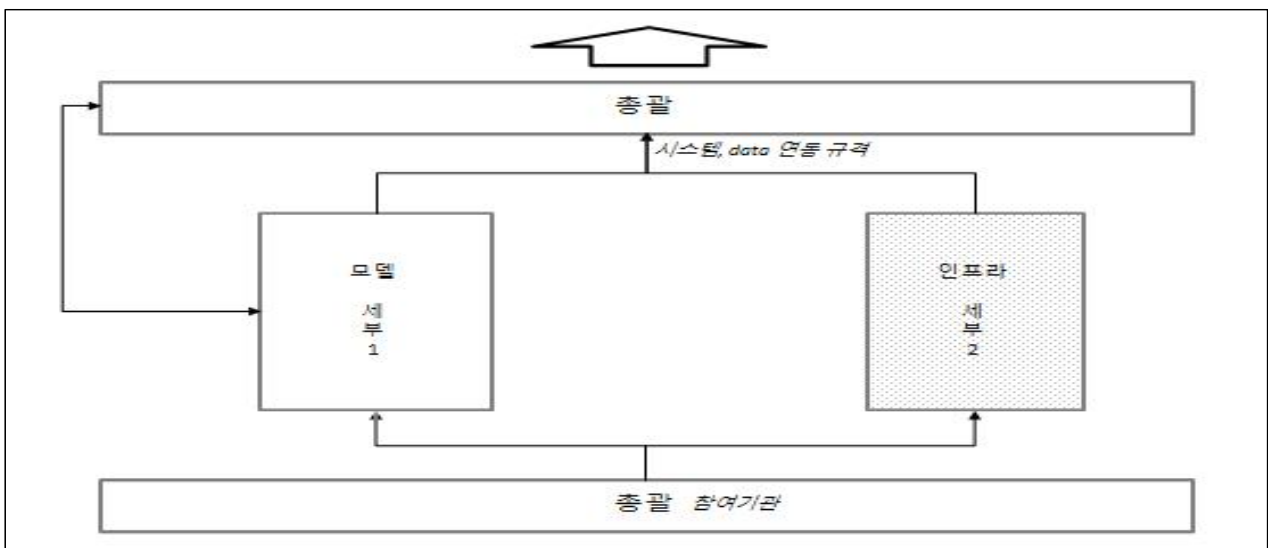


관리번호	2026-피지컬AI-33	(지정공모형, 통합형 세부1)
기술분류	대분류(인공지능) - 중분류(신뢰-산업AI) - 소분류(활용-지원AI) - 세분류(산업 특화 문제 해결 기술)	
중점분야	AI(), AI반도체(), 차세대통신(), 양자(), 사이버보안(), AI·디지털융합(√)	
총괄과제명	정밀제어를 위한 물리지능행동모델 원천기술 개발	
세부과제명	물리지능행동모델 및 통합 프레임워크 기술 개발	

1. 개요

총괄과제

- (총괄개념) 본 총괄과제는 단위 공정의 물리 법칙을 이해하는 PINN(Physics-Informed Neural Network), 전체 멀티 공정을 아우르는 물리 월드 모델(World Model), 초정밀 제어 액션을 생성하는 대규모 행동 모델(LAM)을 하나로 결합한 「통합 물리지능 행동모델(PINN-World Model-LAM) 프레임워크」를 확보하고, 이를 대규모로 학습·실증하는 NPU-GPU 기반 Co-Working 테스트베드 및 초정밀 제어용 학습데이터 플랫폼을 동시에 구축하는 제조 자율제어 원천기술 개발 사업
- 「모델(세부1)-인프라(세부2)」 결합형 병렬 구조로, 세부1이 요구하는 대규모 분산학습·시뮬레이션·실증을 세부2가 전담 지원



<총괄 및 세부과제 간 연계도>

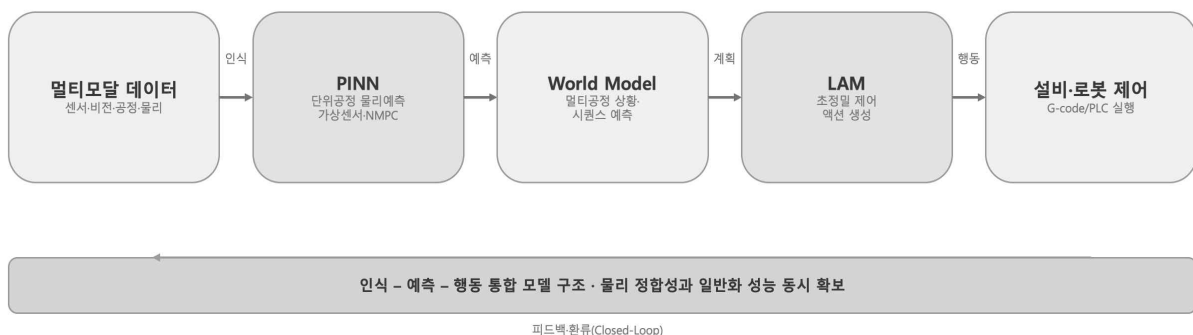
- (총괄목표) 통합형 과제의 목표 설명
 - PINN-World Model-LAM 통합 프레임워크와 이를 지원하는 NPU-GPU 혼성 학습·실증 인프라를 결합하여, 제조 현장에 일반화 적용 가능한 물리지능행동모델 원천기술 확보 — [세부1]통합 모델·프레임워크와 [세부2]학습·실증 인프라를 단일 총괄 관리체계로 결합·운영
 - 2026-피지컬AI-13 과제에 대해 공정별 PINN 모델 개발을 위한 데이터 요구사항 도출(1년차)
 - 2026-피지컬AI-28 과제의 물리행동지능모델의 전 주기 플랫폼 구축을 위한 PI-LAM 모델 연동 규격을 정의(협의)
 - 총괄-세부 과제간 협력체계와 통합 실증 방안 제시 필요
 - 세부과제와 유기적으로 협력하도록 세부과제의 공동기관으로 참여 가능

