rebellions_

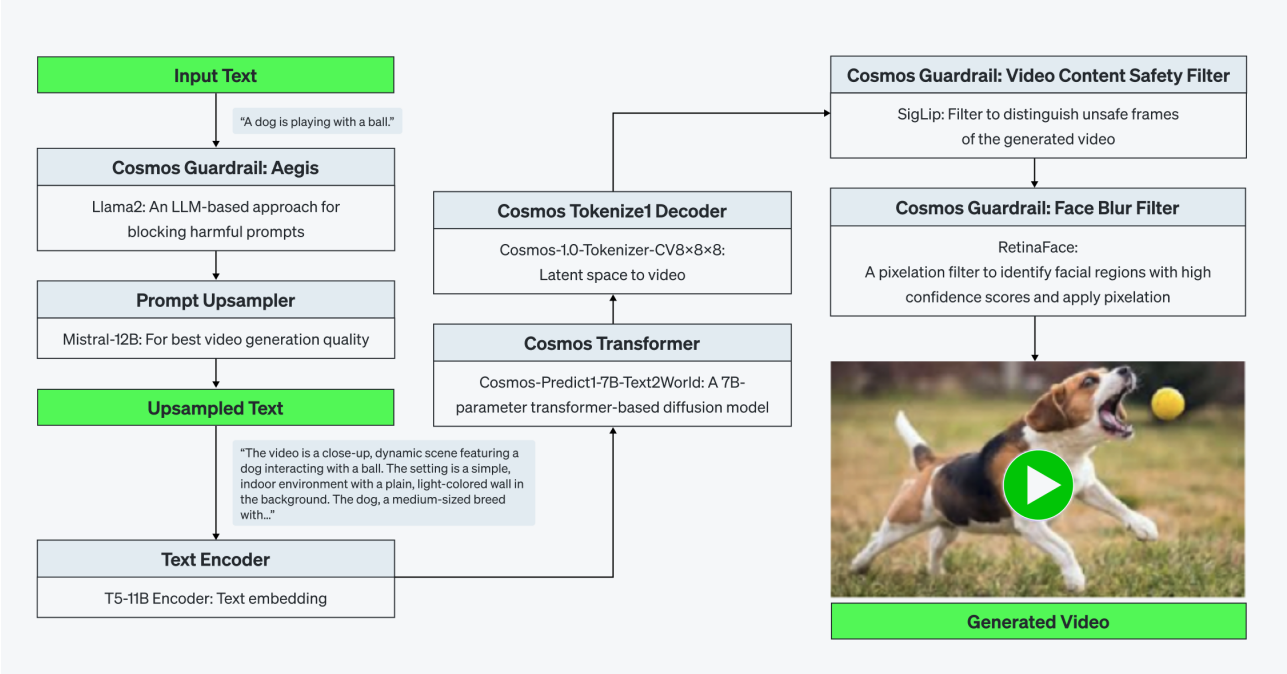
물리적 한계를 넘는 AI: ATOM™으로 Cosmos를 실현하다

May 19, 2025

멀티모달 AI는 인프라의 경계를 새로 쓰고 있습니다. 이제 모델은 더 이상 하나의 입력 형태에만 국한되지 않습니다. 텍스트, 이미지, 영상 등 다양한 입력이 동시에 통합된 세상—그 최전선에는 Cosmos가 있습니다.

Cosmos는 diffusion 기반 텍스트-투-비디오 생성 모델로, 토크나이저(tokenizer), 사전학습된 월드 파운데이션 모델, 가드레일(guardrail) 시스템 등 **5~6개의 독립된 AI 서브시스템**으로 구성되어 있습니다. 각 서브시스템은 고유한 아키텍처적 복잡성과 연산 특성을 가지고 있습니다.

기존에는 이러한 대규모 워크로드를 오직 **고성능 GPU 플랫폼**만이 감당할 수 있었습니다. 이는 GPU가 최적이라서가 아니라, **대체재가 없었기** 때문이었습니다. GPU는 강력한 범용 가속기지만, **고정된 소프트웨어 스택과 경직된 메모리 계층**에 의존하기 때문에 새로운 모델 구조에 대한 적응이 느리고, 비효율적이며, 때로는 과도한 수작업을 요구합니다. 그러나 **Cosmos**의 핵심은 단순한 크기가 아닙니다. 그 구조적 다양성과 동적 시퀀스입니다. 따라서 효율적인 실행은 단순히 연산 속도가 아니라, **시스템 전체의 적응성**의 문제입니다. 이것이 바로 **리벨리온**이 해결하고자 한 과제였습니다.



