Market Insight | 2025.1.16

그림: Midjourney(인간과 AI가 같이 게임하는 세상)

MIRAE ASSET 미래에셋증권

AI Weekly

엔비디아는 근본부터 AI 기업

한종목
chongmok.han@miraeasset.com

김은지
eunji.kim.a@miraeasset.com

Highlight of the Week

I. AI Issue: CES 2025의 주제는 AI고 그것을 리드하는 주인공은 엔비디아

엔비디아는 CES 2025에서 차세대 그래픽 카드 GeForce RTX 50 시리즈를 공개하며, AI 기반 그래픽 기술의 새로운 패러다임을 제시. 특히, 중급형 모델인 RTX 5070이 이전 세대 최상위 모델인 RTX 4090급 성능을 단 549달러에 제공한다는 점은 그래픽 카드 시장에 큰 반향을 일으킬 것으로 예상.

이번 발표의 핵심은 AI 업스케일링 기술인 DLSS의 최신 버전, DLSS 4. DLSS 4는 기존 DLSS 3 대비 4배 향상된 프레임 생성 기술을 제공하며, 특히 새롭게 탑재된 '멀티 프레임 생성(MFG)' 기능은 기존 렌더링 방식 대비 최대 8배까지 프레임률을 향상시킬 수 있음. DLSS 4의 혁신은 생성 AI의 주축 아키텍처인 트랜스포머를 접목했기에 가능.

RTX 5070이 RTX 4090급 성능을 내는 것은 DLSS 4의 '멀티 프레임 생성' 기능을 사용했을 때만 가능하다는 한계는 존재하나, 이는 그래픽 카드 성능 평가의 패러다임이 변화하고 있다는 게 훨씬 중요. 결론적으로, 엔비디아 RTX 50 시리즈와 DLSS 4는 GPU의 순수 쉐이더 성능을 중시하던 시대에서 AI 소프트웨어 기반 그래픽 성능을 중시하는 시대로의 전환을 알리는 신호탄. 엔비디아는 하드웨어와 소프트웨어 기술을 결합하여 그래픽 기술의 미래를 선도하고 있으며, 이러한 변화 속에서 가장 큰 수혜를 입을 기업으로 주목.

한편, 젠슨 황의 CES 2025 기조연설을 복기해보면, '엔비디아의 지난 30년이 AI 네이티브 기업으로서의 미래를 그리기 위함'이었음을 알 수 있음. 젠슨 황은 AI 발전 단계를 인지 AI, 생성 AI, 에이전트 AI, 물리적 AI로 구분하고, 또한 새로운 패러다임인 '테스트 시간 스케일링'이 엔비디아의 컴퓨팅 수요를 증가시킬 것이라 전망. GB200 NVL72 시스템은 이전 대비 4배 향상된 전성비를 제공하며, '테스트 타임 스케일링' 시대에 최적화된 제품.

또한, 에이전트 AI 인프라로 NeMo, NIM, Agent Blueprint를 제시하고, 특히 Agent Blueprint를 통해 AI 에이전트 개발 가속화를 예고. 더 나아가, 물리적 AI 비전의 핵심은 'Cosmos' 모델로, 2천만 시간 분량의 비디오와 Omniverse 3D 데이터를 학습하여 현실에 가까운 가상 세계를 생성. 이는 디지털 트윈, 자율주행, 로보틱스 분야에 혁신을 가져올 것으로 강조되었음.

'프로젝트 DIGITS'는 페타플롭스급 AI 성능을 제공하는 소형 AI 슈퍼컴퓨터로, 3,000달러의 가격에 2025년 5월 출시 예정. 개인용 클라우드 컴퓨팅 시대를 열 것으로 기대.

결론적으로, 엔비디아는 CES 2025를 통해 AI 플랫폼 기업으로 도약하겠다는 포부를 명확히 드러냈음. 물리적 AI를 위한 '월드 파운데이션 모델'인 “Cosmos”는 엔비디아가 하드웨어와 소프트웨어를 넘어 현실과 가상을 연결하고, AI를 통해 새로운 가치를 창출하는 플랫폼 기업으로 도약하고 있음을 보여줌.

II. Paper of the week: ChipNeMo, ChipAlign (by 엔비디아)

엔비디아가 개발한 ChipNeMo는 반도체 설계 관련 전문 지식을 학습한 언어 모델이나 일반 LLM에 비해 지시사항 준수 능력이 부족하다는 한계가 있었음. 이를 해결하기 위해 엔비디아는 ChipAlign이라는 새로운 방법을 제안함. 칩 전문가 모델과 일반 LLM의 가중치를 결합해 두 모델의 장점을 모두 갖춘 새로운 모델을 생성하는 것.

ChipAlign은 LLM 가중치가 리만 다양체 위에 존재한다고 가정하고, 지오데식 보간법을 사용해 가중치를 결합함. 이는 단순 평균이나 가중치 차이를 이용하는 기존 방식과 달리 결합 과정에서 기존 모델의 성능 손실이 거의 없음. 이를 통해 생성된 ChipAlign 모델은 기존 병합 방법보다 우수한 성능을 보였으며, 특히 지시사항이 포함된 질문에서 설계 자동화 성능이 크게 개선됨.

또한 ChipAlign은 기존 ChipNeMo의 도메인 지식 수준을 유지하면서도 프롬프트의 지시에 부합하는 답변을 제공하는 데 성공함. 전문성과 사용자 편의성을 모두 갖춘 모델을 개발하는 데 성공한 엔비디아가 이를 기반으로 제품 개발 효율성을 추가 개선할 것으로 기대.

표 1. AI 관련 주요 일정

일	월	화	수	목	금	토
12	13	14	15	16	17	18
·	·	·	·	· TSMC 실적	·	·
19	20	21	22	23	24	25
·	· 위스트론 실적(예)	·	· APH 실적	· 하이닉스 실적	·	·
26	27	28	29	30	31	1
·	·	· SAP 실적	· Autonomous(~30) · IBM 실적 · TSLA 실적 · MSFT 실적 · META 실적 · NOW 실적	· GOOGL 실적(예) · AAPL 실적 · INTC 실적	· QCOM 실적(예) · AMZN 실적(예)	·

자료: Bloomberg, 미래에셋증권 리서치센터

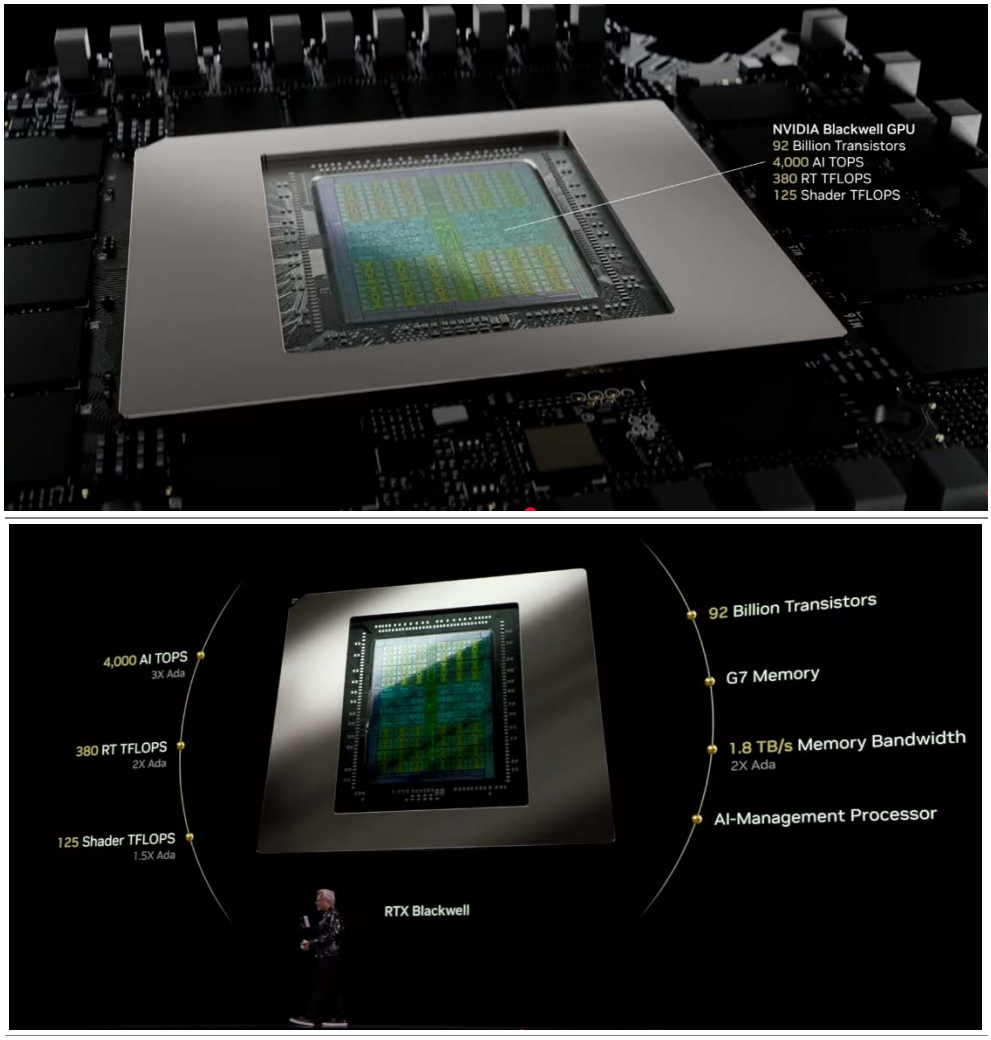
I. AI Issue: CES 2025의 주인공, 엔비디아

1. 소매용 GPU에 AI를 접목해 새로운 표준을 만들어가는 엔비디아

(1) 소비자 가전 박람회답게, 엔비디아의 첫 화두는 "RTX 50" 시리즈

2025년 1월, 라스베이거스에서 열린 세계 최대 전자제품 박람회 "CES 2025"에는 6년만에 찾아온 주인공이 있었다. 바로 엔비디아의 젠슨 황 CEO였다. 그가 키노트 연설에서 공개한 여러 정보 중에서도, 차세대 그래픽 카드인 GeForce RTX 50 시리즈는 그래픽 기술의 새로운 패러다임을 제시하며 전 세계의 이목을 집중시켰다. 젠슨 황 CEO는 이 제품에 대해 **"AI가 GeForce로 귀환했다"**라는 자신감 넘치는 선언을 했는데, 게이머는 물론 AI 업계의 비상한 관심을 끌었다.

그림 1. Blackwell 기반 RTX 50 발표, '이제는 소매 GPU에서도 AI 성능이 가장 먼저 강조됨'



자료: Nvidia, 미래에셋증권 리서치센터

특히, 중급형 버전이라고 할 수 있는 **RTX 5070이 이전 세대 최상위 모델인 RTX 4090에 버금가는 성능을 단돈 549달러에 제공한다는 주장**은 믿기가 힘들 정도로 서프라이즈였다.

그림 2. CES 2025에서 가장 깜짝 놀란 순간, "RTX 5070 성능이 RTX 4090과 동등하다고?"

자료: Nvidia, 미래에셋증권 리서치센터

이번 발표에서 엔비디아는 RTX 5090, 5080, 5070 Ti, 5070 등 총 4종의 모델을 공개했다. 각 모델의 가격은 RTX 5090 1,999달러, RTX 5080 999달러, RTX 5070 Ti 749달러, RTX 5070 549달러로 책정되었다.

또한, 엔비디아는 RTX 50 시리즈의 모바일 버전 GPU들도 대거 공개했다. AMD가 라이젠 AI Max로 이 시장을 공략하는 것과 같은 흐름이다. RTX 5090 모바일 버전은 10,496개의 쿠다 코어와 24GB의 GDDR7 메모리를 탑재했으며, RTX 5080 노트북 GPU는 7,680개의 쿠다 코어, 16GB의 GDDR7을 갖추고 있고 RTX 4080 노트북 GPU 대비 2배의 성능을 약속하고 있다. 한편 RTX 5070 제품군은 두 가지로 나뉘는데, 일반 RTX 5070(모바일)은 4,608개의 쿠다 코어와 8GB의 GDDR7 메모리를 탑재했다. 더 강력한 "형제"인 RTX 5070 Ti(모바일)는 5,888개의 쿠다 코어와 12GB의 GDDR7을 갖추고 있다.

표 2. 엔비디아 RTX 50 시리즈의 가격 등 비교

GPU 모델	가격	버전	CUDA 코어 개수	GDDR7 메모리 탑재량
RTX 5090	1,999달러 (전작 대비 400달러 인상)	데스크톱	21,760	32 GB
		모바일	10,752	16 GB
RTX 5080	999달러 (전작 대비 200달러 인하)	데스크톱	10,752	16 GB
		모바일	7,680	16 GB
RTX 5070 Ti	749달러 (전작 대비 50달러 인하)	데스크톱	8,960	16 GB
		모바일	5,888	12 GB
RTX 5070	549달러 (전작 대비 50달러 인하)	데스크톱	6,144	12 GB
		모바일	4,608	8 GB

자료: Nvidia, 미래에셋증권 리서치센터

먼저, **가격 측면에서 긍정적인 평가를 내리자면 RTX 5080의 가격이 RTX 4080 Super와 동일한 999달러로 책정된 점은 긍정적으로 평가할 만**하다. 이와 관련해, Reddit을 포함한 여러 테크/게임 커뮤니티 여론을 파악해보니 "RTX 5070과 5080이 날개 돋친 듯 팔려나갈 것"이라는 전망도 심심찮게 확인할 수 있었다. 이는 성능 대비 가격 정책이 공격적으로 보이기 때문이다. RTX 5090이 거의 2천 달러이지만, 그 밑의 모델인 5080은 '고작' 1000달러라는 것에 사람들이 안도하고 있는 듯한 여론이 많았다. 이는 5080의 잠재적인 수요가 높을 것을 시사하기도 한다.

