

# 2026 VLSI Review

한국과학기술원 전기및전자공학부 석사과정 이가은

## Session 9 Neural Interface and Brain-Computer Interface Circuits

이번 2026 IEEE Symposium on VLSI Technology and Circuits의 Session 9는 Neural Interface and Brain-Computer Interface Circuits라는 주제로 총 4편의 논문이 발표되었다. 이 세션은 ECoG, EEG 등 뇌신경 신호로부터 유의미한 정보를 실시간으로 추출·해석하는 온칩 지능형 신경 신호 처리 회로를 중심으로, 운동 의도 디코딩, 발작 검출, 수면 조절 등 다양한 뉴럴 인터페이스 응용을 위한 저전력·고집적 SoC 아키텍처를 탐구한다. 네 논문 모두 딥러닝 기반 디코딩 모델을 온칩에 완전히 통합하여 외부 컴퓨팅 자원에 대한 의존을 제거하고, computing-in-memory(CIM) 등 메모리 중심 연산 구조를 통해 극저전력·저지연 동작을 실리콘 검증으로 입증하고 있다는 점에서 공통된 방향성을 보인다.

**#9-1** 본 논문은 EPFL 연구 팀에서 발표한, 최초로 완전 온칩 통합된 트랜스포머 기반 BCI SoC 연구이다. 척수손상 및 뇌졸중으로 인한 사지마비 환자의 운동 기능 회복을 위해서는 외부 컴퓨터에 의존하지 않는 실시간 온칩 디코딩이 필수적이나, 기존 신경 디코딩 SoC들은 오프칩 메모리에 의존하거나 온칩 저장·연산 자원의 한계로 인해 트랜스포머와 같은 고성능 모델의 탑재가 어려웠다는 문제의식에서 출발한다. 특히 기존의 backpack-mounted 외부 디코딩 시스템은 100ms를 초과하는 큰 레이턴시를 유발해 실사용성에 제약이 있었다는 점이 강조된다.

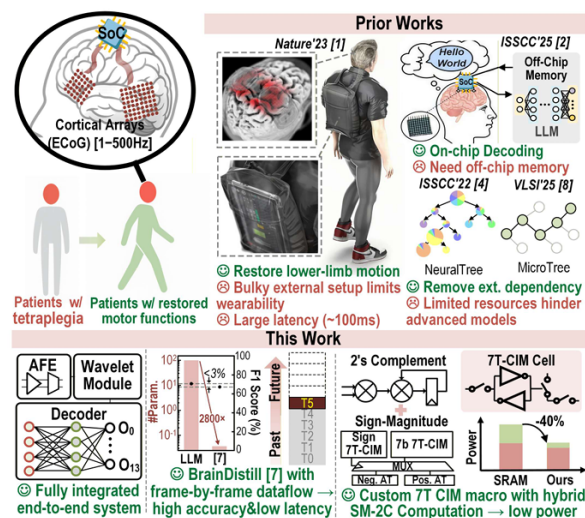
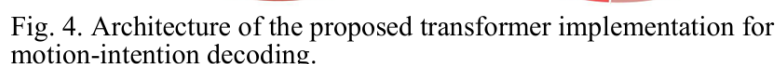


Fig. 1. Prior work (top) and contributions of this work to address shortcomings of prior art (bottom).

[그림 1] 선행연구 대비 본 연구의 contribution 구조도

AFE 부분(Fig. 3)은 4 개의 16 채널 모듈로 분할되어 멀티플렉서를 통해 공유 Gm-C 필터 및 10 비트 SAR ADC(64kS/s, 채널당 4kS/s)로 라우팅되는 구조이며, AC-coupled LNA와 백엔드 자원 공유의 아키텍처 최적화를 통해 기존 대비 30%의 면적 절감을 달성하였다. 측정 결과 채널당 3.56 $\mu$ W의 전력을 소모하면서 33~66dB의 프로그래머블 게인, 45.1dB의 SNDR, 58.1dB의 SFDR, 1kHz에서 10.6M $\Omega$ 의 입력 임피던스, 3.06 $\mu$ V<sub>rms</sub>의 입력 환산 잡음(IRN)을 달성해, 신경 디코딩에 필요한 신호 충실도를 유지하면서도 면적을 줄인 것이 특징이다.