세부1

- (개념) 단위 공정의 물리 법칙을 이해하는 PINN, 전체 멀티 공정을 아우르는 물리 월드 모델, 초정밀 제어를 생성하는 대규모 행동 모델(LAM)을 결합하여, 어떤 환경에서도 일반화 가능한 「통합 물리지능행동모델 프레임워크」를 최종 확보하는 원천기술 개발
 - PINN(단위 공정 물리예측) → World Model(멀티 공정 상황예측) → LAM(초정밀 제어 액션생성)으로 이어지는 인식-예측-행동 통합 모델 구조

[세부1] 통합 물리지능행동모델 — 개념도

PINN → World Model → LAM 인식-예측-행동 통합 (물리법칙 내재화)



※ (PINN→WM→LAM 인식-예측-행동 통합)

<과제 개요>

2. 현황 및 필요성

- (기존 기술현황)
 - (해외 기술동향) 물리 시뮬레이터 · 월드모델 · 로봇 행동모델이 개별적으로 고도화되고 있으나, 제조 공정 물리법칙을 내재화한 통합 프레임워크는 초기 단계
 - (국내 동향) PINN · 예측모델 등 단위 기술은 확보되었으나, World Model · LAM(Large Action Model)과 결합한 End-to-End 통합 학습 프레임워크는 부재
 - 기존 자동화 설비는 사전 정의된 PLC 로직, 로봇 티칭 프로그램, 장비별 전용 인터페이스에 의존하고 있으며, World Model/LAM이 생성한 Action Policy와 실제 로봇 · 장비 제어 명령 사이의 실행 인터페이스가 표준화되어 있지 않음
- (필요성) 물리 정합성과 일반화 성능을 동시에 갖춘 통합 물리지능행동모델 및 분산 학습 프레임워크 확보가 제조 자율제어 실현의 핵심 전제
 - (분야별) 주조 · 금형 · 소성가공 · 용접 · 표면처리 · 열처리 등 뿌리공정과 정밀가공 · 조립 공정은 지배 물리현상(열전달 · 유동 · 소성변형 · 상변태 · 접촉/마찰)이 상이하여 단일 모델로 일반화가 불가능하며, 분야별 지배방정식 · 물리제약을 신경망에 내재화한 PINN과 이를 통합하는 World Model · LAM이 분야 공통 원천기술로 요구됨
 - (경제적 필요성) 제조 불량률(평균 6~8%) · 비계획 정지 · 에너지 손실에 따른 비용 부담이 크며, 물리 정합형 예측 · 제어로 불량 · 에너지 · 리드타임을 동시 절감해

- 제조원가 경쟁력을 확보하고 숙련인력 의존 비용과 품질 편차를 구조적으로 절감할 필요
- (산업적 필요성) PLC 로직·로봇 티칭에 의존하는 기존 자동화의 한계를 넘어 모델이 직접 행동(Action)을 생성·제어하는 End-to-End 통합 프레임워크의 부재가 산업 자율 제어 실현의 병목으로, 글로벌 제조 패권 경쟁에서 원천기술 선점이 산업 경쟁력을 좌우