또한, **5070이 549달러에 발표된 것은 경쟁사 AMD의 신제품 출시 전략에 영향을 미칠 것이라는 분석**이 제기되었다. 아마도 AMD도 550달러 근처에서 가격을 책정했을 가능성이 높은데, 이제 AMD가 새로운 가격 책정으로 다시 고려해야 할 것이라는 예측 또한 가능하다. 실제로, AMD는 본인들의 키노트 행사에서 엔비디아를 의식한 듯 본인들의 소매용 GPU인 RX 9070에 대한 이야기는 완전히 배제했었다.

표 3. 엔비디아 RTX 50 시리즈의 주요 스펙 비교

GPU 엔진 스펙사항	RTX 5090	RTX 5080	RTX 5070 Ti	RTX 5070
아키텍처	Blackwell	Blackwell	Blackwell	Blackwell
CUDA 코어 개수	21,760	10,752	8,960	6,144
부스트 클럭(GHz)	2.41	2.62	2.45	2.51
베이스 클럭(GHz)	2.01	2.3	2.3	2.16
기본 메모리 구성	32GB GDDR7	16GB GDDR7	16GB GDDR7	12GB GDDR7
메모리 인터페이스 너비	512-bit	256-bit	256-bit	192-bit
TGP(전력 소비량)	575W	360W	300W	250W
권장소비자가격(달러)	1,999	999	749	549
출시일	1월 30일	1월 30일	2월	2월

자료: Nvidia, 미래에셋증권 리서치센터

한편, 이번 RTX 50 시리즈 라인업에서 가장 특이할 만한 것은, RTX 5090과 그 아래 모델인 RTX 5080 간의 지나치게 큰 성능 격차다. RTX 5090은 전작인 RTX 4090과 비슷한 커팅 비율을 보여 괴물 같은 스펙을 자랑했지만 RTX 5080은 유저들이 우려했던 대로 5090의 절반 수준의 CUDA 코어 개수만이 확인되었다. 이로 인해 RTX 5080은 1,000 달러라는 가격 차이를 고려하더라도, 상대적으로 상당히 빈약한 제품처럼 느껴지기도 한다.

따라서, RTX 50 시리즈는 사실상 RTX 5090과 그 나머지 라인업으로 양분된 것과 다름없다. 과거에는 최상위 모델과 그 아래 모델 간의 성능 차이가 크지 않아, 예산에 맞춰 유저들이 선택을 할 수 있었다. 하지만, RTX 50 시리즈에서는 이러한 선택지를 선택할 수 있는 헤비유저에게만 해당되는 셈이다. 따라서 "가격이 얼마든 상관없다"는 사람들의 수요를 정조준하기 위해 엔비디아가 내놓은 가격 정책이라고 할 수 있겠다. 이러한 **엔비디아의 의도적인 라인업 양극화는 사실 '대안'이 없기 때문이다. 하이엔드 GPU 시장에 대한 AMD의 빈자리는, 엔비디아에게 '가격 결정권'을 쥐여준 것과 다름없다.**

(2) AI 기반 그래픽 혁명: DLSS 4 공개

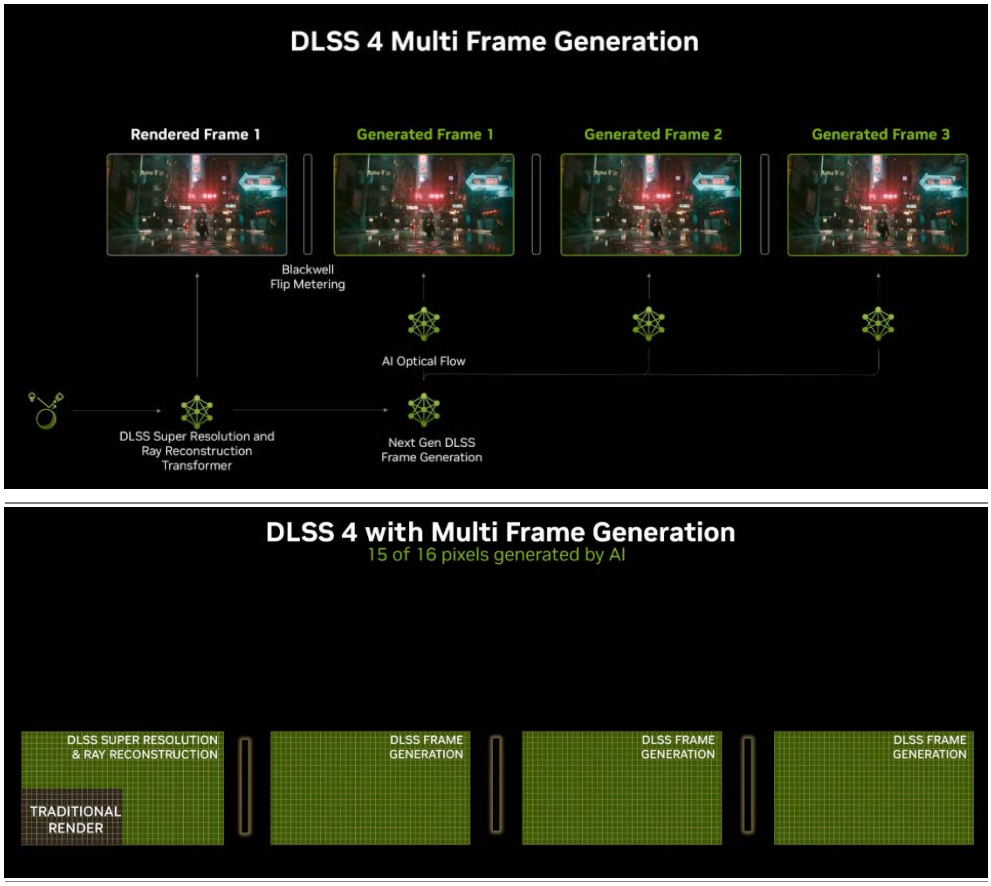
엔비디아는 이번 CES에서 AI 업스케일링 기술의 최신 버전인 DLSS(Deep Learning Super Sampling) 4를 공개하며, 게이밍 그래픽의 새로운 기준을 제시했다. DLSS는 엔비디아가 개발한 인공지능 기반의 이미지 업스케일링 기술이다. DLSS는 저해상도 이미지를 고해상도로 변환하거나, 존재하지 않는 프레임을 AI가 생성하여 게임의 프레임률을 높이는 것을 골자로 한다.

엔비디아는 2018년 9월, GeForce RTX 20 시리즈를 통해 AI라는 용어를 PC 게이밍 시장에 처음 도입한 이래 6년이 넘는 기간 동안 이 AI 기술인 DLSS의 혁신을 거듭해왔다. 그리고 이제, 그래픽 처리 전반에 걸쳐 AI 알고리즘이 핵심적인 역할을 하게 되었다. 다시 말해, **DLSS는 엔비디아 소매용 그래픽 카드의 핵심 경쟁력의 일등공신이 되고 있다**.

그리고 이번 **RTX 50 시리즈와 함께 출시된 "DLSS 4"는 기존의 프레임 생성 기술을 더욱 고도화했다. 이전 세대인 DLSS 3가 렌더링 프레임 한 장당 AI 추가 프레임을 한 장 생성하여 두 배의 프레임을 생성했다면, DLSS 4는 한 장당 세 장의 AI 추가 프레임을 생성하여 프레임을 무려 네 배까지 늘릴 수 있다.** 즉, 하나의 실제 프레임을 렌더링하고, 그 사이에 최대 세 개의 가상 프레임을 끼워 넣는 방식이라는 말이다.

이는 DLSS 4에서 새롭게 탑재된 '멀티 프레임 생성(MFG)' 기능 덕분인데, 기존 렌더링 방식 대비 최대 8배까지 프레임률을 향상시킬 수 있다고 홍보되고 있다.

그림 3. 엔비디아가 이번에 새롭게 발표한 DLSS 4에 탑재된 핵심 신기술 "MFG"



자료: Nvidia, 미래에셋증권 리서치센터

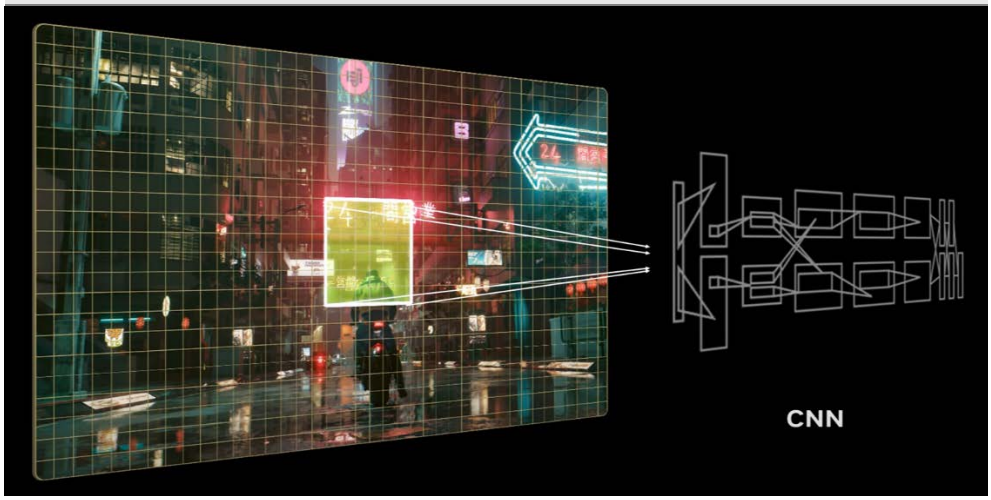
엔비디아 측은 "GeForce RTX 5090 그래픽카드에서의 이러한 거대한 성능 향상으로 4K 화질, 240fps(초당 프레임) 환경으로도, 완전한 레이트레이싱 게이밍이 가능해졌다"고 밝혔다. 믿기 힘들 정도로 어마어마한 게임 성능이다. 실시간 레이트레이싱을 켜둔 채로 4K 해상도에 240 프레임이라는 설정은 게임 방송을 주업으로 삼는 전문 스트리머들도 거의 하지 않는 고급 옵션이기 때문이다.

그렇다면, DLSS 4는 구체적으로 어떤 기술적 혁신을 통해 이러한 놀라운 성능 향상을 이뤄낸 것일까? 여러가지 혁신이 담겨 있지만, 우리 팀은 '**DLSS 4에 생성 AI의 주축 아키텍처인 트랜스포머를 접목한 것'에 주목한다**.

DLSS에서 기존에 사용하던 CNN(Convolutional Neural Network)은 이미지를 마치 조각난 퍼즐처럼, 국소적인 부분에 집중하여 처리하는 방식이었다. 그러니까 주변의 픽셀 정보를 각각의 블록으로 나누어 AI가 인식하면서, 이를 기반으로 다음 프레임을 예측하는 방식이다. 하지만, CNN 방식에서는 전체적인 움직임을 제대로 예측하지 못해 프레임 간 일관성이 떨어지는 문제가 있었다. (예: 화면 깜빡임과 고스팅 현상)

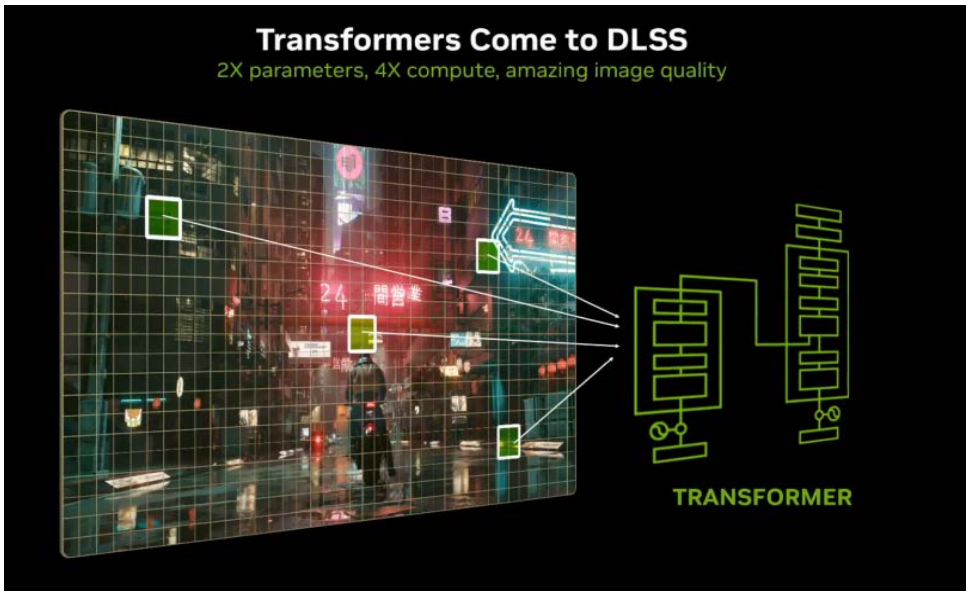
반면, 트랜스포머는 마치 거대한 캔버스 전체를 한눈에 꿰뚫어 보는 화가의 시선처럼, 이미지 전체를 조망하며 픽셀 간의 유기적인 관계를 포착한다. 이를 통해, **트랜스포머는 CNN의 한계를 넘어, 더욱 정교하고 사실적인 이미지 재구성을 가능케 한다**. 특히, 레이 트레이싱과 같이 빛과 그림자가 복잡하게 얽힌 장면에서 더 성능 발휘는 극대화된다.

그림 4. 엔비디아의 DLSS 알고리즘에는 원래 CNN 알고리즘이 쓰였었다



자료: Nvidia, 미래에셋증권 리서치센터

그림 5. "AI가 GeForce(엔비디아 소매용 GPU)에 왔다"는 말의 뜻… 트랜스포머가 여기에도 적용



자료: Nvidia, 미래에셋증권 리서치센터

그림 6. DLSS의 기반 기술이 CNN에서 트랜스포머 아키텍처로 바뀌자 업스케일링의 화질 성능 향상

자료: Nvidia, 미래에셋증권 리서치센터

사실상 **본질적으로 서로 다른 두 AI 업스케일러를 비교하는 것과 같을 정도로 큰 변화다.** 이와 같은 성능 향상은, 이전 CNN 모델보다 비전 트랜스포머 모델이 두 배 더 많은 매개변수를 처리할 수 있게 되었기 때문이다. 우리가 ChatGPT나 Gemini의 놀라운 성능에서 느꼈듯이 트랜스포머의 맥락 이해는 CNN과는 차원이 다르다.