디코더의 핵심은 커스텀 7T-SRAM CIM 매크로이며, 부호-크기(SM)와 2 의 보수(2C) 표현을 혼합 사용하는 하이브리드 SM-2C 연산 기법을 제안한 점이 이 논문의 가장 독창적인 부분이다. 활성화 값이 0 근처에 밀집되는 입력 프로젝션 단계에는 비트

토글이 적은 SM-CIM 을, 넓은 활성화 범위를 갖는 FC1/FC2 단계에는 2C-CIM 을 적용함으로써 전체 토글율을 30% 감소시킨다(Fig. 6). 8 비트 가중치는 부호-크기 비트로 분해되어 8 개의 수직 적층 7T-CIM 셀에 저장되고, isolated read/multiply(VDDH)와 retention/write(VDDL) 전압 도메인을 분리한 구조(Fig. 5)를 채택해 Monte-Carlo 시뮬레이션 기준 VDDH>0.7V 조건에서 신뢰성 있는 쓰기 동작을 확보하였다. 이 구조를 통해 전체 파라미터의 66%, MAC 연산의 72%가 CIM 어레이로 처리되며, 디코더 전력의 40% 절감 효과를 얻는다. 또한 두 개의 feature SRAM 을 두 선형 어텐션 레이어에 재사용함으로써 feature 메모리 사용량을 61.44kb 에서 14.08kb 로 4.4 배 줄였다.

65nm 공정으로 제작된 칩은 전체 10.1mm<sup>2</sup>(디코더 4.7mm<sup>2</sup>, 코어 면적 1.87×2.49mm<sup>2</sup>)를 차지하며, 온칩 SRAM 은 foundry SRAM 9.47kB 와 CIM-SRAM 26kB 로 구성된다(Fig. 8). WIMAGINE 장치로 기록한 6-클래스 상지 운동 의도 데이터셋에서 67.8%의 분류 정확도(F1-score 68%)를 달성하였으며, 손가락 동작 데이터셋에서는 77%(F1-score 75.7%)로 더 높은 정확도를 보였다. 디코더는 47μW(0.224ms 연산 지연, 10.53nJ/classification)만을 소모하며, 이는 68.6 TOPS/W 의 에너지 효율에 해당한다. 전력 breakdown(Fig. 9)을 보면 dynamic power 가 19%, leakage power 가 81%를 차지하고, foundry SRAM 의 leakage 가 전체 leakage 의 81%를 점유해 SRAM 이 여전히 지배적인 전력 소모원임을 알 수 있다. CIM 아키텍처 도입으로 소프트웨어 시뮬레이션 대비 dynamic/leakage 전력이 각각 73%/18% 감소하였고, 6-클래스 confusion matrix 에서도 Rest 클래스가 983 건으로 가장 많은 정분류 건수를 보이는 등 클래스별 편차가 존재함이 확인된다. Fig. 10 에 제시된 실제 예측 확률 궤적에서는 어깨 굴곡, 팔꿈치 신전, 손목 회내, 손 벌림/쥐기 등 6 개 클래스가 안정적으로 구분되는 것을 확인할 수 있다.

Table 1 의 비교에서 본 연구는 기존 CNN-DFT 기반 발표작들(ISSCC'24 Zhong, ISSCC'22 Shin, TBioCAS An 등) 대비 가장 낮은 전력을 유지하면서도 트랜스포머라는 더 큰 모델 규모(35KB)와 더 많은 클래스(14 클래스, 상지 동작) 지원을 동시에 달성하여, 손가락 단위의 단순 제스처 디코딩에 머물렀던 선행 연구들과 차별화된다. 특히 0.224ms 의 연산 지연은 이전 SoC 들의 <1000ms 또는 50ms 수준과 비교해 압도적으로 낮은 값이며, 이는 backpack 형 외부 디코딩 시스템(>100ms) 대비 실시간-비침습적 사용성을 크게 개선한 결과로 해석된다. 다만 정확도(67.8%)는 여전히 개선 여지가 있으며, 이는 6-클래스 상지 동작 분류가 기존 손가락 수준 태스크보다 본질적으로 더 어려운 문제라는 점을 감안할 필요가 있다.

**#9-3** 본 논문은 동일 연구실(EPFL)에서 발표한 발작(seizure) 검출용 초저전력 CIM 프로세서 TARNA 연구이다. 기존 발작 검출 SoC는 환자별 발작 데이터를 이용한 few-shot

학습에 의존하여 실사용이 어렵거나, zero-shot 방식이더라도 전극 위상(topology) 정보와 장기 시간 의존성을 반영하지 못하거나 오프칩 FFT 전처리에 의존해 완전한 온칩 통합이 어려운 한계가 있었다. 저자들은 그래프 합성곱(GCN)과 게이트가 없는 잔차 순환 상태 업데이트를 결합하여, 환자별 발작 데이터 없이도 높은 정확도와 초저전력 동작을 동시에 달성하는 구조를 제안한다.