3. 수요분석 및 기대효과

- (수요분석) 복합 물리 공정의 초정밀 제어 수요, 가상센서(Soft Sensor) 기반 비계측 상태 추정 수요, 제조 AI 통합 운영체제 수요
 - (산업 현장의 수요) 복합 물리공정의 초정밀 제어, 계측이 어려운 내부 상태의 가상 센서(Soft Sensor) 추정, 이상·불량의 사전 예측·자동 보정에 대한 현장 수요가 크며, 공정변수-물성-품질-설비상태를 잇는 물리 인과 기반 실시간 의사결정 수요가 전 공정으로 확대
 - (사회적 수요) 고온·고하중·유해환경 공정의 안전사고 저감과 작업자 안전 확보, 숙련 노하우의 표준 자산화를 통한 인력 단절 대응 등 사회적 수요가 증대하고, 설명 가능하고 신뢰할 수 있는 제조 AI에 대한 기대가 높아짐
 - (고부가가치 수요) 항공·방산·반도체·조선·정밀기계 등 고정밀·고신뢰가 요구되는 산업에서 마이크로미터급 정밀도와 무결성 보증이 필요한 고부가가치 제어 수요가 존재하며, 다품종·고변동 생산의 품질 안정화 수요와도 직결
 - (규제 대응 및 표준화) 정부의 피지컬AI 핵심 경쟁력 확보 전략과 연계하여 국제 표준에 부합하는 보안 기준을 선점하려는 기업들의 움직임이 활발함
- (기대효과) 물리법칙 내재형 고정밀 예측·제어 실현, PINN-World Model-LAM-Simulation으로 이어지는 인식-예측-계획-실행-피드백 전주기 제조 물리지능 프레임워크 완성, 개발되어진 월드모델과 연계되어진 Simulator 확보, 제조 전반 자율제어 패러다임 선도
 - (기술) PINN의 물리법칙 내재화로 데이터 부족 구간의 외삽 한계를 극복하고 World Model·LAM과 결합한 인식-예측-행동 통합 원천기술을 확보
 - (산업) 뿌리산업을 포함한 제조 전반의 자율제어 전환을 가속하여 불량·다운타임·에너지를 구조적으로 저감 및 모델이 행동을 생성·집행하는 패러다임으로 전환하여 OEE·품질·리드타임을 동반 개선
 - (생태계) 모델(세부1)-인프라(세부2)-현장 실증의 연계 생태계를 형성하고 1:N 오케스트레이션·표준 인터페이스로 중소·중견까지 확산 및 PINN 플러그인 등록·재사용 구조로 산학연 공동 자산화 생태계를 조성
 - (산업 경쟁력 강화) 해외 솔루션 대비 물리 정합성과 현장 일반화를 차별점으로 확보하여 글로벌 상업 경쟁력을 강화
 - (비용 절감 및 효율성 증대) 불량 저감·예지보전·에너지 최적화로 직접 원가를 절감하고 가상센서·시뮬레이터 기반 사전검증으로 시운전·셋업 비용과 시간을 단축
 - (사회적 수용성 제고) 물리법칙 위배 행동 자동차단·Fail-safe·Human-in-the-loop·설명가능(XAI) 체계로 안전성과 투명성을 확보하여 현장 수용성을 제고. 위험 공정의 무인·원격화로 작업자 안전과 노동환경을 개선

- (새로운 시장 창출) Physical AI 프레임워크·시뮬레이터·서빙 플랫폼이라는 신시장을 창출하고 데이터·모델·검증을 잇는 신규 밸류체인을 형성. 제조 외 로봇·물류·에너지 등 인접 도메인으로 시장을 확장

4. 연구개발목표

○ 최종목표

- PINN, 물리정보 World Model LAM을 통합한 「통합 물리지능행동모델(PINN-World Model-LAM) 프레임워크」, 개발되어진 월드모델과 연계되어진 Simulator 1식 확보

○ 정량적 연구개발목표

성능지표		단위	연구개발 목표치	연구개발 전 국내 수준	세계 최고수준 (보유국/보유기관)
1	AI모델 안정성 검증 모듈의 위험 입력 탐지 정확도	%	95%	N/A	70% (미국/Meta)
2	신뢰무결성 보장시스템의 입력조작 및 민감정보 탐지정확도	%	99%	N/A	98% (미국/MIT)
3	대규모 행동 모델(LAM) 과업 성공률	%	> 90	N/A	51% (미국, NVIDIA)
4	LAM 추론 지연시간	ms	< 50	N/A	67 ms (미국, NVIDIA)
5	End-to-End Fine-tuning 시간	GPU-Hour	20GPU-Hour 이내	2,688 시간	2,304 시간 (미국, NVIDIA)

- 적대적 입력(FGSM-PGD 등 교란)으로부터 모델을 보호하는 안전시스템의 공격 탐지 정확도로, 물리 일관성 기반 이상탐지를 결합해 강건성 확인
 - (측정방법) 공격·정상 혼합셋에서 (TP+TN)/전체, 공격기법·교란강도(ϵ)·표본수 사전 고정·공개
 - 신뢰무결성 보장시스템 입력조작·민감정보 탐지 정확도 (% 99) : (정의·기술) 입력 변조·민감정보 유출 시도를 탐지해 데이터·모델 무결성을 보장(해시·서명 검증, 이상입력·PII(Personally Identifiable Information) 식별 결합)
 - (측정방법) 조작·정상·민감정보 라벨 테스트셋의 탐지정확도, 오탐(FP)·미탐(FN)률 병기
 - LAM이 부여된 제조로봇 과업을 완수하는 비율(조건부장기 행동시퀀스 생성 후 시뮬레이션/실설비 완료 판정)
 - (측정방법) 정의된 과업셋 N회 중 성공판정 비율, 성공기준(허용오차·시간)·시도횟수 사전 정의
 - 입력→제어 액션 출력까지의 추론 지연으로 실시간 제어 적용 가능성 지표(경량화온디바이스 추론으로 저지연)
 - (측정방법) 동일 입력·배치=1·지정 HW에서 단일 추론 평균·p95, HW·정밀도(FP8 등) 명시
 - 통합 프레임워크 미세조정 누적 GPU 연산시간으로 분산학습·종단간 최적화로 적은 GPU-Hour 수렴. LAM특화 Parameter-Efficient Fine-tuning, 경량 튜닝 기준
 - (측정방법) 동일 데이터셋·수렴기준 하 누적 GPU-Hour=GPU수×시간, 기준 GPU·하이퍼파라미터 고정
- ※ 대규모 분산학습·시뮬레이션·실증 수행을 위해 필요한 연산자원의 확보·활용 계획을 제안서에 구체적으로 제시하여야 함

