고종환 교수

성균관대학교 전자전기공학부

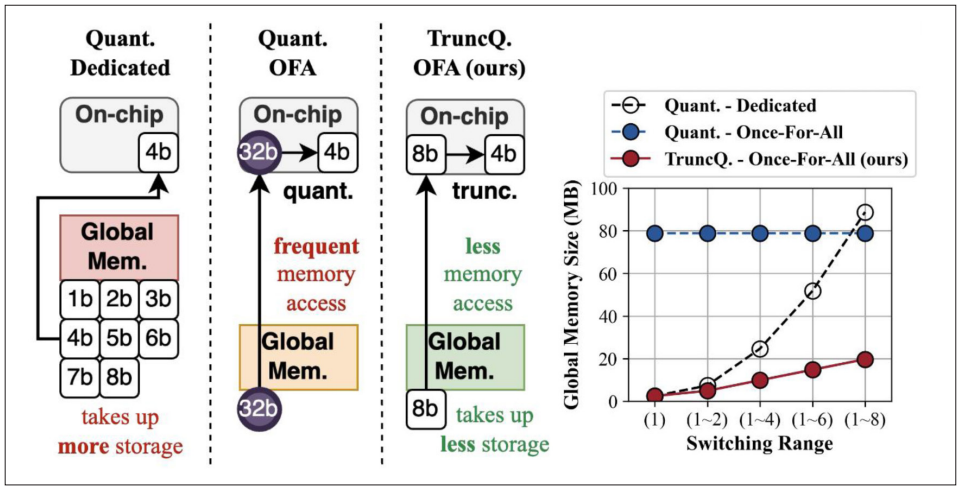
I. 저자 소개

고종환 교수는 현재 성균관대학교 전자전기공학부 교수로 재직 중이며, 인공지능 모델과 HW 시스템을 통합적으로 연구하는 IRIS(Intelligent and Resource-efficient Image/video processing Systems) 연구실을 운영하고 있다. 연구실은 인공지능 알고리즘 및 소프트웨어 설계뿐 아니라, 이를 실제 디바이스에서 효율적으로 구현하기 위한 하드웨어 및 시스템 구조까지 함께 고려하는 SW-HW 통합(co-design) 연구를 수행하고 있으며, 이를 위해 모델 설계, 데이터 표현, 그리고 IMC/PIM 및 in-sensor 기반 하드웨어 아키텍처를 아우르는 end-to-end 연구를 진행하고 있다. 특히 최근에는 LLM, Diffusion 등 멀티모달 생성 모델에 대하여 모델의 효율성과 안전성을 동시에 확보하면서 실제 환경에서 동작 가능한 AI 시스템 구현에 중점을 두고 있다. 연구 성과는 NeurIPS, ICML, CVPR, ICCV 등 인공지능 분야뿐만 아니라 DAC, DATE, ISCA, VLSI 등 시스템 및 HW 분야 주요 국제 학술대회 및 저널에서 발표되고 있으며, 산업체와의 협력을 통한 기술 이전 및 특허 확보에도 적극적으로 참여하고 있다.