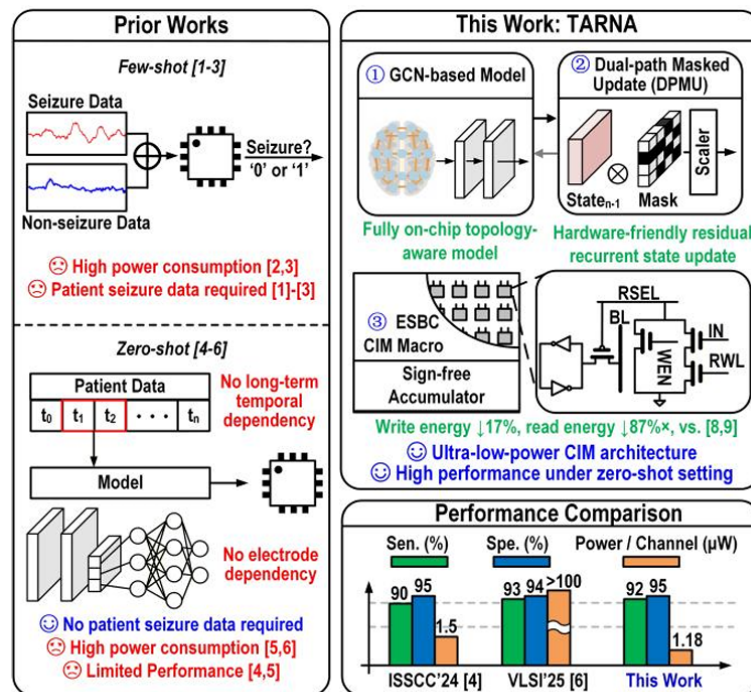


Fig. 1. Limitations of existing designs and contributions of this work.

[그림 3] 선행연구의 한계와 TARNA의 contribution 요약

TARNA의 모델 파이프라인은 전극별로 32개 샘플(0.125s, CHB-MIT 기준)을 하나의 그래프 노드로 구성하고, 4개 주파수 대역의 스펙트럴 에너지(SE)를 밀집층으로 투영한 뒤 이전 상태(State<sub>n-1</sub>)와 잔차 결합하여 시간적 문맥을 보존한다(Fig. 2). zero-shot fine-tuning이나 추가적인 데이터 기반 기법 없이도 기본 모델만으로 83% 민감도, 84% 특이도를 달성해 선행 zero-shot 연구 대비 향상된 일반화 성능을 보인다. 전극의 물리적 배치를 반영한 sparse edge matrix 기반 GCN 레이어(Fig. 3)는 공간적으로 먼 채널 간 연결을 가지치기하여, 예를 들어 첫 채널의 공간적 집계에서 포인트와이즈 컨볼루션이 19개의 부분곱을 생성하는 것과 달리 인접한 5개 채널만으로 연산을 수행함으로써 동일 크기의 CNN 레이어 대비 MAC 연산을 76% 줄이면서도 민감도·특이도를 각각 6.6%, 5.2% 개선한다.

상태 업데이트는 dual-path masked update(DPMU) 기법을 사용한다(Fig. 4). 학습 경로에서는 무작위 이진 마스크를 저장된 상태에 적용해 정규화 효과를 얻고, 추론

경로에서는 이 마스크를 우회함으로써 별도의 의사난수 생성기 없이도 비선형 게이팅(GRU 등)이 요구하는 추가 연산과 제어-메모리 오버헤드를 모두 제거한다. 상태 크기를 정렬하기 위한 스케일링 팩터가 잔차 결합 전에 적용되며, 이 gate-free 접근법은 CHB-MIT 데이터셋에서 82.6% 민감도, 84.0% 특이도를 달성해 동일 조건의 GRU 베이스라인 대비 민감도 2.6%p, 특이도 2.8%p 우수한 성능을 보인다.

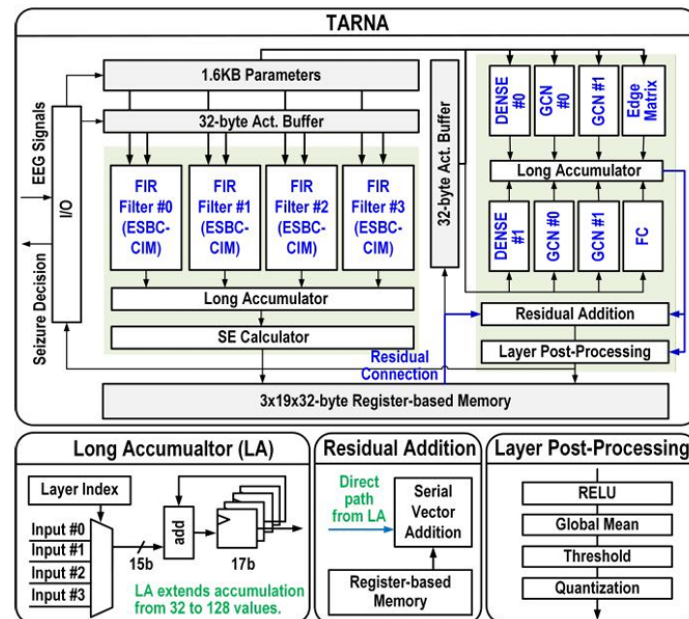


Fig. 5. Hardware architecture of the TARNa.