또한 트랜스포머는 구글의 AlphaFold와 OpenAI의 Sora, 테슬라의 FSD에서 보듯이 그 확장성이 매우 뛰어나다. 따라서 **엔비디아에서도 트랜스포머를 기반으로 한 DLSS가 앞으로도 학습을 진전시켜 그 품질을 계속 더욱 향상시킬 것임을 강조했다**.

사실, 지금의 AI 혁명은 "트랜스포머의 시대"라고 해도 전혀 과언이 아니다. 그리고 그것이 드디어 컴퓨터 그래픽 시장에 도래한 것이라고 쉽게 이해하면 된다. **트랜스포머의 유행에 있어 대대적인 GPU 수요라는 수혜를 맛보는 중인 엔비디아가 본인들의 솔루션에도 트랜스포머를 쓴 것**은 어찌보면 자연스러운 일이기도 하다. 애초에 엔비디아의 텐서코어와 HBM이라는 구조가 트랜스포머 아키텍처에 안성맞춤이 되도록 고안된 구조이기 때문이다.

(3) RTX 5070 = RTX 4090? 그래픽 카드 성능의 새로운 패러다임!

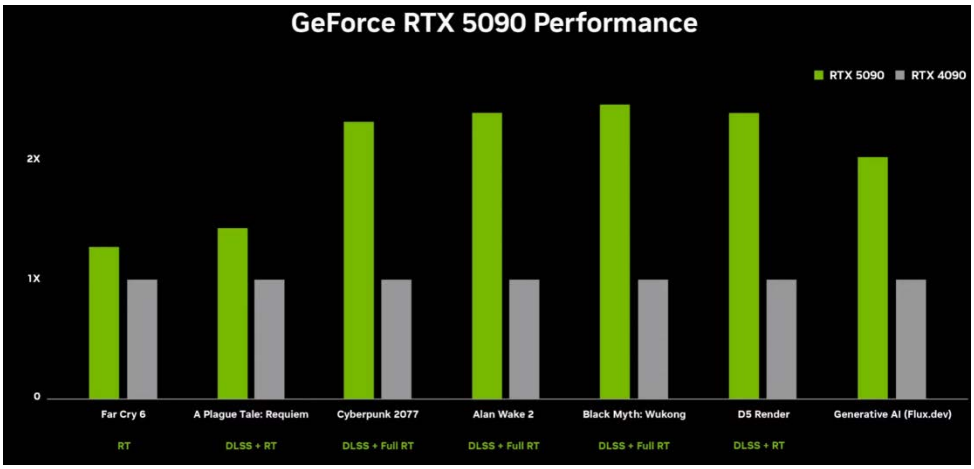
표 4. RTX 5070과 RTX 4090 스펙 비교, '하드웨어 상으로는 훨씬 못 미치는데 게임 성능이 어떻게 비슷할까? 답은 DLSS 4'

특징	RTX 5090	RTX 5080	RTX 5070 Ti	RTX 5070	RTX 4090
아키텍처	Blackwell	Blackwell	Blackwell	Blackwell	Ada Lovelace
DLSS	DLSS 4	DLSS 4	DLSS 4	DLSS 4	DLSS 3
멀티프레임 생성(MFG) 지원	O	O	O	O	X
텐서 코어 (AI) 성능(단위: TOPs)	3,352	1,801	1,406	988	1,321
텐서 코어 세대	5세대	5세대	5세대	5세대	4세대
RT 코어 성능(단위: TFLOPs)	318	171	133	94	191
RT 코어 세대	4세대	4세대	4세대	4세대	3세대
메모리 구성	32GB GDDR7	16GB GDDR7	16GB GDDR7	12GB GDDR7	24GB GDDR6X
메모리 대역폭	1792 GB/초	960 GB/초	896 GB/초	672 GB/초	1008 GB/초

자료: Nvidia, 미래에셋증권 리서치센터

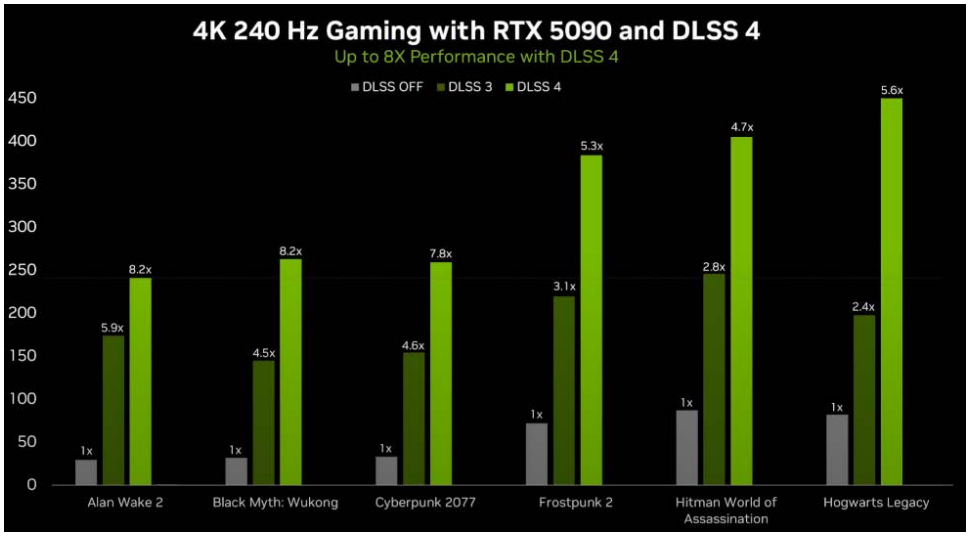
젠슨 황 CEO가 “RTX 5070이 단돈 549달러에 RTX 4090급 성능을 제공한다”고 외칠 때 많은 유저들은 입을 쩍 벌리며 놀랐을 것이다. 하지만, RTX 5070이 RTX 4090급 성능을 내는 것은 오로지 DLSS 4의 '멀티 프레임 생성(MFG) 기능을 사용했을 때만 그렇다는 것이 게임 업계의 맹점으로 떠올랐다.

그림 7. RTX 5090과 전작인 RTX 4090의 각 게임별 성능 비교, ‘DLSS 없으면 성능 향상폭은 적음’



자료: Nvidia, 미래에셋증권 리서치센터

그림 8. RTX 5090의 4K 240Hz 환경에서의 각 게임 성능, ‘DLSS 4가 있고 없고는 엄청난 차이’



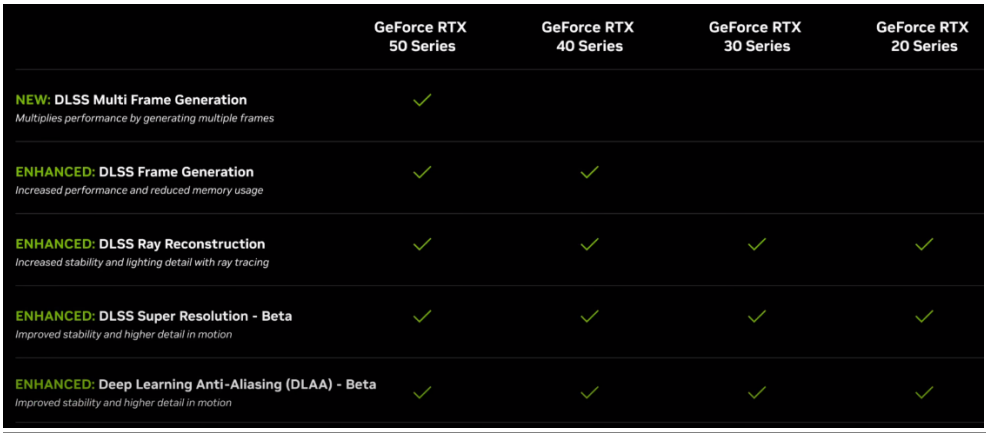
자료: Nvidia, 미래에셋증권 리서치센터

DLSS 4가 아닌 DLSS 3만 지원하는 게임으로 RTX 50 시리즈 GPU를 테스트했을 경우(즉, '멀티 프레임 생성' 기능 없이 비교할 경우), RTX 5090의 성능 향상은 RTX 4090 대비 50% 미만으로 엔비디아의 주장과는 큰 차이를 보이기도 한다. 다시 말해, 새로운 버전의 AI 기능 없이는 불가능한 주장이라는 말이다.

그래서 엔비디아가 공개한 성능 그래프들을 보면서, DLSS 4를 활용한 “프레임 뻥튀기 결과”라면서 냉소적인 유저 반응도 꽤 있었다. 심지어, 소위 “깡성능”이라고 불리는 픽셀 쉐이더 레스터 성능은 역대 최저 수준이라는 우려까지 있을 정도다.

** 깡성능이란? 그래픽 처리 파이프라인의 후반부 단계를 레스터라이제이션이라고 부름. 이 단계에서는 버텍스 쉐이더로부터 받은 3D 객체의 좌표 정보를 받아 모니터 화면에 그려질 각 픽셀의 최종 색상과 밝기를 계산.*

그림 9. DLSS의 각 기능과 이를 지원하는 GPU 라인업, '새로운 MFG 기능은 RTX 50만 지원'



자료: Nvidia, 미래에셋증권 리서치센터

그럼에도, 엔비디아의 주장을 굳이 평가 절하할 필요는 없다. 영국의 비디오 게임 전문 웹진인 PCGamesN의 CES 2025 현장 보도는 DLSS 4에 대해 호의적인 평가를 내렸다. **실제로 MFG를 켜놓은 상태의 RTX 5070이 게임 성능에 있어서, RTX 4090보다 훨씬 높은 프레임률을 기록했고, 전반적으로 비슷한 성능을 보였다**고 보도한 것이다. 또한, 유럽의 저명한 게임 전문 매체인 유로게이머에서도 RTX 5080에 탑재된 DLSS 4 기술을 직접 체험한 후, "사이버펑크 2077 게임을 풀 패스트레이싱으로도 120fps를 훌쩍 넘기며 더욱 생생하게 경험할 수 있게 되었다"고 평가하며, 이 기술에 대한 극찬을 아끼지 않았다.

사실, 게이머들이 "실상은 멀티 프레임 생성이라는 눈속임 기술에 의존한 결과일 뿐"이라고 엔비디아의 주장을 절하하는 것은 이제 큰 의미가 없다. 경쟁자인 AMD도 본인들의 머신러닝 기반 업스케일링 기술인 FSR에 사력을 다 하고 있다. 엔비디아의 DLSS나 AMD의 FSR 덕분에 실제로 게임 체감 성능이 올라간 것은 사실이기도 하다.

우리 팀은 **엔비디아가 강조하는 소매용 GPU의 성능 제시 방법이 새로운 패러다임이라고 생각한다. GPU의 순수 쉐이더 성능만을 바라보던 하드웨어 관점의 시대가 저물고, "딥러닝 슈퍼 샘플링"으로 대표되는 AI 소프트웨어 관점의 시대가 온 것이다**. 다시 말해, 이제 순수한 소위 "깡성능(레스터 성능)"을 비교하는 낭만의 시대는 지났다고 본다.

엔비디아 RTX 50 시리즈와 DLSS 4는 새로운 패러다임의 시작을 알리는 신호탄이다. 이제 그래픽 기술은 큰 셰이더 코어를 가진 거대 GPU보다는, AI를 통한 프레임 생성과 레이 트레이싱 기능에 집중하는 방식으로 발전할 것으로 예상된다. 그리고 이는 **게임 개발의 새로운 표준으로 자리 잡을 가능성이 높고, 게임 산업뿐만 아니라 영화, 디자인, 의료 등 다양한 분야에 혁신적인 변화를 가져올 것으로 예상된다**.

그리고 이 시장의 룰을 정하는 것은 엔비디아이기 때문에 경쟁자들은 이를 따라야 할 것이다. 무엇보다 최신 반도체 제조 공정 비용이 계속 오르고 있기 때문에, 전통적인 방식으로 '더 큰 성능'을 원하면 그만큼 더 높은 비용을 지불해야만 한다. 이처럼 애초에 GPU의 순수 연산 성능만으로는 시장의 요구를 충족시키기 어렵기도 하거니와, **무엇보다 AI 기술을 활용하는 것이 기업의 이익을 극대화하는 방향일 수 있다.**

이전에도 엔비디아가 선도한 "레이 트레이싱" 기술은 처음에 게임 커뮤니티의 비판을 받았지만, 결국에는 게임 개발자들의 핵심 키워드로 계속해서 남아 있다. 이번에도 마찬가지일 것으로 사료된다. 즉, 그래픽 카드 성능을 따질 때 AI 성능을 따지게 되는 문화는 현재의 주류로 자리 잡아 가는 상황이고, 근래에는 더욱 그럴 것으로 사료된다. 그리고 **그런 분위기 변화 속 가장 수혜를 입을 것은 하드웨어 기업이자 소프트웨어 기업인 엔비디아다.**

저명한 게임 전문 매체인 "PC Gamer"의 최근 기사 내용을 공유하며 글을 마친다.

"AI 기반 프레임 생성이 미래의 길이 될 것이며, 이는 좋은 일이라는 걸 인정할 때다.

엔비디아 RTX 50 시리즈와 개발자 키트는 래스터화가 구시대 기술이고,
우린 이제 AI 렌더링 시대에 확실히 진입했음을 보여준다.