○ 공통 요구사항

- (무빙타겟) 기술 진화가 빠르게 전개되는 분야임을 감안하여, 기술 환경 변화를 반영하기 위한 연구개발 추진 체계 제시
- (사업화·상용화) 연구개발 결과물의 산업 확산에 대한 사업화·상용화 방안 제시

5. 연구개발내용

- ① **(물리 법칙 내재 PINN 모델 개발)** 각 단위 공정의 복합 물리 현상 예측 한계를 분석하고, 지배 방정식을 신경망에 내재화한 고정밀 PINN 모델 개발
 - (설계·초기모델) 질량·운동량 보존 법칙 및 열유체 흐름 지배방정식 등 핵심 물리 법칙을 신경망 손실 함수의 제약 조건으로 명시적으로 포함하여 외삽(Extrapolation) 한계를 극복하고 물리적 인과관계를 신경망에 내재화
 - (고도화) 학습된 PINN을 가상 센서(Soft Sensor)로 활용해 측정이 어려운 공정 내부 상태를 실시간 모니터링, 운전 중 데이터로 재학습하여 지배방정식 파라미터를 보정하고 비선형 모델 예측 제어(NMPC) 아키텍처 구축
- ② **(대규모 다중 모달리티 물리 월드 모델 개발)** 센서·비전 등 이기종 데이터를 융합하여 멀티 공정 전반의 상황 변화와 장기 시퀀스를 예측하는 월드 모델(예측모델, 불확실성 모델링, 계층적 구조 설계) 개발
- ③ **(상황 인지 기반 대규모 행동 모델(LAM) 개발)** 월드 모델의 인식·예측 결과를 바탕으로 초정밀 설비 제어·로봇 작업을 위한 구체적 액션(레시피 등)을 자동 생성(다중 모달리티 처리, 조건부·장기적 행동순서 생성)
- ④ **(End-to-End 학습 프레임워크 통합)** 데이터 수집부터 PINN-World Model-LAM 추론까지 유기적으로 결합하여 GPU/NPU 클러스터 환경에서 분산 학습 및 종단 간 최적화가 가능한 프레임워크(통합 학습 알고리즘·파이프라인) S/W 개발
 - (통합 프레임워크) 핵심 모듈 통합 및 모듈 간 데이터·제어 신호를 표준 인터페이스로 정의, 센서 입력 → 상태 추정·월드모델 예측 → 행동 모델 제안 → 제어신호 출력의 최종 의사결정 구조 완성
- ⑤ **(단일 World Model & LAM-다수 PINN 오케스트레이션, 1:N)** 단일 LAM이 공정·제조사별로 서로 다른 다수의 PINN(1:N)을 표준 인터페이스로 오케스트레이션하여 멀티 공정을 아우르는 통합 행동계획을 생성하고, PINN 표준 규격(입출력 스키마·지배방정식 표현), 오케스트레이션 로직, 개발 SDK를 통해 신규 PINN을 코드 변경 없이 동적으로 등록·연동
 - (표준 규격) 모든 PINN이 동일하게 따르는 입출력 스키마와 지배방정식 표현 규격을 정의하여 공정·제조사 간 상호운용성 확보
 - (오케스트레이션 로직) LAM이 공정 상황에 따라 필요한 PINN 집합을 선택·조합·호출하여 통합 행동계획으로 융합
 - (개발 SDK·동적 등록) 신규 PINN을 표준 규격에 맞춰 등록하면 코드 변경·재빌드 없이 런타임에 자동 인식·연동(플러그인형 확장)

- ⑥ (AI 시뮬레이터 개발) 개발된 물리지능행동모델(PINN-World Model-LAM)들이 하나의 통합된 AI 운용공간에서 인식·계획·행동을 수행하도록 하는 시뮬레이터 개발(가상 공정 환경에서 모델 간 상호작용·행동계획·제어 결과를 통합 검증)

6. 지원기간/예산/추진체계/특기사항

○ 연구개발기간 : 5년 이내(' 26.8월~' 30.12월)

