서울대학교 SEOUL NATIONAL UNIVERSITY

보도자료

배포 일자
2026.10.28.(수)

보도 희망 일시	☑ 배포 이후(엠바고 없음)		☐ 별도 기재	
담당/문의	기관명	AI연구원		
	연락처	02-880-4111	이메일	hyunoh@snu.ac.kr

서울대-포스텍-노타 공동연구팀, 10년을 기억하는 차세대 옴니모달 AI 에이전트 개발 나선다

- 경량·저전력AI한계극복기술개발사업 수행

- ICML·EMNLP·ECCV 등 국제 최우수학회 논문 9편 성과

서울대학교는 포항공과대학교(포스텍), AI 경량화 전문기업 (주)노타가 공동연구팀을 구성해 10년 이상 사용자를 기억하고 보조하는 차세대 옴니모달 AI 에이전트 기반 기술 개발에 나선다. 이번 연구는 과학기술정보통신부와 정보통신기획평가원(IITP)의 경량·저전력AI한계극복기술개발사업으로 추진되며, 과제명은 '차세대 옴니모달 에이전트를 위한 초장기 계층형 기억·추론 아키텍처'다. 총 연구개발비는 약 96억 원, 연구 기간은 2026년부터 2030년까지 5년이다. 연구팀은 지난 8월 19일 서울대 해동첨단공학관 3층에서 '초장기 계층형 기억·추론 AI연구센터' 현판식과 간담회를 열고 본격적인 연구 개시를 알렸다.



▲ 다양한 모달리티를 실시간으로 이해하고 10년 이상 사용자를 보조하는 초장기 옴니모달 에이전트

연구팀은 비디오·오디오·텍스트·심박수 등 다양한 모달리티를 실시간으로 이해하고, 축적되는 사용자 경험을 바탕으로 장기간 사용자를 보조하는 에이전트를 목표로 한다. 연산·메모리·전력이 제한된 웨어러블 기기의 한계를 넘기 위해 ▲뇌의 계층적 기억 체계를 모사한 초장기 기억·추론 기술 ▲상태공간모델(SSM)과 트랜스포머를 결합한 차세대 모델 아키텍처 ▲뇌 신호·심박수까지 아우르는 옴니모달 에이전트 응용의 세 축으로 연구를 수행한다.

과제는 서울대가 주관하고 포스텍과 노타가 공동으로 참여하며, AI반도체 기업 (주)퓨리오사에

이아이가 협력기관으로 함께한다. 연구책임자인 송현오 교수(서울대 컴퓨터공학부)를 비롯해 서울대 이재욱·김건희·심재웅 교수와 포스텍 모상우 교수 등 참여 교수 5명과 연구원 41명이 알고리즘부터 하드웨어·시스템까지 아우르는 풀스택 연구팀을 구성했다. 서울대 AI연구원이 위치한 해동첨단공학관에는 공동연구센터(SNU-POSTECH-Nota-Furiosa Joint AI Lab)를 설치해 참여기관 연구원이 상주하며 연구한다.

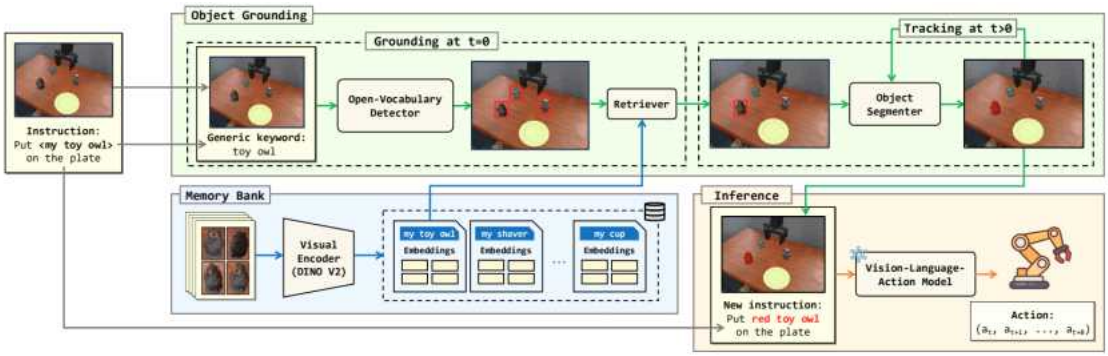
현판식에는 이재욱 AI연구원장을 비롯한 연구팀이 참석해 센터의 출범을 축하했다. 이어 열린 간담회에서는 연구 개발 추진 계획을 중심으로 향후 연구 계획을 논의했다.

상용화는 노타가 맡는다. 모델 경량화·추론 효율화 기술은 노타의 AI 모델 최적화·배포 플랫폼 NetsPresso에 기능 모듈로 탑재하고, 옴니모달 기술은 엣지 기기용 비전-언어 AI 솔루션 NVA(Nota Vision Agent)의 멀티모달 기능을 고도화하는 데 활용할 계획이다. 또한 협력기관인 퓨리오사에이아이의 AI반도체 RNGD에서 개발 기술을 구동·검증하고, 성능이 입증되면 퓨리오사 NPU SDK 통합을 검토해 국산 NPU 기반 상용 레퍼런스를 확보할 계획이다.