[Figure 1. Cosmos-Predict1-7B + Cosmos Guardrail 도식

새로운 패러다임: 불가능을 가능하게 하다

리벨리온은 **상용 NPU 최초로 Cosmos-Predict1-7B를 실시간으로 구동**했습니다. 단순히 성능을 밀어붙인 것이 아니라, **본질적으로 유연하고 확장 가능한 스택**을 기반으로 이루어낸 것입니다. 그리고 여기서 끝이 아닙니다.

혁신의 구조: 풀스택 역량의 총합

Cosmos가 **ATOM™** 위에서 실행될 수 있었던 이유는, 컴파일러부터 런타임, 실리콘까지 **스택의 모든 계층이 아키텍처 다양성에 적응하도록 설계**되었기 때문입니다. 다른 이들이 하나의 모델에 최적화할 때, 리벨리온은 **다양한 모델을 동시에 수용하는 유연성**을 구축했습니다. 이 유연성은 사후(add-on)가 아니라 **설계에 내재된 구조적 속성**입니다.

AI 모델이 끊임없이 진화하는 만큼, 그에 따라 인프라도 함께 진화해야 합니다. GPU 시스템은 여전히 강력하지만, **레거시 툴체인과 분리된 소프트웨어 계층, 공급업체 종속 구조(vendor lock-in)** 로 인해 유연한 대응이 어렵습니다.

이에 리벨리온은 **시스템 레벨 접근(System-level Approach)** 을 택했습니다. 컴파일러, 런타임, 하드웨어를 모두 사내에서 직접 개발함으로써, **모델이 인프라에 맞추는 것이 아니라, 인프라가 모델에 맞춰 진화하는 플랫폼**을 구축했습니다.이 철학이야말로 Cosmos on ATOM™을 **가능하게 했을 뿐 아니라, 확장 가능하게 만든 핵심 원리**입니다.

[Figure 2. 리벨리온 풀스택]

소프트웨어 혁신: 적응성을 위한 디자인

- **모듈형 컴파일러 스택:** 적응성은 모듈성에서 시작됩니다. 리벨리온의 컴파일러는 200개 이상의 모델을 지원하는 풍부한 연산 프리미티브 라이브러리를 갖추고 있으며, 트랜스포머(Transformer), 컨볼루션, 디퓨전(Diffusion) 모델을 모두 처리할 수 있는 프론트엔드, 그리고 하드웨어 최적화 코

드를 자동 생성하는 백엔드로 구성되어 있습니다. 개별 튜닝 없이도 모델별 커널을 효율적으로 변환합니다.

- **통합 런타임 오케스트레이션:** 모델마다 실행 패턴이 다릅니다. 리벨리온의 런타임은 모델의 동작 특성을 분석해 메모리·연산·커널 실행 흐름을 동적으로 조율합니다. 이를 통해 Cosmos처럼 레이어별 부하가 달라지는 워크로드에서도 최대 활용률을 유지합니다. 또한 Predictive DMA 스케줄링을 도입해, 데이터를 "필요하기 직전"에 미리 로드함으로써 비동기·다단계 워크로드에서도 연산 지연을 방지합니다.
- **확장형 통신 인프라:** 확장은 선택이 아니라 필수입니다. "Rebellions Scalable Design"은 고대역폭 통신 계층과 텐서 병렬 실행을 통해 멀티카드 추론을 가능하게 합니다. 이를 통해 Cosmos는 여러 NPU에 걸쳐 분산 실행되면서도 지연이나 병목 없이 동작합니다.