○ 정부지원연구개발비

구분	기간	개월수	정부지원연구개발비(억원)
1년차	26.08~26.12	5개월	202.38
2년차	27.01~27.12	12개월	317.45
3년차	28.01~28.12	12개월	300.82
4년차	29.01~29.12	12개월	354.98
5년차	30.01~30.12	12개월	307.37
합계	26.08~30.12	53개월	1483.00

* 연차별 정부 지원 연구개발비는 당해연도 예산 심의 결과에 따라 변동될 수 있음

○ 주관기관 : 제한없음

- 다수 참여기관과의 연계-협력을 주도할 수 있는 주관기관 역할 필요

○ 통합형 과제 유의사항

- 본 과제는 통합형 과제로, 총괄 및 세부 과제 컨소시엄을 구성하여 신청
- 선정평가 시, 총괄 및 세부 과제 통합 선정

* (2026-피지컬AI-33)총괄/세부1 ~ (2026-피지컬AI-34)총괄/세부2

○ 수행체계 : 주관 및 참여기관의 역할 제시 필수

- 국가균형발전 연계 : 비수도권 제조 클러스터 소재 기관의 참여와 지역 제조 생태계와의 실증 방안 제시 권고
- 국무회의 의결 근거, 경남 제조업 현장에 대해 피지컬AI 학습·실증을 위한 과제 수행체계와 지역에 대한 기대효과 제시 필요
- 사업 목적 달성을 위해 선정 이후 주관/참여기관으로 구성하는 협의체 운영 예정
- 2026-피지컬AI-13 과제에 대한 융합데이터 요구사항 도출 및 반영 절차 제시
- 2026-피지컬AI-28과제와 PI-LAM 모델 연동 규격 협의 절차 제시

연구유형	기초연구 (), 응용연구 (), 개발연구 (√)	TRL (5)~(7)단계
과제특징	국무회의 의결 예비타당성조사 면제와 입지적정성(경남) 검토를 거쳐 추진	
구분		기술분야명/팀명
책임PM		피지컬AI/지역AX R&D센터
담당팀장		피지컬AI/지역AX거점팀
		성명
		이준우
		윤정섭

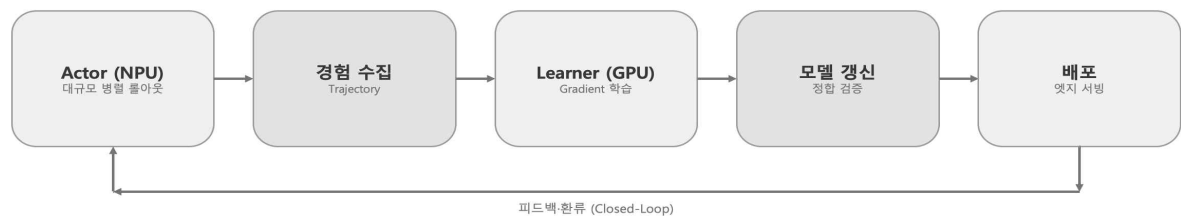
관리번호	2026-퍼지컬AI-34	(지정공모형, 통합형 세부2)
기술분류	대분류(인공지능) - 중분류(신뢰-산업AI) - 소분류(활용-지원AI) - 세분류(산업 특화 문제 해결 기술)	
중점분야	AI(), AI반도체(), 차세대통신(), 양자(), 사이버보안(), AI·디지털융합(√)	
총괄과제명	정밀제어를 위한 물리지능행동모델 원천기술 개발	
세부과제명	물리지능행동모델을 위한 co-working 테스트베드 기술개발	

1. 개요

- (개념) 물리지능행동모델의 원천 핵심 기술 개발 및 실증을 지원하기 위한 학습데이터 센터 및 NPU-GPU 기반의 Co-Working 테스트베드를 구축 세부1의 물리지능행동 모델(PINN · World Model · LAM) 원천기술을 대규모로 학습 · 시뮬레이션 · 실증할 수 있도록, 국산 NPU-GPU 기반 학습데이터센터와 Co-Working 테스트베드를 구축하는 인프라 개발 · 실증 사업
 - 규모 몰아웃은 NPU(Actor)에 학습은 GPU(Learner)로 분리하는 이기종 강화학습 구조로 학습 속도 · 에너지 효율 극대화하는 인프라
 - 다수 참여기업의 공동 활용(가상화 기반 기업별 독립 컴퓨팅 환경 제공)을 동시에 지원하는 공유 인프라

[세부2] Co-Working 테스트베드·학습 인프라 — 개념도

Actor(NPU) 몰아웃 ↔ Learner(GPU) 학습 분리 강화학습



※ 몰아웃=GPU(Actor) 학습=GPU(Learner) 분리 이기종 강화학습

<과제 개요>

2. 현황 및 필요성

○ (기존 기술현황)

- (해외 기술동향) 대규모 AI 학습은 GPU 클러스터 중심이며, 추론 · 엣지 배포에 NPU를 결합한 이기종 강화학습 인프라가 부상
- (국내 동향) 국산 NPU · AI 가속기 기술은 확보 단계이나, 제조 AI 대규모 학습 · 실증을 위한 NPU-GPU 혼성 클러스터 및 전용 학습데이터센터 레퍼런스는 부재