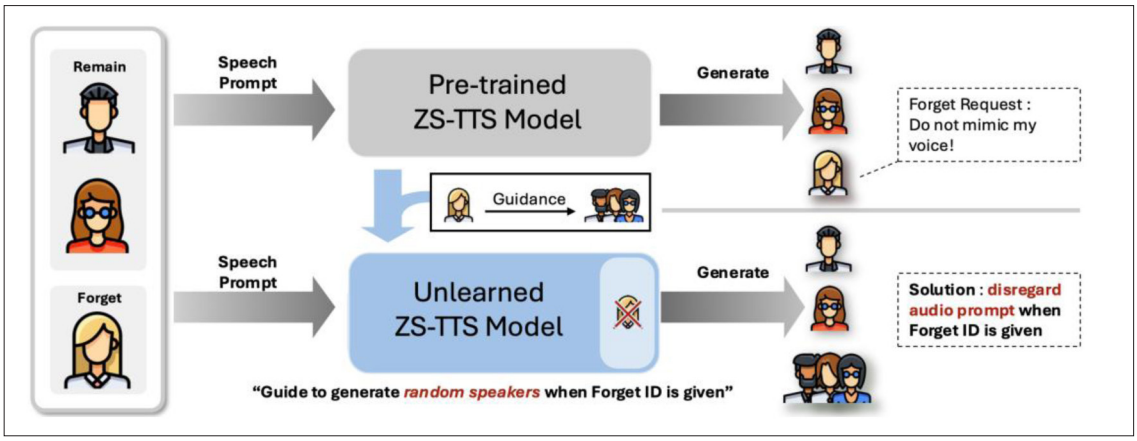
II. 연구 분야

1. 생성 모델 경량화 및 unlearning 연구

최근 대규모 딥러닝 모델, 특히 생성 모델(LLM, Diffusion 모델 등)은 다양하고 복잡한 생성 태스크에 대해 높은 성능을 달성하고 있으나, 막대한 연산량과 메모리 요구량으로 인해 실제 환경에서의 활용에는 여전히 제약이 존재한



<그림 1> 추론 시 적응적으로 정밀도를 변경할 수 있는 TruncQuant 기법 [ISLPED '25]



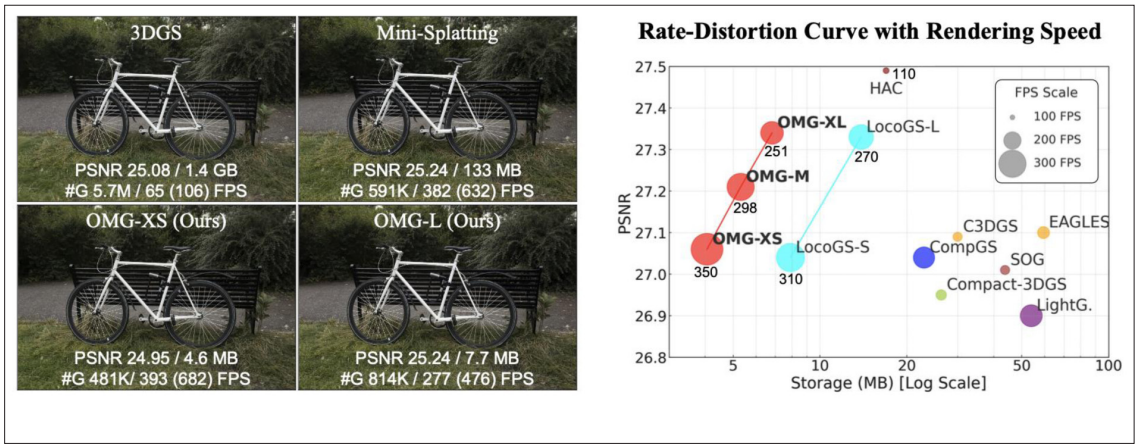
<그림 2> Speech-to-text 모델에서의 Identity unlearning 연구 [ICML '25]

다. 이에 연구실에서는 모델의 경량화 및 가속을 위한 다양한 접근과 더불어, 입력 및 환경 변화에 따라 동적으로 경량화를 수행하는 multi/mixed precision 양자화, dynamic sparsity, test-time adaptation 등의 기법을 연구하고 있다.

또한 생성 모델의 확산과 함께 개인정보 침해, 유해 영상 생성 등 안전성 문제 역시 중요한 이슈로 부각되고 있어, 생성 모델에서 특정 정보나 개념을 제거하기 위한 machine unlearning 및 concept erasing 기술을 기반으로 신뢰할 수 있고 제어 가능한 생성형 AI 모델을 구현하는 것을 목표로 한다.

2. 효율적인 데이터 표현 및 압축 연구

이미지와 비디오뿐 아니라 3D 및 4D 데이터와 같은 고차원 데이터는 정보량이 매우 크며, 이를 효과적으로 저장하

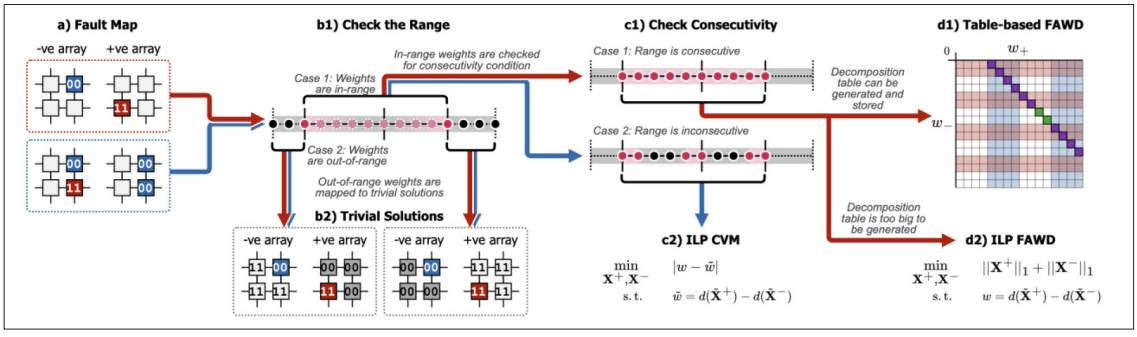


<그림 3> Gaussian splatting 기반의 3D 장면 표현 효율화 연구 [NeurIPS '25]