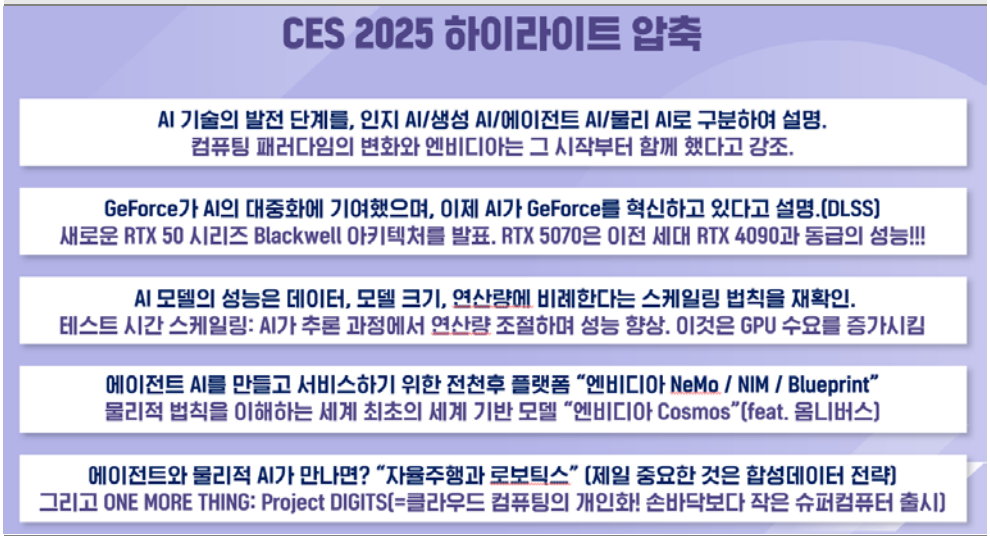
AI와 레이 트레이싱을 사용한 게임 렌더링이 시각적으로 완벽한 수준에 도달하고
(유저들이) 받아들일 수 있는 성능으로 구동되는 시점이 가까워지고 있다"

2. CES 2025 젠슨 황 키노트

(1) 키노트 요약

RTX 50 시리즈가 젠슨 황 키노트의 백미였지만, 사실 훨씬 더 거대하고 근본적이며 대담한 변혁에 대한 것은, RTX 50 발표 이후의 이야기로 쭉 이어졌다. 그의 발언을 순차적으로 간략하게 리뷰해볼 필요가 있어 이번주 자료에서 정리를 해보려 한다.

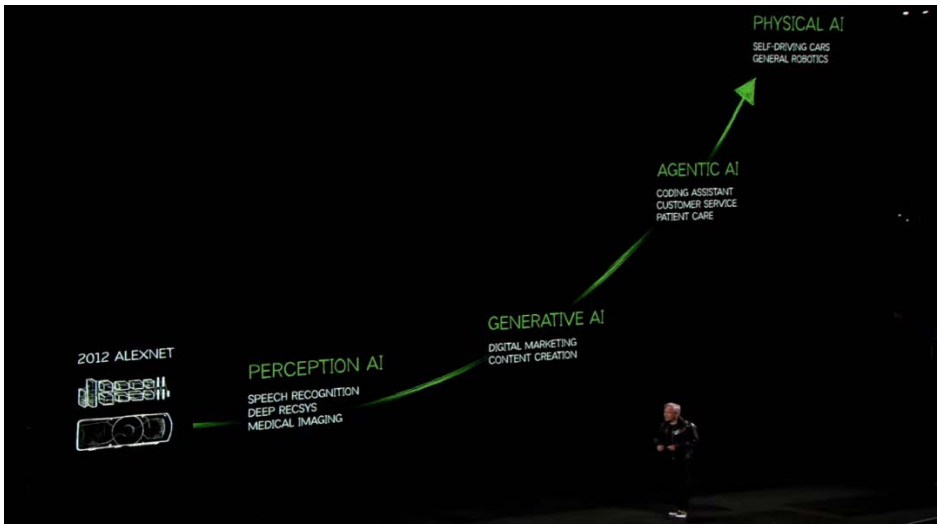
그림 10. CES 2025에서의 젠슨 황 CEO의 키노트 내용을 압축



자료: Nvidia, 미래에셋증권 리서치센터

그는, 엔비디아가 쌓아온 지난 30년의 길이 곧 "그래픽과 AI가 만나는 여정이었다"고 회고하며 운을 뗐다. 1993년에 등장한 NV1부터 프로그래머블 GPU, CUDA, 그리고 2012년의 AlexNet이 보여준 AI 혁명까지, 엔비디아가 걸어온 길이 곧 기술 패러다임의 변화와 함께 해왔다는 말이었다. **AI native 기업으로서 애초에 준비된 기업이 엔비디아라는 뜻이다**.

그림 11. 젠슨 황 'AI의 발전 단계는 4단계. 이제는 에이전트 AI와 물리적 AI로 가는 길'



자료: Nvidia, 미래에셋증권 리서치센터

먼저 그의 말 중 흥미로웠던 부분은, **AI의 발전 단계에 대해서 정리한 대목이다. '인지 AI'와 '생성 AI'를 넘어, 이제 '에이전트 AI'와 '물리 AI' 분야에까지 빠르게 발전하고 있다면서, 컴퓨팅의 모든 계층이 변혁의 시대를 맞고 있다는 점을 재차 강조했다**. 그리고 젠슨 황은 AI 개발자들 사이에서 항상 많이 회자되는 '스케일링 법칙'에 대해서도 자세히 언급했다. 스케일링 법칙은 "데이터/모델 파라미터/연산량이 늘어날수록 모델 성능이 기하급수적으로 좋아진다"는 법칙이다.

여기서, 흥미로운 부분은 '**테스트 시간 스케일링(test-time scaling)'이라는 개념을 젠슨 황이 직접 언급한 것**이다. 이 개념은 OpenAI가 o1 모델을 내놓으면서 새롭게 부상한 개념이다. AI가 답변을 내놓는, 즉 "추론 시점(test time)"에 AI가 스스로 "얼마나 깊게 계산할 것인지" 동적으로 판단하는 기술이다. 간단한 질문에는 빠르게 대답하고, 복잡한 문제는 더 많은 토큰을 쓰면서 단계별로 고민한다는 방향성이다. 그리고 그는 이러한 스케일링의 영역 확장은 **결국, 엔비디아의 컴퓨팅 수요를 더욱 높이고 있으며, Blackwell 아키텍처가 그 요구를 충족할 것이라는 예측도 덧붙였다**.

GTC 2024에서 처음 소개된 Blackwell을 기반으로 하는 대형 시스템인 GB200 NVL72가 그래서 다시 강조되었다. GB200 NVL72는 이전 세대 대비 전성비가 4배, 비용당 성능이 3배 개선되어 1.4 ExaFLOPs이 넘는 연산 성능을 제공한다는 점도 공개되었다. 전 세계 슈퍼컴퓨터가 도합해야 겨우 1ExaFLOPs를 넘기던 시대와 비교하면, 말 그대로 '한 캐비닛 안에 들어가는 슈퍼컴퓨터'로 생각되기도 한다.

그런데, 엔비디아가 단순히 실리콘 레벨이 아닌 NVL72라는 시스템 레벨로써 강조한 이유는, 결국 "테스트 타임 스케일링"이 향후에 주류가 될 것임을 이미 알고 있었기 때문이었고 이를 지원하기 위해서 이런 제품이 나왔다고 우리 팀은 이해하고 있다. (이에 대한 힌트는 이미 AI Weekly에서 자주 다룬 바 있다)

Blackwell 시스템은 전 세계 45개 공장에서 생산 중이라면서 "**Blackwell in full production"이라고 직접 장표까지 띄워 젠슨 황이 발표**하는 내용은 소소한 웃음 포인트이기도 했다. Blackwell 생산 지연 소문에 대해서 이제 그만 의심해달라는 뉘앙스로 비쳐졌기 때문이다.

생성 AI에 대한 챕터를 끝내자, 그는 곧바로 에이전트 AI에 대한 이야기를 이어 나갔다. 젠슨 황이 정의한 에이전트는 "스스로 추론하고, 여러 모델 및 툴을 결합해 문제를 단계별로 해결하는 AI"다. 이를 위한 인프라로 엔비디아는 NeMo 프레임워크와 NIM(Nvidia Inference Microservices)을 제시했다. 사실 이것은 GTC 2024에서 이미 소개된 개념인데, 이번에는 **새롭게 "Agent Blueprint"라는 것이 소개되었다. 이것은 다양한 업무 및 산업 분야에 특화된 AI 에이전트 예시를 엔비디아가 레퍼런스 형태로 제공하는 것이다**.

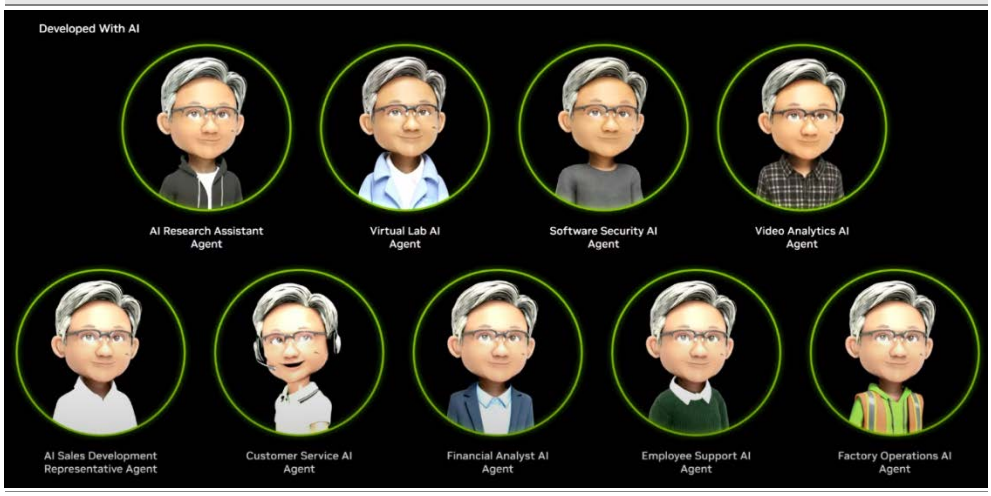
그러니까 쉽게 말해, **NeMo, NIM, Agent Blueprint 모두 AI 에이전트와 관련된 것이다. Agent Blueprint로 AI 에이전트의 기본 틀을 만들고, Nemo로 훈련하고 관리하며, NIM으로 실제 추론 기능으로 실행**하는 형태가 되는 것이다.

표 5. 에이전트 AI 시대를 위해 엔비디아가 제시한 세가지 레이어

구분	NeMo (AI 에이전트의 "육아" 과정)	NIM (AI 에이전트의 실력 발휘 방법)	Agent Blueprint (AI 에이전트 개발 지침)
기능	- AI 에이전트 훈련 (Training) - AI 에이전트 온보딩 (Onboarding) - AI 에이전트 평가 (Evaluation) - AI 에이전트 배포 (Deployment) - AI 에이전트 가이드레일 설정 (Guardrails)	- AI 모델 배포 (Deployment) - 추론 최적화 (Inference Optimization) - 다양한 모델 지원 (Model Support) - 클라우드 및 온프레미스 환경 지원	- AI 에이전트 템플릿 제공 - 오픈 소스 제공 - 맞춤화 지원 (Customization) - 모듈화 지원 - 개발 가이드 제공
의의	- AI 에이전트 개발 생태계의 기반 - 엔터프라이즈 AI 에이전트 개발 가속화 - AI 에이전트의 라이프사이클 관리	- AI 모델의 실용성 극대화 (Practicality) - AI 추론의 효율성 향상 (Efficiency) - AI 에이전트의 실질적 가치 실현 (Realization)	- AI 에이전트 개발 진입 장벽 해소 - AI 에이전트 개발 기간 단축 - AI 에이전트 생태계 확장
상호 관계	AI 에이전트의 생성, 훈련, 관리, 배포를 총괄	훈련된 AI 모델을 실제 환경에다가, 배포하고 실행하도록 지원	AI 에이전트의 초기 설계 및 개발 가이드 제공

자료: 미래에셋증권 리서치센터

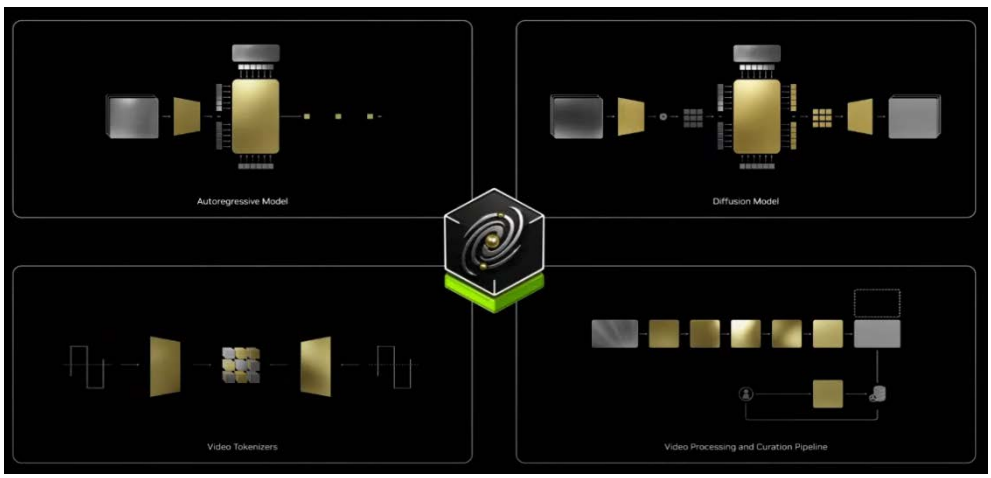
그림 12. 엔비디아가 내놓은 Agent Blueprint, '그냥 이거 가져다 써서 에이전트 편하게 쓰라는 것'



자료: Nvidia, 미래에셋증권 리서치센터

에이전트 AI, 그 이후의 비전도 제시했다. 현실 세계의 물리 법칙을 이해·시뮬레이션할 수 있는 'Physical AI(물리적 AI)'가 그 주인공이었다. **엔비디아는 "세계 최초의 물리적 AI 모델"이라고 내세운 Cosmos라는 새로운 기반 모델을 내놓았다**.