○ (필요성) 국산 NPU 기반 제조 AI 전용 학습 플랫폼과 현장 실증 테스트베드 확보가 모델 자립 · 확산의 필수 기반

- (분야별) 대규모 학습은 GPU 클러스터, 추론 · 엣지 배포는 국산 NPU가 강점으로 분야별 최적 가속기가 상이하여, 학습(Learner=GPU) · 몰아웃(Actor=NPU)을 분리한 이기종 강화학습 인프라와 분야 공용 학습데이터센터가 분야 공통 기반기술로 요구됨
- (경제적 필요성) 해외 GPU 종속에 따른 도입 · 운영 · 전력 비용 부담이 크며, 국산 NPU 기반 저전력 추론 · 혼성 클러스터로 학습 · 추론 TCO를 절감하고 공동 활용(가상화) 인프라로 참여기업의 중복 투자를 회피하여 자원 효율을 극대화할 필요

- (산업적 필요성) 제조 AI 대규모 학습·실증을 위한 NPU-GPU 혼성 클러스터·전용 학습데이터센터 레퍼런스가 부재하여, 국산 AI 가속기 생태계 육성과 모델 자립·확산, 공급망 안정성·기술주권 확보 차원의 산업적 필요성이 큼

3. 수요분석 및 기대효과

- (수요분석) 대규모 행동모델 학습용 고성능 컴퓨팅 수요, 국산 NPU 기반 저전력 추론·엡지 배포 수요, 제조 데이터 보안 수요
 - (산업 현장의 수요) 대규모 행동모델 학습용 고성능 컴퓨팅, 국산 NPU 기반 저전력 실시간 추론·엡지 배포, 제조 공정 기밀을 보호하는 데이터 보안에 대한 현장 수요가 크며, 다수 기업의 공동 활용·격리 환경 수요가 동반
 - (사회적 수요) AI 인프라의 전력·탄소 부담 저감(저전력 추론)과 국산 기술 자립에 대한 사회적 기대가 높고, 데이터 주권·보안에 대한 신뢰 요구가 증대하여 안전한 공동 활용 인프라 수요가 확대
 - (고부가가치 수요) 초대규모 파라미터 모델 학습·시뮬레이션, 고신뢰·고가용(99.9%+) 운영이 요구되는 고부가가치 컴퓨팅 수요가 존재하며, 기업별 전용 영역·페타바이트급 스토리지 등 프리미엄 인프라 수요와 직결
 - (규제 대응 및 표준화) 정부의 퍼지컬AI 핵심 경쟁력 확보 전략과 연계하여 국제 표준에 부합하는 보안 기준을 선점하려는 기업들의 움직임이 활발함
- (기대효과) 대규모 행동모델을 수용하는 제조 AI 전용 학습 플랫폼 구현으로 국산 AI 가속기(NPU) 인프라 기술력 입증 및 국산 NPU 기반 AI 학습플랫폼 글로벌 레퍼런스 확보
 - (기술) Learner(GPU)-Actor(NPU) 분리형 이기종 강화학습 구조로 학습 속도·에너지 효율을 극대화하는 인프라 기술을 확보. NPU 모델 이식·경량화·SLO 스케줄링·정합 검증(WYTIWYS) 등 운영 원천기술을 정립
 - (산업) 국산 NPU 기반 제조 AI 전용 학습 플랫폼을 구현하여 국산 AI 가속기 인프라 기술력을 입증하고 산업 적용을 확대.
 - (생태계) 국산 NPU-GPU 혼성 클러스터의 글로벌 레퍼런스를 확보하여 AI 반도체·인프라·제조가 연계된 산학연 공동 활용 생태계를 형성
 - (산업 경쟁력 강화) 저전력·고효율 추론(perf/watt)과 무중단 고가용 운영으로 해외 GPU 종속 대비 비용·전력 경쟁력을 확보.
 - (비용 절감 및 효율성 증대) vGPU·가상화·공유 스토리지로 자원 활용률을 높이고 NPU 저전력 추론으로 전력비를 절감. 공동 활용 인프라로 참여기업의 설비투자(CapEx)·운영비(OpEx)를 동반 절감
 - (사회적 수용성 제고) 다층 보안(Defense-in-Depth)·데이터 주권·정보보호 인증으로 기밀 보호 신뢰를 확보하여 공동 활용 수용성을 제고. 저전력·친환경 AI 인프라로 사회적·환경적 수용성을 향상
 - (새로운 시장 창출) 국산 NPU 기반 제조 AI 학습·추론 인프라라는 신시장과 'AI 인프라 as a Service' 모델을 창출. 테스트베드·공정셀 레퍼런스를 통한 실증·확산 서비스 신규 수익원을 형성