[그림 4] TARNa의 전체 하드웨어 아키텍처

전력 효율의 핵심은 energy-saving bit cell(ESBC) 기반 CIM 매크로이다(Fig. 6). 매크로는 1k 개의 4 비트 가중치를 저장하며, 4 개의 비트 셀이 하나의 공유 control-and-compute unit(CCU)을 사용해 bit-cell group(BCG)을 구성한다. 매크로 내부에는 8 개의 sign-free accumulator가 통합되어 있으며, 한 사이클에 32 개 가중치로부터 32 비트를 패치해 부분곱을 생성한 뒤 시프트-누적한다. 부호 있는 곱셈을 정확히 처리하기 위해 이후 사이클에서 최대 4 개의 캐리("+1") 비트를 주입하는 방식을 채택해, 누산기 입력이 비트 반전만으로 동작하도록 하고 누산기당 3 회의 덧셈 연산을 제거하였다. 읽기 시에는 CCU가 우측 접근 트랜지스터를 차단해 별도의 감지 증폭기 없이 고정 노드 전압을 형성하는 sparsity-aware read pattern(SARP)을 사용하고, 쓰기 시에는 셀을 준안정 상태로 두어 최대 4 배에 달하는 비트라인 충전 에너지를 회피한다. 이를 통해 비트당 평균 27.3%의 읽기 에너지 절감을 달성하며(Fig. 7), CCU 공유 자체만으로도 면적과 누설 전력을 각각 35%, 33.4% 절감한다. 종합적으로 기존 CIM 설계([8],[9]) 대비 쓰기 에너지 1.17 배, 읽기 에너지 1.87 배, 읽기 지연 10.2 배의 개선을 보인다.

65nm 공정으로 제작된 TARNA 는  $3\text{mm}^2(1.0\text{mm}\times3.0\text{mm})$  면적에 5,416 개의 가중치와 5.64kB 의 온칩 SRAM 만으로 구성되며, 레지스터 기반 메모리를 사용해 0.65V 의 서브스레숄드 동작을 지원한다.  $22.5\mu\text{W}(1.18\mu\text{W}/\text{channel})$ 의 전력을 소모하며, CHB-MIT 소아 발작 데이터셋 전체 환자를 대상으로 평가한 결과 100%의 이벤트 기반 민감도, 91.6%의 프레임 기반 민감도, 95.3%의 특이도, 0.048/hr 의 위양성률, 0.54s 의 검출 지연을 달성하였다(Fig. 9, Fig. 10). 전력·면적 breakdown 을 보면 분류기(classifier)가 전력의 54%, 특징 추출이 19.2%, 메모리 및 기타가 26.8%를 차지하며, 면적 측면에서는 메모리 및 기타가 36.7%로 가장 큰 비중을 차지해 전력과 면적의 지배 요인이 서로 다를 것을 보여준다. 또한 TUSZ 데이터셋(315 명 전체 환자) 평가에서도 사전 fine-tuning 없이 0.88 의 AUROC 를 기록해 진정한 zero-shot 일반화 성능을 입증하였다.

Table I 비교에서 TARNA 는 기존 발표작(ISSCC'22, JSSC'22, VLSI'25, ISSCC'23, VLSI'25, ISSCC'24) 대비 채널당 전력이  $1.18\mu\text{W}$  로 가장 낮으면서도, 학습에 사용한 테스트 환자 데이터의 비율이 0%인 진정한 zero-shot 조건에서 위양성률과 민감도 모두 우수하거나 동등한 성능을 보인다. 특히 few-shot 기반 선행 연구들이 환자 데이터의 20~35%를 학습에 사용한 것과 대비된다는 점에서, 저전력·고정확도·환자 비의존성이라는 세 가지 상충 조건을 동시에 만족시킨 사례로 평가된다.

## 저자정보



### 이가은 석사과정 대학원생

- 소속 : 한국과학기술원 전기및전자공학부
- 연구분야 : 디지털 회로 설계
- 이메일 : gelee@ics.kaist.ac.kr
- 홈페이지 : <https://idec.or.k>