하드웨어 혁신: 유연성을 위해 설계된 아키텍처

유연한 연산 구조

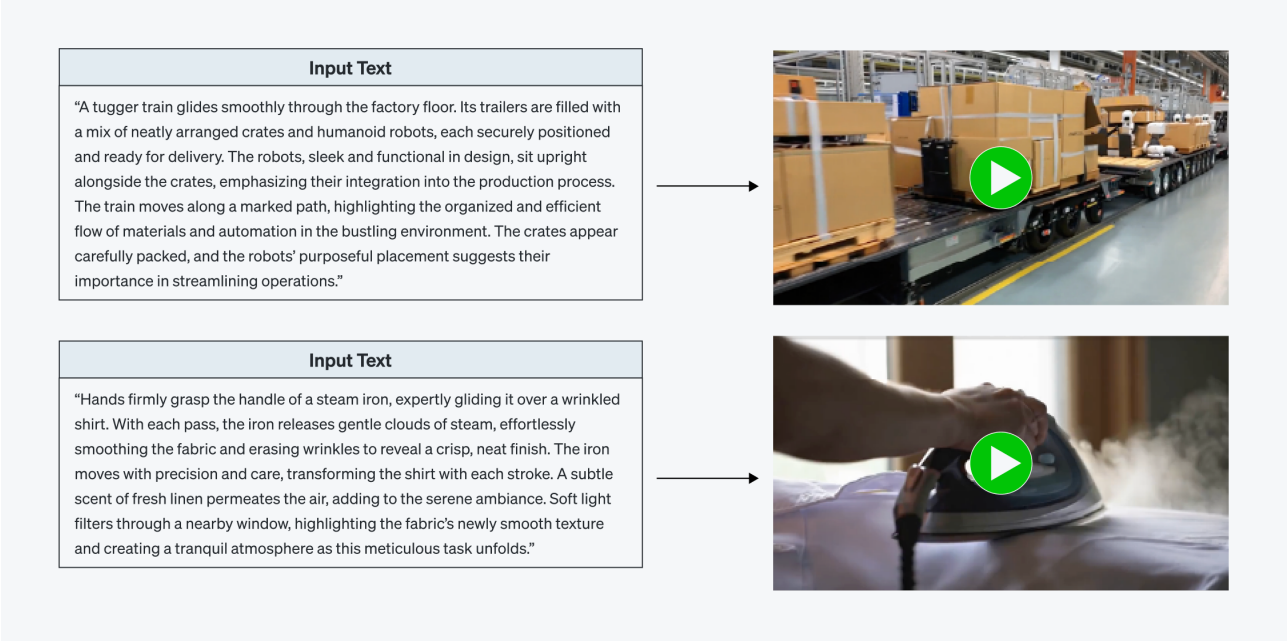
하드웨어는 모델과 함께 움직여야 합니다. 리벨리온은 ATOM™의 연산 코어를 **다양한 연산자와 동적 실행 경로를 폭 넓게 지원**하도록 설계했습니다. 이를 통해 Cosmos의 다양한 연산 유형을 **커널 재작성 없이 그대로 실행**할 수 있었습니다.

메모리-연산 동기화

현대 AI 모델은 선형적으로 연산하지 않습니다. 리벨리온은 **런타임 동작을 추적하고, 메모리 흐름을 실시간으로 조정하는 메모리-연산 조정 로직**을 설계했습니다. 이를 통해 Cosmos와 같은 **비정형·버스트성 연산 단계**를 별도의 특수 엔지니어링 없이 처리할 수 있습니다.

고대역폭 인터커넥트 기반 멀티카드 스케일링

NPU 간 확장을 위해 **시스템을 처음부터 다시 설계할 필요는 없습니다.** 리벨리온은 자체 개발한 고속 인터커넥트 IP를 통해 텐서 병렬 워크로드에서도 **저지연·고처리량 통신**을 보장합니다. **Cosmos**와 같은 복합 구조의 모델도 대규모 스케일에서 **병목 없이 동시 실행**이 가능합니다.



[Figure 3. 프롬프트에서 영상으로: ATOM™에서의 Cosmos-Predict1-7B 추론]

추론 결과 보기

결론: 새로운 AI 시대의 개막

Cosmos on ATOM™은 단순한 기술 시연이 아닙니다. **“시스템이 모델에 적응해야 한다”** 는 리벨리온의 아키텍처 철학이 실현된 사례입니다. 리벨리온은 하나의 모델을 위한 칩을 설계하지 않았습니다. **앞으로 등장할 모든 모델을 위한 플랫폼을 설계했습니다.** 모델 복잡도가 기하급수적으로 늘어나는 시대— 빠르게 적응하고, 유연하게 확장하며, 정밀하게 실행하는 인프라만이 미래를 주도할 것입니다. 리벨리온은 바로 그 미래를 만들고 있습니다. **Cosmos는 그 시작일 뿐입니다.**