4. 연구개발목표

○ 최종목표

- 물리지능행동모델을 실증할 수 있는 학습인프라로 NPU-GPU 혼성 연산장비와 공정 실증을 위한 테스트베드 구축

○ 정량적 연구개발목표

	성능지표	단위	연구개발 목표치	연구개발 전 국내 수준	세계 최고수준 (보유국/보유기관)
1	시스템 가용성	%	> 99.9	99.9 (NAVER Cloud)	99.9 (미국, AWS)
2	데이터 파이프라인 처리량	%	> 30	N/A	N/A
3	추론 인프라 전력 효율	% (perf/watt)	>40%	N/A	기준 대비 50% 우위 (미국, Google)
4	LAM 추론 지연시간	ms	< 50	N/A	N/A
5	연산 효율	(TPS/랙)	26,400	N/A	82,749 TPS (미국, NVIDIA)
6	실증 공정 환경 유형수	종	4	N/A	N/A

- 무중단 서비스 가능 시간 비율(연 다운타임 $\approx 8.76\text{h}$ 이내)로 HA·이중화·24/365 운영으로 달성
 - (측정방법) 가용시간/총시간 $\times 100$, 다운타임 분류코드·측정기간 정의
- 기준 시뮬레이션-추론-학습 루프 대비 처리효율 향상률(Actor-Learner 분리·병렬·캐싱)
 - (측정방법) (개선후 처리량-기준)/기준 $\times 100$, 기준 파이프라인·데이터·HW 고정
- 단위 전력당 추론 성능=성능÷소비전력으로 국산 NPU로 전력효율 극대화
 - (측정방법) 과제 종료시점 기준 GPU 대비 성능/소비전력 비율, 30B 규모의 모델을 공랭식 15kW 랙에서 측정한 것을 기준으로 함.
- 행동모델 추론 지연으로 경량화·온디바이스(NPU) 추론으로 저지연 달성
 - (측정방법) 지정 NPU·배치=1 단일 추론 평균·p95, HW·정밀도 명시
- 동일 전력등급 랙 기준 초당 처리 토큰수(국산 NPU·대규모 모델 기준)
 - (측정방법) 대표 LLM·FP8·동일 전력등급 랙 평균 TPS, 모델·시퀀스·전력등급 사전 고정
- 실증을 위해 구축된 상이한 공정 환경(대상 공정, 테스트베드)의 유형수(4종)
 - (측정방법) 실증환경 구축·운영 완료 종수, 실증환경 정의·완료기준 사전 명시

○ 공통 요구사항

- (무빙타겟) 기술 진화가 빠르게 전개되는 분야임을 감안하여, 기술 환경 변화를 반영하기 위한 연구개발 추진 체계 제시
- (사업화·상용화) 연구개발 결과물의 산업 확산에 대한 사업화·상용화 방안 제시

5. 연구개발내용

- (하이브리드 클라우드 기반 학습데이터 센터 구축) 온프레미스의 장점과 클라우드 확장성을 결합한 하이브리드 환경 구축, 워크로드 기반 리소스 효율 운영, 24시간 365일 무중단 운영 목표
 - (기반환경 설비) 학습데이터 센터 위치의 전력·하중 등 공간적 기초환경 확인 및 기반설비·운영통제실 구축

- (컴퓨팅 기반환경) GPU/CPU 서버, 클러스터링 SW 운용서버·SW, 고속네트워크 스위치, 워크스테이션 등 단계적 도입 및 가상화 환경 구성
- (가상화 서버) 하이퍼바이저 기반 VM 또는 컨테이너 오케스트레이션으로 참여 기업별 독립 컴퓨팅 환경 제공
- (고성능 GPU 인프라) 최신 고성능 GPU 자원을 동적 할당하는 가상 GPU(vGPU) 기술 적용
- (스토리지 가상화) 페타바이트급 대규모 데이터 처리를 위한 고속 플래시 기반 고성능 공유 스토리지, 기업별 전용 영역 논리 분리

② (다층 보안 아키텍처(Defense-in-Depth) 설계) 참여 제조기업 핵심 기밀인 제조 공정 데이터에 대한 인프라·데이터·접근 보안 아키텍처 설계

③ (이기종 GPU-NPU 강화학습 클러스터 구축) 대규모 병렬 롤아웃을 NPU(Actor)에 할당하고 Gradient 업데이트를 GPU(Learner)에 분리하는 아키텍처로 학습 속도·에너지 효율 극대화