공동연구팀은 과제 착수와 함께 국제 최우수학회 논문 9편을 발표했다. 이 가운데 과제 목표와 가장 맞닿은 대표 성과로는 개인화 로봇 에이전트 기술 'VAP'와 대형언어모델 초경량 압축 기술 'Q-Strata'가 꼽힌다.

■ 대표 성과 ① 개인화 VLA 에이전트 기술 'VAP'

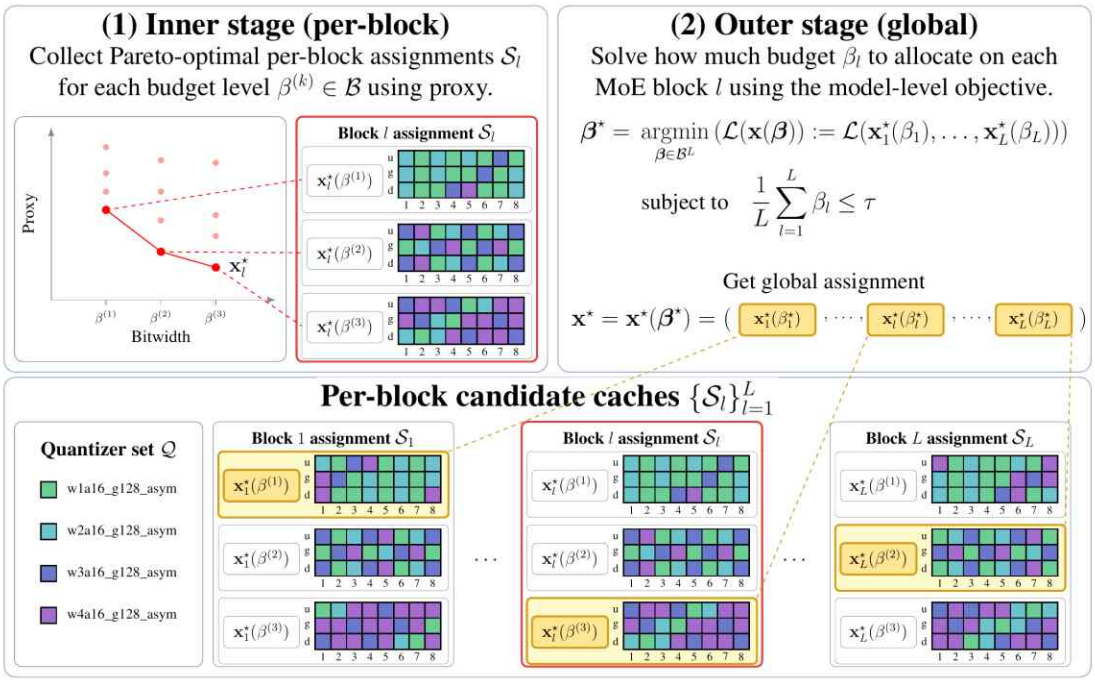
포스텍 모상우 교수 연구팀은 로봇이 "내 컵을 가져와" 같은 개인화된 명령을 수행하도록 돕는 'VAP(Visual Attentive Prompting)' 기술을 개발했다. 사용자가 제공한 몇 장의 이미지를 시각 메모리로 삼아 목표 물건을 찾고 강조 표시함으로써, 시각-언어-행동(VLA) 모델이 추가 학습 없이도 비슷한 물건들 가운데 사용자의 물건을 정확히 집어낼 수 있게 했다(ICML 2026).



▲ 소수의 기준 이미지로 시각 메모리를 구축해 목표 대상을 찾고 강조 표시하는 VAP 프레임워크

■ 대표 성과 ② MoE 대형언어모델 초경량 압축 기술 'Q-Strata'

서울대 송현오 교수 연구팀은 전문가 혼합(MoE) 구조의 대형언어모델을 극저비트로 압축하는 혼합 정밀도 양자화 기술 'Q-Strata'를 개발했다. MoE 모델은 성능이 뛰어나지만 수많은 전문가 모듈을 모두 메모리에 올려야 해 저자원 기기에 탑재하기 어렵다. Q-Strata는 블록 내부의 층별 비트 배분과 블록 간 비트 배분을 두 단계로 나눠 최적화해, 같은 메모리 예산에서 성능 손실을 최소화한다. 그 결과 Mixtral-8×7B 모델을 평균 1.75비트로 압축했을 때 기존 최고 기법 대비 언어모델 성능 지표(퍼플렉서티)의 오차를 절반 이하로 줄였고, DeepSeek-V2-Lite 모델의 크기를 31.4GB에서 5.0GB로 약 6분의 1 수준까지 줄였다. 이는 제한된 메모리와 전력으로 대형 모델을 구동해야 하는 온디바이스 에이전트의 핵심 기반 기술이다(EMNLP 2026).



▲ 블록 내부 단계(1)에서 비트 예산별 최적 배분 후보를 구한 뒤, 블록 간 단계(2)에서 모델 전체 성능을 기준으로 블록마다 예산을 배분하는 Q-Strata의 2단계 비트 배분 구조

송현오 교수는 "이번 과제는 제한된 자원에서도 사람처럼 오래 기억하고 추론하는 인공지능의 기반 기술을 확보하려는 도전"이라며 "산학연이 한 공간에서 협력하는 공동연구센터를 중심으로 세계 최고 수준의 초장기 옴니모달 에이전트 원천기술을 개발해 나가겠다"고 밝혔다.