그림 13. 엔비디아가 내놓은 Cosmos, '트랜스포머와 디퓨전 모델 기반, 비디오 처리에 특화'

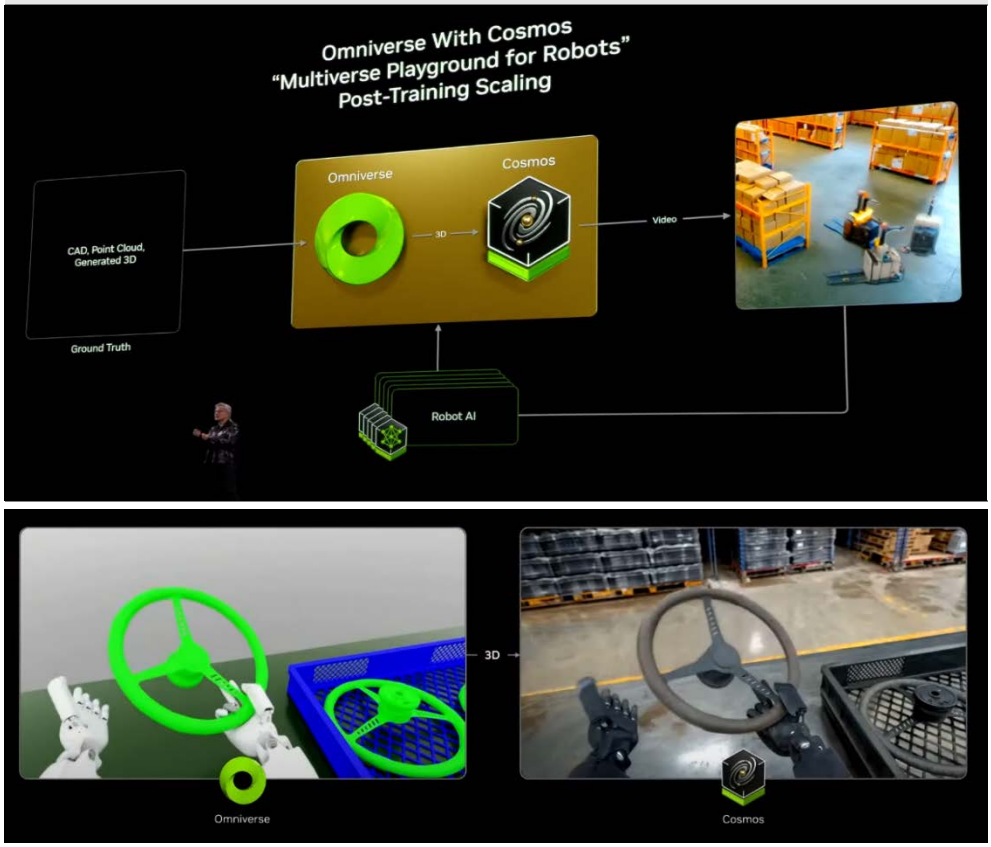


자료: Nvidia, 미래에셋증권 리서치센터

Cosmos는 무려 2천만 시간 분량의 비디오를 학습한 모델이고, 이를 통해 **실제 환경에 가까운 가상 세계를 생성하고 여러 물리 현상을 재현할 수 있다**고 그는 강조했다. 그리고 비디오 데이터뿐만 아니라, Omniverse 플랫폼으로부터 얻은 3D asset 데이터까지 Cosmos에 결합되어 있다는 점도 매우 주목할 만하다.

로보틱스나 산업 현장에서 좀 더 정확한 시뮬레이션 및 합성 데이터를 얻을 수 있기 때문이다. Cosmos는 오픈소스 라이선스로 세가지 모델이 공개된다고 한다. 따라서, 로봇 공학이나 자율주행, 제조업 등 다양한 분야에서 새롭게 시도되는 응용 사례가 곧 늘어날 것으로 예상된다.

그림 14. Cosmos와 Omniverse가 만나서, 'AI 모델은 이 우주가 3차원이라는 것을 이해하게 된다'



자료: Nvidia, 미래에셋증권 리서치센터

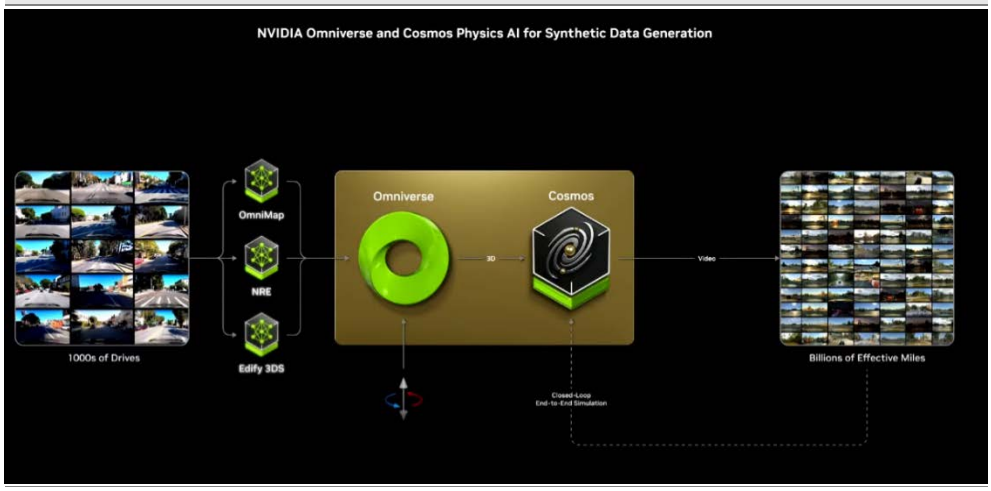
물리적 AI가 구축되면 가장 효과적으로 쓰일 영역으로 젠슨 황은 3가지를 짚었다. 첫째는 디지털 트윈이다. 즉, 제조업을 디지털화하여 생산성을 극대화하는 것이다. 엔비디아에 따르면 창고·물류센터 시장의 규모는 무려 1조 달러 규모다. 엔비디아는 세계 최대 규모의 물류자동화 업체인 Kion과 세계 최대 컨설팅 업체인 Accenture와의 파트너십을 통한, 산업·물류 자동화 시나리오를 소개했다. 공장, 창고, 물류센터를 디지털 트윈으로 구현한 뒤, 해당 공간에서 무수히 많은 시뮬레이션을 돌려 최적의 운영 방안을 찾아내는 방식이다.

한편, 자율주행차 분야에서는 새 프로세서 'Thor'가 공개됐고, 기존의 Orin 대비 20배 성능이 향상되었다고 주장했다. **젠슨 황 CEO는 "자율주행은 아마도 수조 달러 규모가 될 첫 번째 로보틱스 산업일 것이다. 엔비디아는 이 분야에서 40억 달러 규모의 비즈니스를 하고 있다. 올해는 50억 달러 수준의 run rate가 예상된다"고 말했다**. 또한, "이미 전 세계 거의

모든 자동차 회사가 엔비디아와 협력하고 있다"면서, **최근에는 Toyota도 엔비디아와 함께 다음 세대 자율주행 시스템을 만들기로 발표한데서 청중들의 박수 소리가 커지기도 했다**.

그리고, 자율주행 영역에도 Omniverse와 Cosmos가 적극적으로 활용되게 된다는 점도 중요하다. 젠슨 황은 물리적 AI 구축의 어려움으로 '데이터 확보의 어려움'을 들었는데, 이를 극복할 수 있는 '합성 데이터 생성'이라는 전략을 이 둘의 조합으로 가능하다는 게 골자다. 이에 대해 젠슨 황은 "수천 번의 주행 데이터만 가지고 수십 억 마일로 확장할 수 있다. 자율주행차에게 필요한 방대한 학습 데이터를 안정적으로 공급할 수 있다. **물론 실제 주행도 계속해야 하지만, 합성 데이터를 통해 엄청난 학습 기회를 얻을 수 있다"고 밝혔다.**

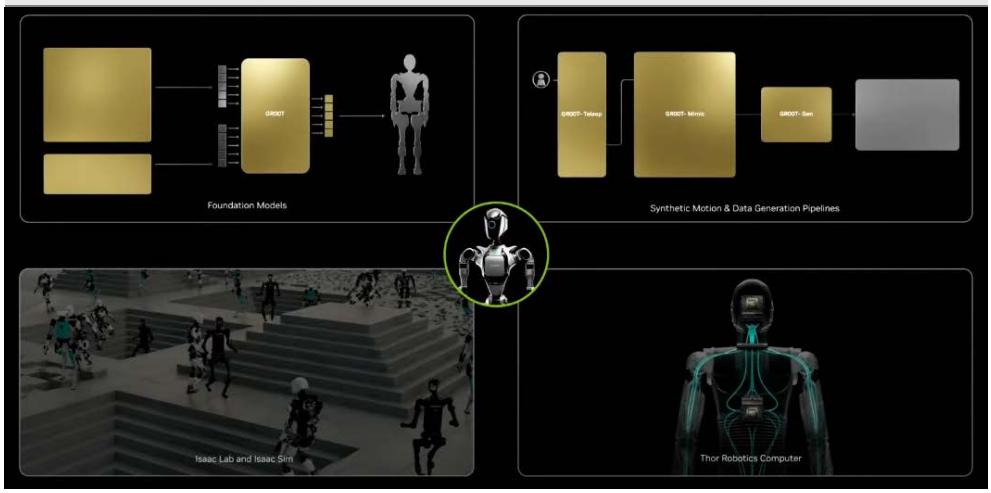
그림 15. 실제 주행 영상을 Omniverse로 3D로 재구성하고, Cosmos에게 주면 '합성 데이터' 생성



자료: Nvidia, 미래에셋증권 리서치센터

물리적 AI의 궁극적 지향점은 로봇이다. 에이전트 로봇, 자율주행차, 휴머노이드 로봇이라는 '3가지 로봇'이 거대한 변화를 맞을 것이라 언급했다. 특히 엔비디아가 Isaac GR00T라는 툴체인을 통해 합성 데이터 기반 로봇 훈련을 가속화하겠다는 계획을 밝혔다. 범용 휴머노이드 모델을 개발하려면 엄청난 양의 실제 데이터가 필요하지만, 이를 수집하고 정제하는 데는 많은 비용이 들기 때문에 **GR00T라는 것으로 필요한 것을 다 제공해주겠다는 뜻이다**.

그림 16. GR00T: 로봇을 위한 기반 AI 모델 + 합성 데이터 생성 + 강화학습 환경 + 온디바이스 칩



자료: Nvidia, 미래에셋증권 리서치센터

그리고 젠슨 황은 놀랍게도 여기서 끝내지 않았다. 마치 **스티브 잡스의 재림과 같이 그는 "ONE MORE THING"을 말했다. 여기서 나온 것이 바로 '프로젝트 DIGITS'라는 이름의 새 AI 슈퍼컴퓨터였다.** 이것은 손바닥 만한 크기의 컴퓨터로 마치 애플의 Mac mini를 연상시키는 듯했다. 그런데 그 성능은 애플의 M4 시리즈와 비교하기 미안할 정도였다. **DIGITS의 AI 성능은 (FP4 기준이긴 하지만) 무려 페타플롭스이며, 경쟁이 없는 수준이다**.

DIGITS는 엔비디아의 DGX 클라우드가 제공하는 모든 AI 소프트웨어를 구동할 수 있는 시스템으로써, 즉, 개인용 클라우드 컴퓨팅 플랫폼처럼 활용할 수 있다. 여기에 탑재된 GB10 칩은 MediaTek과 협력해 만든 특별한 버전의 CPU가 접목된다고 했다.

2025년 5월 출시 예정이며, 거기다 가격은 3,000달러에 불과하다. 따라서, 이 장치를 여러 대 연결해 DGX 스타일의 클러스터로 쓰는 수요층이 꽤 많으리라 짐작해볼 수 있다. 집에서 로컬로 AI를 훈련해보려는 개발자들, 소형 AI 연구소/스타트업/정부기관 등이 주요 타겟층으로 자리잡을 수 있다고 본다. GPU가 너무 비싸 가속컴퓨팅으로의 전환에 대해서 회의론을 가진 사람들에 대한 엔비디아의 대답은, 바로 이것인 듯하다.

'GPU 기반 AI 공장의 구축비용이 비싸다고? 아예 책상 위의 AI 공장 시대를 열겠다."

그림 17. Project DIGITS: 엔비디아의 AI 서버인 "DGX"를 소형화/개인화한 신제품

자료: Nvidia, 미래에셋증권 리서치센터

(2) 지난 GTC 2024와 이번 CES 2025를 비교

GTC 2024가 엔비디아 여정의 시작이었다면, CES 2025는 그 여정의 한 이정표를 세운 것이라 평가하고 싶다. GTC 2024를 돌이켜보면, 엔비디아는 Blackwell, NIM, Omniverse라는 세 개의 주요포인트를 제시했었다. 그리고 CES 2025에서는 그 기술들이 RTX 50 시리즈, 엔비디아 에이전틱 AI, 그리고 Cosmos라는 구체적 결과물로 이어졌음을 보여주었다.

Blackwell의 확장: 데이터센터에서 게이밍까지

GTC 2024에서 첫 선을 보인 Blackwell은, 이번 CES 2025에서 Blackwell은 게이밍 GPU인 RTX 50 시리즈에 적용되며 그 활용 범위를 넓혔다. RTX 5070이 549달러라는 가격에 RTX 4090급 성능을 제공한다는 것은 Blackwell의 성능과 전력 효율성이 시장에 미칠 파급력을 보여준다. 이는 "AI 팩토리"의 표준화를 추구했던 GTC 2024의 전략과 연결되며, 엔비디아가 Blackwell을 통해 전방위적인 시장 지배력을 강화하려는 야심을 드러낸다.

NIM의 진화: AI 모델 배포를 넘어 에이전틱 AI 개발 플랫폼으로!