- (NPU 혼성 클러스터) ① 학습(Learner) 계층: 초고성능 학습용GPU 서버군으로 World Model·LAM 사전학습·풀파인튜닝(초대규모 파라미터급), ② 물리 시뮬레이션 계층: 고성능 시뮬레이션GPU로World Model 기반 물리 시뮬레이션·합성데이터 생성, ③ Actor/Edge 계층: 국산 추론 가속기(NPU) 다수로 대규모 롤아웃 병렬·실시간 제어를 담당
- (네트워크) NPU(Actor)를 GPU Farm에서 분리함에 따라 발생하는 네트워크 복잡도를 개선하는 운영
- (현장실증) 대규모 모델 학습부터 로봇·제조 현장 배포까지 연결되는 전주기 물리 지능행동모델 확산 인프라 체계 구현
- (연산 인프라 단계별 구축) 1단계(시뮬레이션GPU+국산NPU+공통 인프라 우선 가동→폐루프 연동 검증→GPU 학습 계층 추가→3계층 통합 운영), 2단계(자원관리·SLO 스케줄링·다수 기업 공동활용 확대·운영 표준 정착)
- (운영·플랫폼R&D) GPU에서 학습된 모델은 그대로 국산 NPU에서 동작하지 않으므로, ①국산 NPU 모델 이식·경량화·최적화 기술 개발(이식 파이프라인→양자화·컴파일→병목 분석·커널 최적화→학습-배포 정합성(WYTIWYS) 검증→확산), ②이종 GPU·NPU 자원을 하나의 체계로 운용하기 위한 다음 기술을 개발(통합SW 아키텍처→자원 풀·모니터링→SLO 스케줄링→운영 표준화). 운영·보안은 석·박사급 전담 운영인력, 공동활용(신청-심사-배분), 접근제어·자원격리·암호화·로그, 정보보호 인증, 24시간 모니터링·장애 대응

④ (공정셀 및 로봇 테스트베드 구축) 실제 공정 외란·설비 상호작용·현장 조건에서 물리지능행동모델의 제어 성능을 검증하는 최종 실증 기반 구축

- (구성) 제조공정 실증 셀, 다공정 연계 로봇·이송, 외란·이상 검증 설비, 안전·성능관리 운영시스템, End-to-End 통합제어

6. 지원기간/예산/추진체계/특기사항																															
○ 연구개발기간 : 5년 이내(' 26.8월~' 30.12월) ○ 정부지원연구개발비 <table border="1"> <thead> <tr> <th>구분</th> <th>기간</th> <th>개월수</th> <th>정부지원연구개발비(억원)</th> </tr> </thead> <tbody> <tr> <td>1년차</td> <td>26.08~26.12</td> <td>5개월</td> <td>62.03</td> </tr> <tr> <td>2년차</td> <td>27.01~27.12</td> <td>12개월</td> <td>103.57</td> </tr> <tr> <td>3년차</td> <td>28.01~28.12</td> <td>12개월</td> <td>132.97</td> </tr> <tr> <td>4년차</td> <td>29.01~29.12</td> <td>12개월</td> <td>50.02</td> </tr> <tr> <td>5년차</td> <td>30.01~30.12</td> <td>12개월</td> <td>38.01</td> </tr> <tr> <td>합계</td> <td>26.08~30.12</td> <td>53개월</td> <td>386.60</td> </tr> </tbody> </table> * 연차별 정부 지원 연구개발비는 당해연도 예산 심의 결과에 따라 변동될 수 있음 ○ 주관기관 : 제한없음 - 다수 참여기관과의 연계-협력을 주도할 수 있는 주관기관 역할 필요 ○ 통합형 과제 유의사항 - 본 과제는 통합형 과제로, 총괄 및 세부 과제 컨소시엄을 구성하여 신청 - 선정평가 시, 총괄 및 세부 과제 통합 선정 * (2026-피지컬AI-33)총괄/세부1 ~ (2026-피지컬AI-34)총괄/세부2 ○ 수행체계 : 주관 및 참여기관의 역할 제시 필수 - 국가균형발전 연계 : 비수도권 제조 클러스터 소재 기관의 참여와 지역 제조 생태계와의 실증 방안 제시 권고 - 국무회의 의결 근거, 경남 제조업 현장에 대해 피지컬AI 학습·실증을 위한 과제 수행체계와 지역에 대한 기대효과 제시 필요 - 사업 목적 달성을 위해 선정 이후 주관/참여기관으로 구성하는 협의체 운영 예정				구분	기간	개월수	정부지원연구개발비(억원)	1년차	26.08~26.12	5개월	62.03	2년차	27.01~27.12	12개월	103.57	3년차	28.01~28.12	12개월	132.97	4년차	29.01~29.12	12개월	50.02	5년차	30.01~30.12	12개월	38.01	합계	26.08~30.12	53개월	386.60
구분	기간	개월수	정부지원연구개발비(억원)																												
1년차	26.08~26.12	5개월	62.03																												
2년차	27.01~27.12	12개월	103.57																												
3년차	28.01~28.12	12개월	132.97																												
4년차	29.01~29.12	12개월	50.02																												
5년차	30.01~30.12	12개월	38.01																												
합계	26.08~30.12	53개월	386.60																												
연구유형	기초연구 (), 응용연구 (), 개발연구 (√)		TRL (5)~(7)단계																												
과제특징	국무회의 의결 예비타당성조사 면제와 입지적정성(경남) 검토를 거쳐 추진																														
구분		기술분야명/팀명	성명																												
책임PM		피지컬AI/지역AX R&D센터	이준우																												
담당팀장		피지컬AI/지역AX거점팀	윤정섭																												