · 평가 task 자체가 모달리티에 종속**	Kanana 4 variants가 **이미 차원 분해의 부분 구현** · 라벨 분류 → 차원 점수 표로 갈아엎기	핵심 · 2장
F4	역방향 접근 (output → input 추론)	본 발표 범위 밖, 추후 별도로 정리할 방향성으로 소개	outline 단계

F1·F2·F3는 본 발표를 거치며 plan에 직접 반영했고, F4는 다음 작업의 방향성으로만 두고 본 발표 범위에서는 뺐 두었다.

F1 · image output 제외 근거

답변

- 비교 모델 5종(GPT-4o · Gemini 2.5 · Qwen2.5/3/3.5-Omni) 중 **동일 파이프라인 안에서 이미지 output을 안정적으로 지원하는 모델 부재**
- any-to-any를 별도 모델군으로 둘 경우, 본 연구의 통제 변수(input × output × task) 붕괴. 비교 가능성이 우선
- 이미지 output은 **future work에 명시**, 동일 파이프라인이 안정되면 그대로 확장 가능한 frame 유지

현재 SCOPE

OUTPUT MODALITIES

text · audio

5모델 모두 안정 지원. output 변수로 다루기에 충분한 통제 가능.

자료: /posts/remode-feedback-1/

F2 · XModBench와의 차별점

UNIQUE 1

output modality를 변수로

XModBench 등은 input modality 변경에 집중. 본 연구는 **output까지 직접 변수화**해 input × output 그리드 측정.

UNIQUE 2

과제 특성별 역전

모달리티 우열이 **과제에 따라 뒤집히는** 패턴(말투 단서가 중요한 과제 vs. 텍스트 비중 큰 과제 등)을 직접 측정 대상으로 설정.

UNIQUE 3

mismatch cost

입력과 출력 modality가 어긋날 때 발생하는 **비용**을 정량화. 단순 정답률을 넘어 "어떻게 틀리는가"까지 측정.

XModBench·Cross-Modal Consistency 영역은 본 연구의 부분집합으로 위치시키고, 위 세 축에서만 차별 주장. 자료: `/posts/remode-feedback-2/`

F3 · 결과를 가정한 분류, 비판과 진단

CRITIQUE

비판 요지

"modality A에서 정답률 70%, B에서 50%"라고 비교할 때, task 자체가 한 modality에 편향되어 있으면 결과는 **모델 능력이 아니라 task 정의를 측정**하고 있는 것. 분류 기준이 결과를 미리 가정하는 셈.

DIAGNOSIS

진단 · 4 variants는 이미 요인 분해

Kanana 실험의 `text-bare / image-bare / audio_neutral / audio_emotion`은 단순 "모달리티 4종"이 아니라 **(기호적 vs. 지각적) × (말투 단서 유무)**의 차원 분해.

비판은 옳음. 동시에 우리 설계는 이미 그 비판이 요구하는 방향으로 부분 이동 중.

VARIANT 1
text-bare
기호적 · 말투 없음

VARIANT 2
image-bare
지각적 · 말투 없음

VARIANT 3
audio_neutral
지각적 · 말투 평탄

VARIANT 4
audio_emotion
지각적 · 말투 살아있음

F3 · 비판을 어떻게 plan 문서에 반영했는가

검토한 개선안 4가지

"라벨 분류" 비판에 대응하기 위해 고려한 선택지와 그 결과.

- **① 라벨은 유지하고 주석만 추가**, 가장 가볍지만 비판이 그대로 남음 · 기각
- **② 모달리티마다 별도 task 정의**, 모달리티 간 비교 자체가 불가능해짐 · 기각
- **③ 요인 점수 표 도입**, task를 라벨 대신 (기호적·지각적·말투 의존도 …) 요인 좌표로 표현 · **채택**
- **④ task 비교를 빼고 모달 일관성만 측정**, 본 연구의 contribution이 약화됨 · 기각

③번 채택 → PLAN 문서 개정

개선안 ③을 plan 문서의 세 섹션에 어떻게 반영했는가.

PLAN 섹션	원안	개정 후
§과제 유형	라벨 묶음 (modality-neutral / audio-advantage 등)	**요인 점수 표, task별 좌표**
§risk	일반 risk 목록	**"결과를 가정한 분류"를 risk #1로 명시 · 완화책으로 요인표 제시**
§평가 지표	정답률 단일 지표	**요인별 점수 + mismatch cost 함께 보고**

근거 자료: `/posts/remode-feedback-3/`

F4 · 역방향 접근, 다음 단계의 방향성

F4 · 역방향 접근

- 입력 modality → 출력 정확도라는 일방향 외에, **출력 양상에서 입력 modality를 추론**하는 역방향 가능성
- 본 연구의 mismatch cost 정의와 자연스럽게 결합되며, 같은 데이터에서 부산물로 분석 가능
- 본 발표 범위에서는 방향성만 공유하고, 본 연구의 결과 위에서 추가 분석으로 이어 갈 예정

다음 단계

- **Part 2 plan**을 본 발표에서 받은 의견을 반영해 구체화한다
- **Kanana classify 본 배치**는 발표 후 진행, 본 연구의 결정타로 쓰지 않으므로 일정에 여유를 둔다
- 요인 점수 표의 차원 후보를 plan 본문 수준으로 lock-in한다

END · Q&A

2026.04.28