GTC 2024에서 첫 선을 보인 NIM은 기업이 AI 모델을 쉽게 배포하고 운영할 수 있도록 돕는 마이크로서비스 플랫폼이다. CES 2025에서 NIM은 한 단계 더 진화했다. 단순한 모델 배포의 목적을 넘어, 에이전틱 AI 개발을 위한 핵심 인프라로 자리매김하려는 전략으로 해석된다. 특히, '엔비디아 Llama Nemotron Language Foundation Model'을 공개한 것은, 오픈소스 생태계와의 협력을 강화하고, NIM의 활용도를 높이려는 의도로 풀이된다.

Omniverse의 도약: 디지털 트윈을 넘어, 현실과 가상을 잇는 플랫폼으로

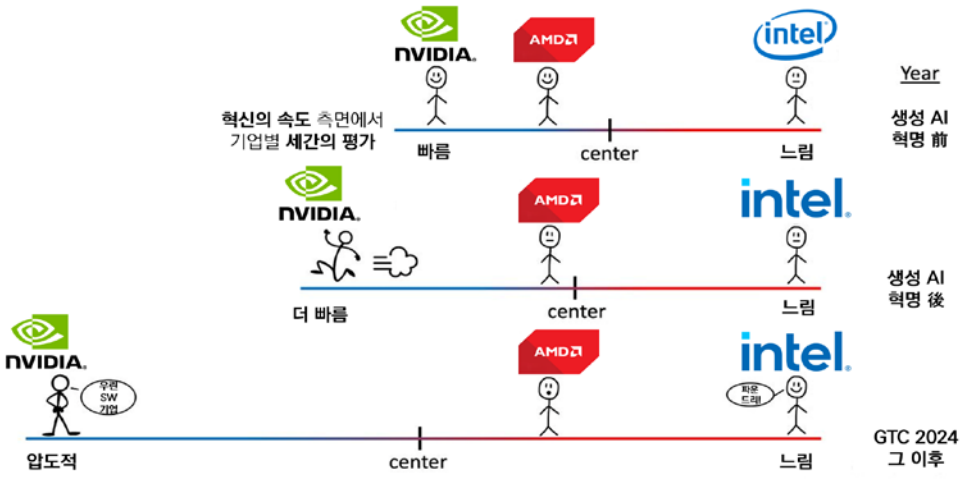
GTC 2024에서 주목받았던 Omniverse는 현실 세계를 가상 공간에 그대로 재현하는 '디지털 트윈'을 구축하는 플랫폼이다. CES 2025에서 Omniverse는 새로운 모델인 ' Cosmos'와 결합되어, 물리적 세계를 이해하고 시뮬레이션하는 AI 개발 플랫폼으로 도약했다. 또한, Cosmos의 오픈소스 형식의 공개는 Omniverse 생태계 확장과 AI 개발 접근성 향상을 위한 전략일 것이다. 그리고 "물리적 AI"라는 개념은, AI가 현실 세계와 맺는 관계의 패러다임을 재정립하려는 시도로써 'AI의 끝은 결국 메타버스'라는 우리 팀의 생각을 강화해줬다.

GTC 2024 때 느꼈던 의문 및 우려 사항의 해소

GTC 2024에서 언급되었던 CUDA 플랫폼의 지배력, 데이터센터 수준의 혁신, 소프트웨어 중심 전략, 시뮬레이션/메타버스의 중요성은 CES 2025에서 더욱 강화됐다.

먼저, GTC 2024가 열릴 당시만 하더라도, AMD와의 경쟁 심화 가능성이 제기됐었다. 하지만, CES 2025에서 엔비디아는 RTX 50 시리즈의 공격적 가격 정책과 Blackwell의 성능 우위, 그리고 압도적인 SW 경쟁력으로 경쟁 우위를 더욱 공고히 했다고 본다. RTX 50 시리즈에 청중은 열광했고, 일각에서는 경쟁사인 AMD의 소극적인 자세의 소매용 GPU의 미래에 관한 애도의 목소리가 나올 정도였다.

그림 18. 우리 팀이 "GTC 2024 탐방기" 자료에서 예상했던 것, 'AI 역량으로 회사간 차이는 크다'



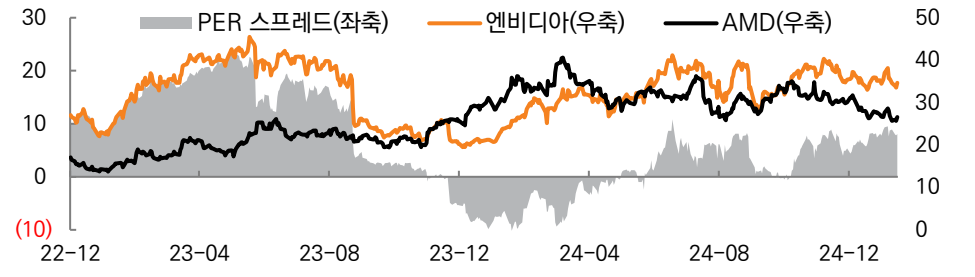
자료: X(@SwipeWright), Nvidia, AMD, Intel, 미래에셋증권 리서치센터

한편, GTC 2024에서는 모델 학습에 필요한 제조업 데이터가 부족하다는 지적이 제기됐다. 그러나, 이번 CES 2025에서 합성 데이터 생성으로 해당 문제를 해소하려는 전략을 진일보 시켰다는 점은 놀라운 부분이었다. 게다가, 가상 데이터가 현실 세계를 완벽하게 반영하지 못한다는 우려에 대해서도, 엔비디아의 솔루션은 뚜렷했다. **새로운 기반 모델인 Cosmos라는 "월드 모델"의 물리 법칙 이해 능력에다가 기존의 Omniverse(3D 가상현실 제작 협력 플랫폼)의 융합을 꾀한 것**이다. 다시 말해, '물리적으로 정렬된', 즉 현실 세계의 물리 법칙을 따르는 가상 세계를 생성함으로써 문제를 극복하려는 전략으로 풀이된다.

결론적으로, 엔비디아는 AI 플랫폼 기업으로 도약하겠다는 포부를 분명히 드러냈다. CES 2025 기조연설은 GTC 2024에서 제시된 엔비디아의 비전과 전략을 더욱 구체화하고 확장했을 뿐만 아니라, AI, 로보틱스, 그리고 메타버스 분야에서의 구체적인 청사진을 제시했다. 이는 엔비디아의 미래 성장 동력을 명확히 보여준 셈이다.

특히, **'Physical AI'와 '월드 파운데이션 모델'이라는 개념은 엔비디아가 단순히 하드웨어와 소프트웨어를 넘어, 현실 세계와 가상 세계를 연결하고, AI를 통해 새로운 가치를 창출하는 플랫폼 기업으로 도약하고 있음을 보여준다.** 기조연설이 끝난 후, 다시 한번 기술 리더로서의 입지를 굳혔다는 평가가 지배적이었다고 풀이된다.

그림 19. ChatGPT 출시 이후, 엔비디아와 AMD의 PER 추이와 둘 간의 PER 차이



자료: Bloomberg, 미래에셋증권 리서치센터

기조연설을 마무리하며 젠슨 황이 던진 'Happy New Year!'라는 인사말은, 우리에게 던진 새해 인사이기도 하지만… AI 시대를 맞은 본인들에게 하는 말과 다름없지 않을까?

표 6. CES 2025에서의 엔비디아 젠슨 황 CEO의 키노트 직후 들려온 각종 협력 소식 (일부만 정리한 것…)

협력 당사자	분야	NVIDIA와의 협력 내용 및 핵심 기술	기대 효과
Innoviz	자율주행 (LiDAR)	- NVIDIA DRIVE AGX Orin 플랫폼에서 실행되는 LiDAR 소프트웨어 개발 - 차량 환경 실시간 정보 처리, 물체 감지/분류/추적	자율주행 안전 향상, 정밀한 LiDAR 센싱 LiDAR 솔루션 접근성 및 호환성 향상
Arbe	자율주행 (Radar)	- NVIDIA DRIVE AGX 플랫폼과 통합된 4D 이미징 레이더 기술 제공 - 고해상도, 장거리 탐지, 기존 레이더 한계 극복	악천후/어두운 환경에도 정확한 감지 고성능, 고해상도 레이더 솔루션
Aurora	자율주행 (상용트럭)	- NVIDIA DRIVE Thor 고성능 AI 칩, DRIVE OS 활용 무인 트럭 개발 - 콘티넨탈과 협력, 2025년 텍사스 무인 트럭 상용 서비스, 2027년 대량 생산	24시간 운행, 인건비 절감, 정확한 배송 사고 위험 감소, 안전 운송 보장
Uber	자율주행 (모빌리티 서비스)	- NVIDIA의 Cosmos, DGX Cloud과 Uber의 방대한 주행 데이터 결합	AI 알고리즘 훈련, 성능 개선 가속화
Illumina, IQVIA	헬스케어 (신약 개발, 유전체)	- Illumina: 멀티-오믹스 분석 소프트웨어에 Nvidia 가속 컴퓨팅 활용 - IQVIA: NVIDIA AI Foundry 사용 맞춤형 기반 모델 구축	신약 개발 가속, 맞춤 의료 실현 로봇 수술 보조, 환자 모니터링 자동화, 병원 운영 효율 개선
Apple/Valve	게임 (클라우드 게이밍)	- NVIDIA GeForce NOW 클라우드 게임 서비스 확장 - Apple Vision Pro, Meta Quest 3/3S, Pico 등 혼합현실 헤드셋 지원 - Steam Deck용 GeForce NOW 네이티브 앱, 최대 4K/60fps HDR 스트리밍	혼합 현실 헤드셋에서 고품질 게임 제공, GeForce NOW의 서비스 지역 확장
Logitech	게임 (라이브 스트리밍)	- NVIDIA ACE를 활용한 AI 스트리밍 어시스턴트 출시 - GeForce RTX GPU 기반 오디오 대 얼굴 애니메이션, 3D 렌더링 기술 제공 - Streamlabs 앱 스토어 통해 직관적 애플리케이션 제공	AI 개인방송 보조로 독창적 콘텐츠 제작, 스트리머의 방송 운영 부담 감소
크래프톤 (배틀그라운드)	게임 (캐릭터 상호작용)	- NVIDIA ACE를 기반으로 협동 플레이 가능 캐릭터(CPC) "Ally" 개발 - NVIDIA ACE를 통해 Mistral-Nemo-Minitron-8B-128k-instruct SLM(소형언어모델)을 활용하여 AI 팀원의 고도화된 게임 플레이 구현 (효과적인 의사소통, 전략 추천, 전리품 공유, 운전, 전투 등)	실제 사람과 협동하는 듯한 경험, AI와 고도화된 협동 플레이
크래프톤 (inZOI)	게임 (캐릭터 상호작용)	- NVIDIA ACE를 기반으로 CPC "스마트 조이" 개발하고, 주변 환경 및 사건에 반응하는 현실적이고 반응성 있는 캐릭터 구현 - NVIDIA ACE의 온디바이스 LM 기술을 활용하여 캐릭터 상호 작용 및 행동 개선	깊은 상호작용으로 캐릭터 유대감 형성, 게임 세계에 몰입도 극대화
NetEase (NARAKA)	게임 (캐릭터 상호작용)	- NVIDIA ACE를 기반으로 로컬 추론 방식 "AI 팀원" 기능 개발 - NVIDIA ACE를 통해 AI 팀원이 전투 지원, 아이템 찾기, 장비 교체, 기술 제안, 승리 전략 제안 등 플레이어 보조	개인 맞춤형 지원, 진입 장벽 완화
위메이드 (MIR5)	게임 (캐릭터 상호작용)	- NVIDIA ACE를 기반으로 최초의 AI 보스 개발 - NVIDIA ACE를 통해 AI 보스가 플레이어의 행동을 지속적으로 학습하고 적응하여 동적으로 변화하는 공격 패턴 생성. AI 보스가 플레이어의 이전 전투를 학습하고, 플레이어의 전략, 스킬, 장비에 적응하여 매번 다른 전투 경험 제공	예측 불허의 반복 플레이, 지루함 해소, 게임 수명 연장

자료: Nvidia, 외신, 미래에셋증권 리서치센터

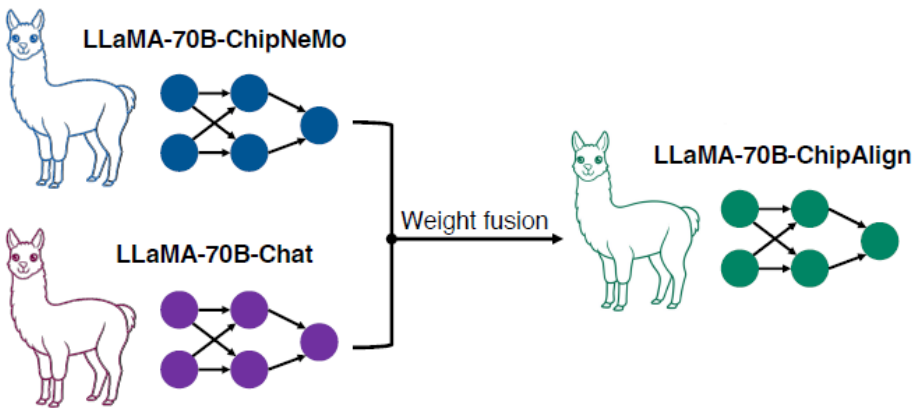
II. Paper of the Week

내 말을 더 잘 이해하는 ChipNeMo, ChipAlign – 엔비디아

AI를 선도하는 기업으로 평가받는 엔비디아는 AI를 지원하는 제품과 서비스를 확대하고 있을 뿐 아니라, 자신들의 개발 효율 향상을 위해 자체적으로도 AI를 매우 적극적으로 도입했다. 대표적인 예가 바로 ChipNeMo다. 반도체 개발을 위해 고안된 언어 모델인 ChipNeMo는 설계 자동화 스크립트, 버그, 회로 등 칩 설계와 관련된 전문 지식을 학습한 모델이다.