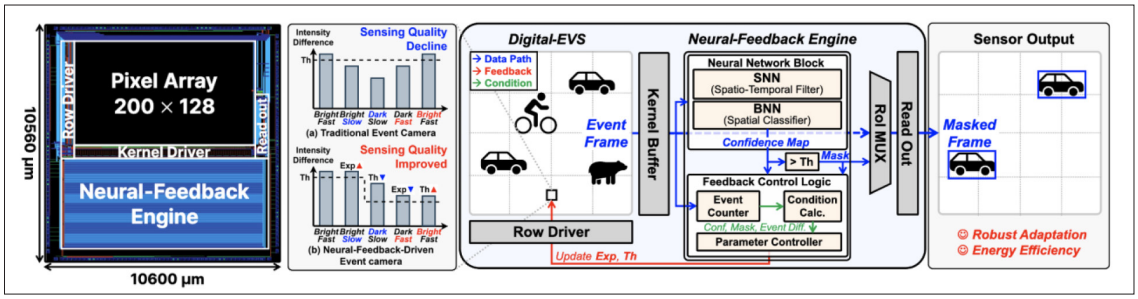
고 전송하기 위해서는 효율적인 표현과 압축 기술이 필수적이다. 연구실에서는 2D image 및 video 데이터에 대한 task-oriented representation뿐 아니라, NeRF 및 Gaussian Splatting과 같은 neural representation 기반의 3D/4D 장면 표현을 연구하고 있다. 특히 단순한 시각적 품질이 아닌 인식 및 추론 성능을 고려하여 데이터 표현을 최적화하는 방향에 주목하고 있으며, 다양한 downstream task 응용 환경에서 효율적으로 활용 가능한 representation 및 compression 기술을 개발하고 있다.

3. In-sensor / In-memory computing 연구

기존의 시스템 구조에서는 센서, 메모리, 프로세서 간 데이터 이동으로 인해 성능 및 에너지 효율 측면에서 한계가 존재하며, 이러한 문제를 해결하기 위한 새로운 컴퓨팅 패러다임이 요구되고 있다. 이에 따라 연산을 센서 내에서 수행하는 in-sensor computing과 메모리에서 수행하는 In-memory computing 기술을 연구하고 있다. 이를 통해



<그림 4> 효율적이고 강인한 in-memory computing을 위한 weight 매핑 기법 [ICCAD '25]



<그림 5> In-sensor 물체탐지 및 센서 피드백을 통한 효율적인 dynamic vision sensor [VLSI '26]

센서부터 메모리, 프로세서까지 이어지는 end-to-end 최적화를 달성하여, 데이터의 이동을 최소화하면서도 높은 효율성과 안정성을 동시에 확보하는 것을 목표로 한다.

Ⅲ. 연구 실적

- CVPR, NeurIPS, ICCV, ICML 등 ML/CV 분야 탑 컨퍼런스 20편 이상 발표 (ECCV '22 oral, CVPR '24 highlight 2편, ICLR '24 spotlight 포함)
- DAC, DATE, ICCAD, ISCA, VLSI 등 HW/System 분야 탑 컨퍼런스 20편 이상 발표
- ECCV '20 VisDrone 챌린지 우승, DAC '24 System Design Contest 3위

저 자 소 개

고 종 환

- 2004년 : 서울대학교 컴퓨터공학, 기계항공공학 학사
- 2006년 : 서울대학교 전기컴퓨터공학 석사
- 2018년 : Georgia Institute of Technology 전자컴퓨터공학 박사
- 2006년 ~ 2013년 : 국방과학연구소 선임연구원
- 2019년 ~ 현재 : 성균관대학교 부교수
- 주관심분야 : 모델 경량화/가속, 데이터 압축/표현, 인센서/인메모리 컴퓨팅