이 모델은 전문 지식은 뛰어나지만 일반 LLM에 비해 지시사항을 따르는 능력은 부족한 것으로 알려졌었다. 이는 실제로 엔지니어들이 사용하는 데 불편감을 주는 요소다. 자신이 의도한 바를 모델이 명확하게 이행하지 못한다면, AI를 사용함으로써 생기는 이득도 있지만 그만큼 수고로움도 커지기 때문이다. 이러한 문제를 해결하기 위해 엔비디아 연구진은 ChipAlign이라는 새로운 방법을 제안했다. 이는 칩 전문가 모델에 비해 일반적인 LLM이 지시사항을 더 잘 따른다는 점에서 착안해, 두 모델의 가중치를 결합해 두 모델의 장점을 모두 갖춘 새로운 모델을 생성하는 방식을 취했다.

그림 20. 모델 병합 과정 도식
ChipNeMo와 지시 정렬 모델을 미세조정한 뒤, 두 모델의 가중치를 융합해 ChipAlign 모델을 생성



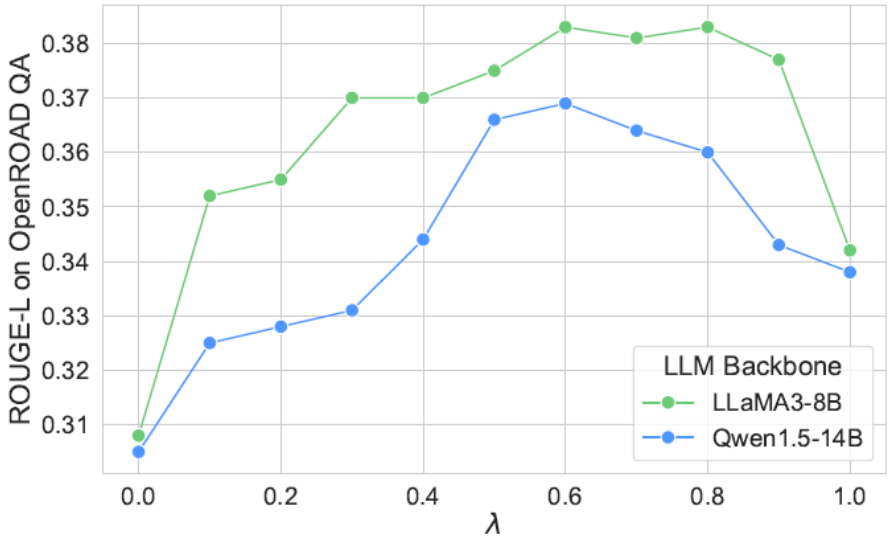
자료: arXiv, 미래에셋증권 리서치센터

이전에도 두 모델의 가중치를 통합해 새로운 모델을 생성하고자 하는 연구는 있었다. 그러나 **이전 모델과 ChipAlign의 차이는 가중치를 결합하는 방식에 있다.** 기존 연구는 가중치를 단순 평균하거나, 모델의 가중치 차이를 이용해 새로운 모델을 생성하는 방식을 사용했다. 반면 ChipAlign은 LLM의 가중치가 갖는 특성을 고려했다. LLM의 가중치는 고차원 공간 위에 존재하는 점으로 볼 수 있다. 또한 신경망의 복잡한 구조, 비선형 활성화 함수, 학습 과정 등을 생각한다면, 가중치들이 존재하는 이 공간은 평평한 유클리드 공간이 아닌 곡률을 갖는 공간일 가능성이 높다. 따라서 이 논문에서는 LLM의 가중치가 리만 다양체 위에 존재할 것이라고 가정했다.

곡률이 있는 표면 위에 가중치가 존재한다고 가정하는 경우, 이전 연구로 언급됐던 단순 평균법을 사용하는 것은 부적절하다. 예를 들어 서울에서 뉴욕의 좌표를 직선으로 잇는다면 중간 지점은 지각 밑이 되므로, 지표면 위에서의 중간 지점을 계산하기 위해서는 다른 방법을 사용해야 하는 것과 같다. 따라서 본 연구에서는 단순 평균 대신 **리만 다양체 위에 존재하는 두 점을 잇는 최단 경로인 지오데식을 따라 내분하는 지오데식 보간법(Geodesic Interpolation)을 사용해 새로운 가중치를 생성**했다.

칩 전문가 LLM과 지시 정렬된 LLM의 가중치를 각각 W_{chip}과 $W_{instruct}$라고 할 때, 두 가중치 사이의 지오데식 상에 있는 병합된 가중치 W_{merge}는 보간 파라미터인 λ에 따라 지오데식 상의 임의의 점을 나타내게 된다. **이 때 λ는 ChipAlign의 유일한 하이퍼파라미터로, 두 LLM의 가중치를 얼마나 혼합할지 결정하는 변수**다. LLaMA3-8B와 Qwen1.5-14B의 두 모델에 대해 λ 값을 0에서 1까지 변화시키면서 생성된 모델에 대한 벤치마크 성능을 측정한 결과, **λ가 0.6일 때 두 모델 모두에서 최상의 성능을 보여, 연구진들은 이 값이 일반적으로 최적 하이퍼파라미터로 사용될 수 있을 것으로 예상**했다.

그림 21. 두 LLM의 혼합비율을 나타내는 λ가 0.6일 때 LLaMA3와 Qwen 1.5 모델 모두 최고성능



자료: arXiv, 미래에셋증권 리서치센터

위의 방법을 통해 만들어진 다양한 파운데이션 모델을 가지고 연구진들은 실험을 수행했다. Qwen1.5-14B-Chat, LLaMA3-8B-Instruct 모델을 OpenROAD QA 데이터로 미세조정해서 칩 전문가 모델을 생성한 뒤, λ를 0.6으로 두고 병합해 Qwen1.5-14B-ChipAlign, LLaMA3-8B-ChipAlign 모델을 각각 생성했다.

그 뒤 생성된 답변과 정답 사이의 유사도를 측정해 모델의 성능 개선 여부를 확인했다. 그 결과, **ChipAlign 모델들이 모든 기존 병합 방법보다 우수한 성능을 보였으며, 특히 지시 사항이 포함된 질문에서 설계 자동화(EDA) 성능이 크게 개선**되었다. 이는 ChipAlign 방법이 기존에 부족했던 모델의 지시 정렬 능력을 효과적으로 강화했음을 의미한다.

그러나 지시 정렬 능력을 키우는 대가로 칩 설계 전문성이라는 도메인 지식을 희생할 수는 없다. 따라서 도메인 지식에 대한 평가도 함께 수행했다. 설계 자동화 스크립트, 버그, 회로 등에 대한 객관식 QA 벤치마크를 사용해 만들어진 ChipAlign 모델군을 평가한 결과, ChipAlign 모델들은 ChipNeMo와 유사한 도메인 지식 수준을 유지하고 있음을 확인할 수 있었다. 모델이 가지고 있던 전문 지식은 유지하면서 프롬프트의 지시에 부합하는 답변을 제공하는 모델을 만드는 데 성공한 셈이다.

엔비디아는 이러한 방법을 통해 기존의 ChipNeMo 성능을 한층 더 끌어올리는 데 성공했다. 처음 ChipNeMo를 만들었을 당시 엔비디아가 제시한 목표는 '1년차 직원이 질문했을 때 30년 경력의 엔지니어처럼 답변해주는 모델'이었다.

그리고 지난 2023년 11월 처음으로 ChipNeMo 논문이 공개된 이후 약 1년 2개월동안 엔비디아가 이뤄낸 것들을 생각하면, **30년의 경력자와 맞먹는 지식을 갖춘 모델이 내 생각과 질문을 정확하게 이해한 뒤 결과물을 제공할 수 있게 된 지금 엔비디아의 작업 효율성은 또 한 번 도약할 것으로 기대된다.**

III. AI Peer Table

표 7. Peer Table (1)

종목명	티커	주가 (현지통화)	시가총액 (조원)	기간별 수익률(%) 1W	1M	3M	6M	1Y	YTD	실적발표 예정일	AI 밸류체인 관련 한줄평
최종 소비자(온디바이스, 앱)											
애플	AAPL US	237.87	5,208	-2.0	-4.1	1.8	1.7	28.6	-5.0	01-30	엔드유저 데이터 보유 및 서비스 배포능력
테슬라	TSLA US	428.22	2,001	8.4	-1.8	95.0	69.5	95.6	6.0	01-29	FSD v13 공공 출시 임박 및 1Q25 중국 출시 가능성
삼성전자	005930 KS	53,700	321	-4.3	-3.4	-9.7	-38.4	-25.0	0.9	04-04	온디바이스 AI, HBM, 파운드리 영역에서의 가능성
퀄컴	QCOM US	164.41	266	3.4	3.7	-5.1	-20.2	19.6	7.0	01-31	생성 AI 처리 강화용 온디바이스 CPU 설계(X Elite)
스냅	SNAP US	11.84	29	-1.7	4.7	7.2	-27.0	-27.9	9.9	02-04	이미지 생성 모델을 탑재한 증강현실 서비스
비즈니스 효율화											
어도비	ADBE US	417.28	264	-0.5	-10.4	-17.9	-26.2	-30.0	-6.2	03-14	세계 최고의 미디어편집 툴. 서비스 배포능력(Firefly)
세일스포스	CRM US	321.16	447	-1.8	-9.2	11.5	27.4	18.8	-3.9	02-28	세계 최고의 CRM 업체. AI로 사용성 강화(Einstein)
서비스나우	NOW US	1,054.13	316	0.5	-6.0	13.6	37.3	44.6	-0.6	01-29	워크플로우 자동화. AI로 사용성 강화(Now Assist)
크라우드스트라이크	CRWD US	362.02	130	0.9	-1.6	19.8	-4.1	27.8	5.8	03-05	기업 고객 대상 엔드포인트 보안(Charlotte AI)
IBM	IBM US	220.03	296	-1.4	-4.7	-4.8	22.3	37.5	0.1	01-29	기업 고객 대상 AI 모델 개발, 배포 플랫폼(watsonx)
액센츄어	ACN US	349.73	319	-2.2	-2.5	-5.1	10.5	-0.3	-0.6	03-20	각 산업별 맞춤형 AI 에이전트를 만들기 위한 첨병
SAP	SAP GY	253.60	467	3.6	5.2	20.9	35.1	77.6	7.3	01-28	기업용 SW 솔루션에 AI를 내장한 에이전트(Joule)
인포시스	INFO IN	1,949.65	136	0.9	-2.5	0.6	15.5	21.7	3.7	01-16	각 산업별 맞춤형 AI 에이전트를 만들기 위한 첨병
타타컨설턴시	TCS IN	4,249.60	259	3.4	-5.0	3.5	2.4	10.9	3.8	04-11	각 산업별 맞춤형 AI 에이전트를 만들기 위한 첨병
유니티	U US	22.04	13	-4.0	-9.8	0.4	30.2	-36.3	-1.9	02-26	차세대 게임엔진에 AI 기능 대거 탑재(Sentis, Muse)
모더나	MRNA US	34.77	19	-19.5	-16.9	-39.3	-71.4	-67.2	-16.4	02-14	신약 개발 등 모든 업무에 생성 AI 활용(Dose ID)
루닛	328130 KS	63,400	2	-8.4	-23.4	56.9	35.8	-16.5	-0.6	03-21	AI 기반 암 검출/진단 보조 솔루션(인사이트)
AI Ops											
몽고DB	MDB US	242.91	26	-2.3	-9.1	-14.7	-3.8	-38.0	4.3	03-07	AI 모델 구축시 핵심 인프라인 '벡터 DB' 기능(Atlas)
클라우드플레어	NET US	114.48	57	-1.1	0.8	22.0	37.2	44.2	6.3	02-06	실시간 AI 추론에 CDN 수요증가 가능성(Workers AI)
팔란티어	PLTR US	68.14	226	-0.1	-10.4	60.6	137.7	306.6	-9.9	02-03	비즈니스 현장의 의사결정을 돕는 AI 플랫폼(AIP)
포스패러다임	6682 HK	40.85	4	-4.8	-22.9	89.6	-15.9	-19.9	-19.8	03-20	중국의 팔란티어(Sage)

자료: Bloomberg, 미래에셋증권 리서치센터
주: 실적발표 일정은 변경될 수 있음.

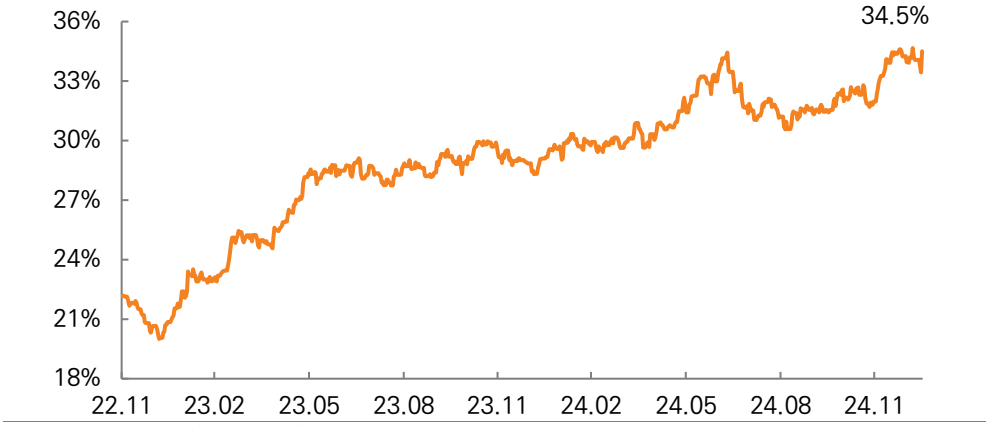
표 8. Peer Table (2)

종목명	티커	주가 (현지통화)	시가총액 (조원)	기간별 수익률(%) 1W	1M	3M	6M	1Y	YTD	실적발표 예정일	AI 밸류체인 관련 한줄평
파운데이션 모델											
알파벳	GOOGL US	195.55	3,497	0.8	3.0	18.3	5.1	37.6	3.3	01-30	OpenAI와의 AGI 경쟁이 가능한 업체(Gemini)
메타 플랫폼스	META US	617.12	2,269	1.0	-0.4	5.3	24.6	65.4	5.4	01-29	오픈소스 AI 개발의 선구자(Llama 등)
알리바바	9988 HK	79.95	285	-0.8	-5.2	-19.2	5.9	19.4	-3.0	02-07	중국의 CSP이자 중국 최고의 언어모델(Qwen)
바이두	9888 HK	78.50	41	-0.4	-10.0	-14.4	-13.9	-22.0	-5.1	02-28	NLP 및 자율주행 모델의 전통강호(Ernie Bot, Apollo)
센스타임	20 HK	1.33	9	1.5	-10.7	-11.3	-0.7	29.1	-10.7	03-26	중국의 멀티모달 AI의 다크호스(SenseNova)
네이버	035420 KS	206,500	33	2.0	-3.5	16.7	20.3	-10.2	3.8	02-03	한국형 LLM 개발 및 AI 서비스(HyperCLOVA)
클라우드											
마이크로소프트	MSFT US	426.31	4,614	0.4	-4.7	2.0	-5.7	10.6	1.1	01-29	OpenAI 모델 라이선스 독점권 보유한 세계 2위 CSP
아마존 닷컴	AMZN US	223.35	3,419	0.5	-1.8	19.0	15.9	44.5	1.8	01-31	최고의 CSP로서 Anthropic에 수 조원 투자
오라클	ORCL US	158.31	645	-2.7	-8.5	-8.8	11.2	50.1	-4.8	03-11	AI 데이터센터 capex 경쟁에 진입한 전통적 강자
소프트뱅크	9984 JP	9,001.00	123	-3.5	-4.5	-1.0	-21.0	41.6	-2.0	02-12	'비전펀드'는 AI에 집중. 일본 최고의 AI 슈퍼컴퓨터.
하드웨어 인프라											
엔비디아	NVDA US	136.24	4,857	-2.8	1.5	3.5	6.1	149.1	1.5	02-26	AI 모델 훈련 및 추론에 필수인 GPU계의 현존 최강자
브로드컴	AVGO US	228.00	1,556	-0.6	1.7	29.9	33.8	108.5	-1.7	03-07	이더넷 기반 네트워킹 반도체의 최강자
AMD	AMD US	119.96	283	-1.5	-5.5	-23.4	-33.3	-18.1	-0.7	02-04	서버용 CPU의 최강자. AI 가속기 분야 패스트팔로어
인텔	INTC US	19.72	124	-0.8	-3.0	-13.0	-42.4	-57.6	-1.6	01-30	파운드리로서 소버린 AI 미 지정학적 가치 부상
마이크론	MU US	103.19	167	3.8	0.8	-0.9	-21.0	25.8	22.6	03-20	SK하이닉스 추격 중인 "미국"의 메모리 반도체
SK하이닉스	000660 KS	198,200	144	-3.3	10.5	5.0	-14.8	50.7	14.0	01-23	HBM 부문 전세계 최강자
시놉시스	SNPS US	507.19	114	1.0	-0.2	-2.0	-17.7	2.6	4.5	02-21	반도체 EDA 부문 리딩 업체, AI 에이전트 적극 활용
암페놀	APH US	70.95	125	-1.2	-3.4	8.2	2.3	48.3	2.2	01-22	엔비디아향 AI용 서버 내 고속 I/O 및 커넥터 제조
버티브	VRT US	132.00	72	2.1	4.9	22.9	50.3	166.9	16.2	02-21	액체 냉각 방식에 있어 글로벌 선두업체
Arm	ARM US	147.40	226	1.5	-3.0	-2.2	-17.0	110.6	19.5	02-05	AI 가속기 설계 위한 다수의 IP를 소유한 팹리스
마벨 테크놀로지스	MRVL US	116.00	146	-1.8	-3.9	46.2	57.8	77.1	5.1	03-07	데이터 인프라용 네트워킹 및 스토리지 ASIC 설계
SMIC	981 HK	37.20	81	18.3	46.7	44.2	105.5	122.0	17.0	02-06	7나노 공정을 달성한 중국 유일의 미세공정 파운드리
TSMC	2330 TT	1,065.00	1,218	-3.2	-1.8	2.3	1.8	86.7	-0.9	01-16	명실상부 반도체 파운드리 최강자
콴타	2382 TT	260.00	44	-9.7	-5.5	-11.9	-19.4	17.8	-9.4	03-17	GPU 기반 보드 및 서버 시스템 조립 담당, 대만 업체
위스트론	3231 TT	101.50	13	-2.9	-4.7	-7.3	-10.6	5.4	-2.4	01-20	GPU 기반 보드 및 서버 시스템 조립 담당, 대만 업체
Alchip	3661 TT	2,920.00	10	-5.8	-3.3	49.4	0.9	-19.7	-11.0	03-03	빅테크들의 AI 가속기 설계를 돕는 대만의 팹리스

자료: Bloomberg, 미래에셋증권 리서치센터
주: 실적발표 일정은 변경될 수 있음.

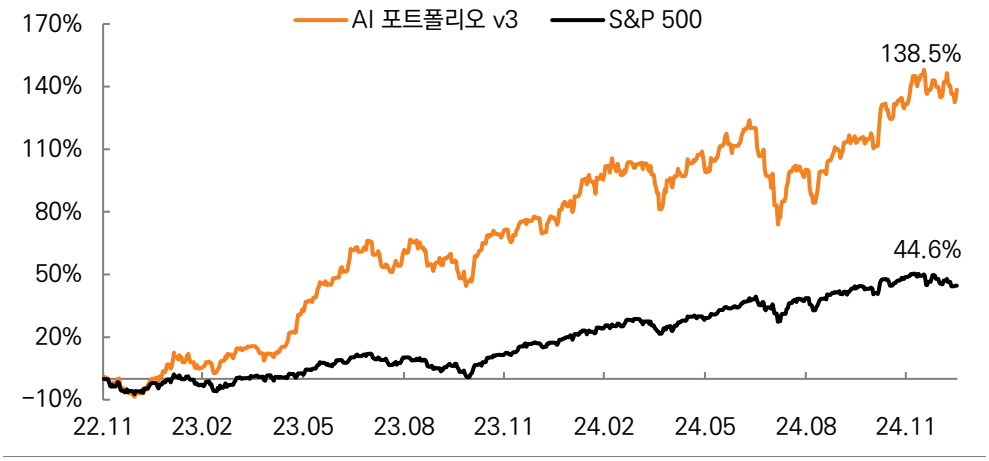
IV. Appendix: Charts

그림 22. ChatGPT 출시일 이후, S&P 500에서 Magnificent Seven이 차지하는 비중(시가총액 기준)



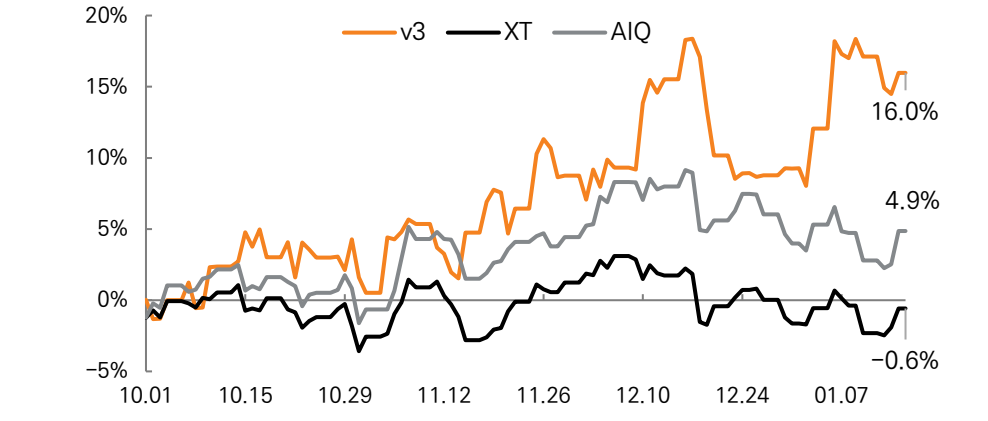
자료: Bloomberg, 미래에셋증권 리서치센터

그림 23. ChatGPT 출시일 이후, AI 밸류체인에 속한 유니버스 종목들의 평균 수익률(동일가중 방식)



자료: Bloomberg, 미래에셋증권 리서치센터

그림 24. 대표 AI ETF vs AI 유니버스(v.3) 수익률 추이(24.09.30 기준)
- Global X Artificial Intelligence & Technology ETF(AIQ), iShares Exponential Technologies ETF(XT)



자료: Bloomberg, 미래에셋증권 리서치센터

표 9. LLM 벤치마크 챗봇아레나의 "Hard Prompts" 기준 성능 상위 모델 10개(동일 모델 생략)

순위 (전 주 변동)	모델명	아레나 점수	기관명	라이선스	지식 컷오프
1(-)	o1-2024-12-17	1365	OpenAI	독점 폐쇄소스	
1(-)	Gemini-exp-1206	1365	구글	독점 폐쇄소스	-
3(-)	Gemini-2.0-flash-thinking-exp-1219	1358	구글	독점 폐쇄소스	2024년 8월
4(-)	o1-preview	1352	OpenAI	독점 폐쇄소스	2023년 10월
5(-)	Gemini-2.0-flash	1351	구글	독점 폐쇄소스	2024년 8월
6(New)	o1-mini	1338	OpenAI	독점 폐쇄소스	
7(▼1)	ChatGPT-4o-latest	1337	OpenAI	독점 폐쇄소스	2023년 10월
8(▼1)	Deepseek-v3	1306	DeepSeek	상업적 사용허가	
9(▼1)	Claude 3.5 Sonnet (20241022)	1304	Anthropic	독점 폐쇄소스	
10(▼1)	Gemini 1.5 Pro	1298	구글	독점 폐쇄소스	2023년 11월

자료: lmarena.ai, 미래에셋증권 리서치센터

주: Hard Prompts는 기존의 일반적인 요청보다 훨씬 더 까다롭고 복잡한 문제를 제시하여 LLM의 한계를 테스트하는 벤치마크

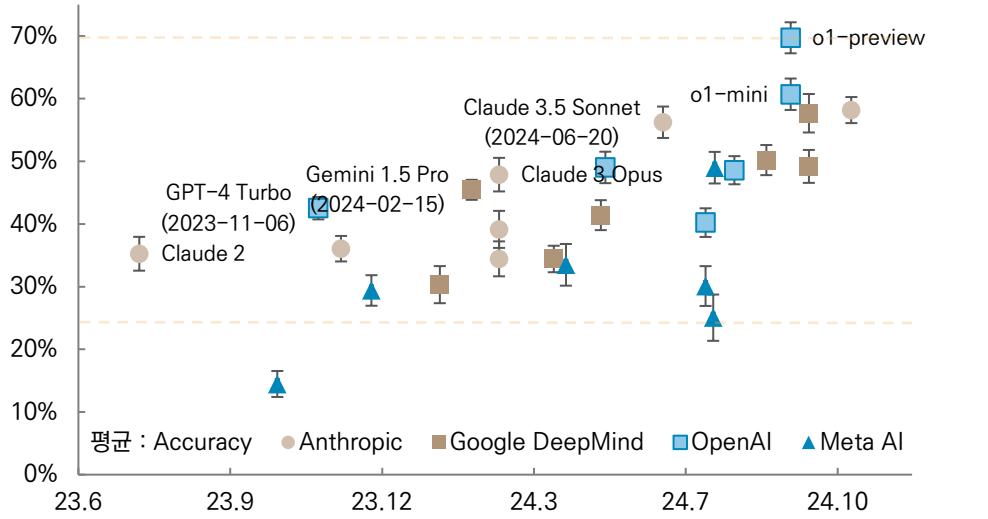
표 10. LLM 벤치마크 LiveBench의 성능 상위 모델 10개(동일 모델 생략)

순위 (전 주 변동)	모델명	전체 평균 점수	기관명	라이선스	지식 컷오프
1(-)	o1-2024-12-17	75.67	OpenAI	독점 폐쇄소스	
2(-)	o1-preview-2024-09-12	65.79	OpenAI	독점 폐쇄소스	
3(-)	Gemini-exp-1206	64.09	구글	독점 폐쇄소스	-
4(-)	Gemini-2.0-flash-thinking-exp-1219	61.43	구글	독점 폐쇄소스	
5(-)	deepseek-v3	60.45	Deepseek	상업적 사용허가	
6(-)	Gemini-2.0-flash	59.26	구글	독점 폐쇄소스	2024년 8월
7(-)	Claude-3.5 Sonnet-20241022	59.03	Anthropic	독점 폐쇄소스	2024년 4월
8(-)	o1-mini	57.76	OpenAI	독점 폐쇄소스	2023년 10월
9(-)	Gemini-exp-1121	57.36	구글	독점 폐쇄소스	-
10(-)	GPT-4o	55.33	OpenAI	독점 폐쇄소스	2023년 10월

자료: LiveBench, Huggingface, 미래에셋증권 리서치센터

표 11. 각 AI 모델들의 GPQA Diamond 기준 성능 비교

GPQA Diamond는 박사 수준의 과학 분야 질문으로, 인간 전문가들도 약 65%의 정확도를 보임



자료: EPOCH AI, 미래에셋증권 리서치센터

Compliance Notice

- 당사는 자료 작성일 현재 조사분석 대상법인과 관련하여 특별한 이해관계가 없음을 확인합니다.
- 당사는 본 자료를 제3자에게 사전 제공한 사실이 없습니다.
- 본 자료를 작성한 애널리스트 김은지은(는) 자료작성일 현재 엔비디아 52주 보유하고 있습니다.
- 본 자료는 외부의 부당한 압력이나 간섭없이 애널리스트의 의견이 정확하게 반영되었음을 확인합니다.