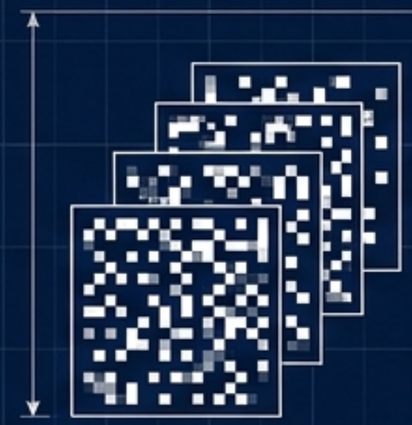
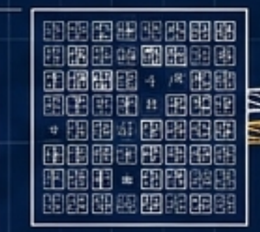
언어에서 행동으로: 로봇 지능과 VLA의 진화 (2022-2026)

아키텍처와 시간축으로 읽는 설계 공간(Design Space)의 해부도
대학원 심층 강의자료 (Explainer & Architectural Deep-dive)



IMAGE_INPUT_MATRIX[1024x768]
PIXEL_STREAM_0x4ASF
RGB_DATA_FLOW::FRAME_N



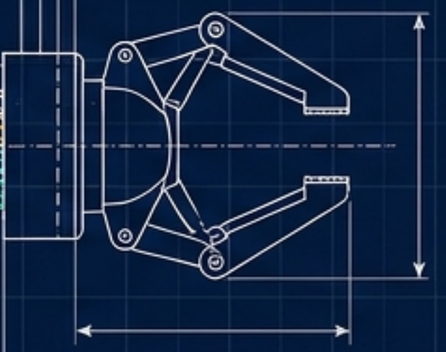


TXT_TOKEN_SEQ[8:255]
NATURAL_LANGUAGE_PRONPT: '찍업 더 레드 큐브'
SEMANTIC_ENBEDDING::VECTOR_1x512
COMMAND_STRING::ACTIVATE_GRIPPER

TRACK_01::VISUAL_FOUNDATIONS (e.g., ViT, CLIP)
TRACK_02::LANGUAGE_MODELS (e.g., BERT, GPT-4)
TRACK_03::MULTIMODAL_FUSION (e.g., Flamingo, PaLM-E)
TRACK_04::ACTION_EXECUTION (e.g., RT-1, RT-2, VLA)

POS(x,y,z):[1.24, 0.56, -0.88]
VELOCITY::VECTOR[0.3m/s]
TORQUE_LIMIT::JOINT_3

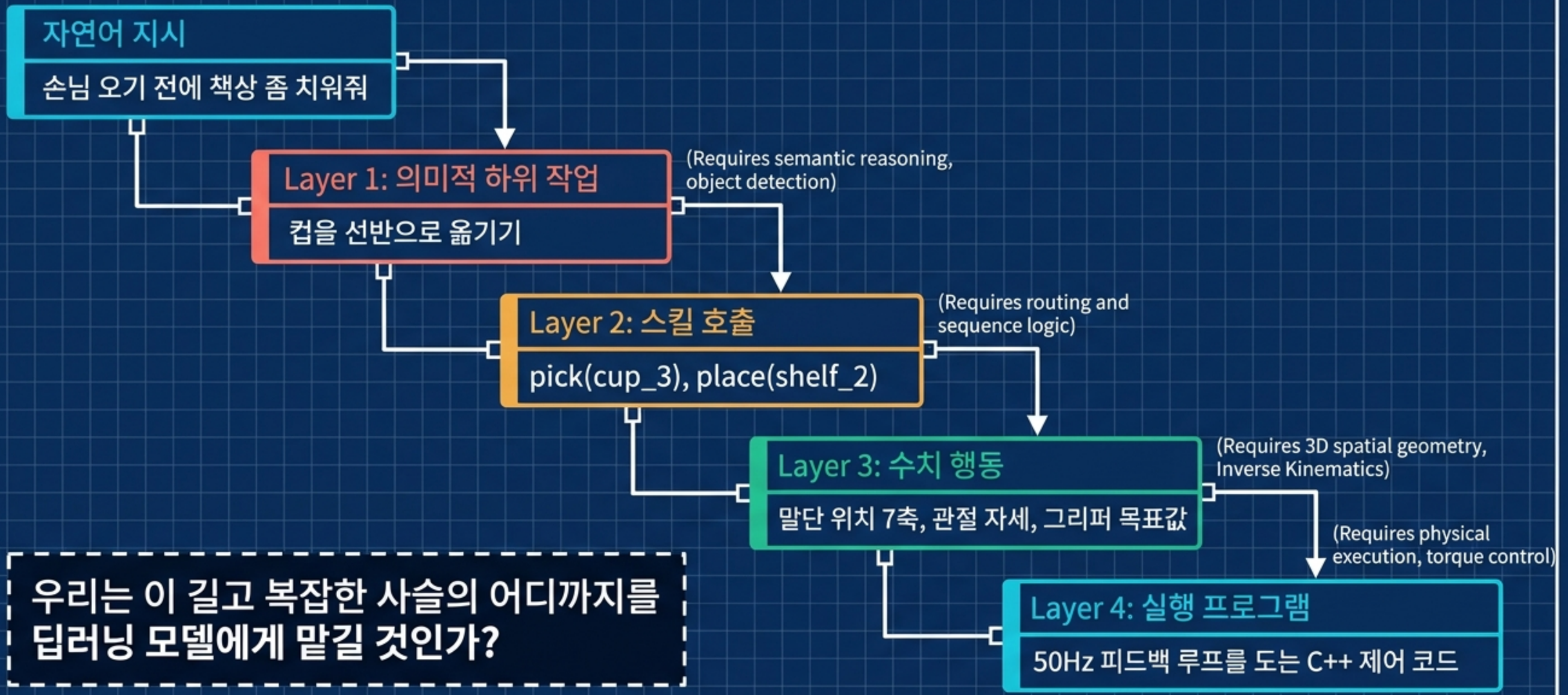
END_EFFECTOR_ACTUATION::GRIP
FORCE_SENSOR_READING: [12.5 N]
OBJECT_CONTACT::CONFIRMED
PHYSICAL_ACTION_COMPLETE



END_EFFECTOR_ACTUATION::GRIP
FORCE_SENSOR_READING: [12.5 N]
OBJECT_CONTACT::CONFIRMED
PHYSICAL_ACTION_COMPLETE

Gemini Notebook

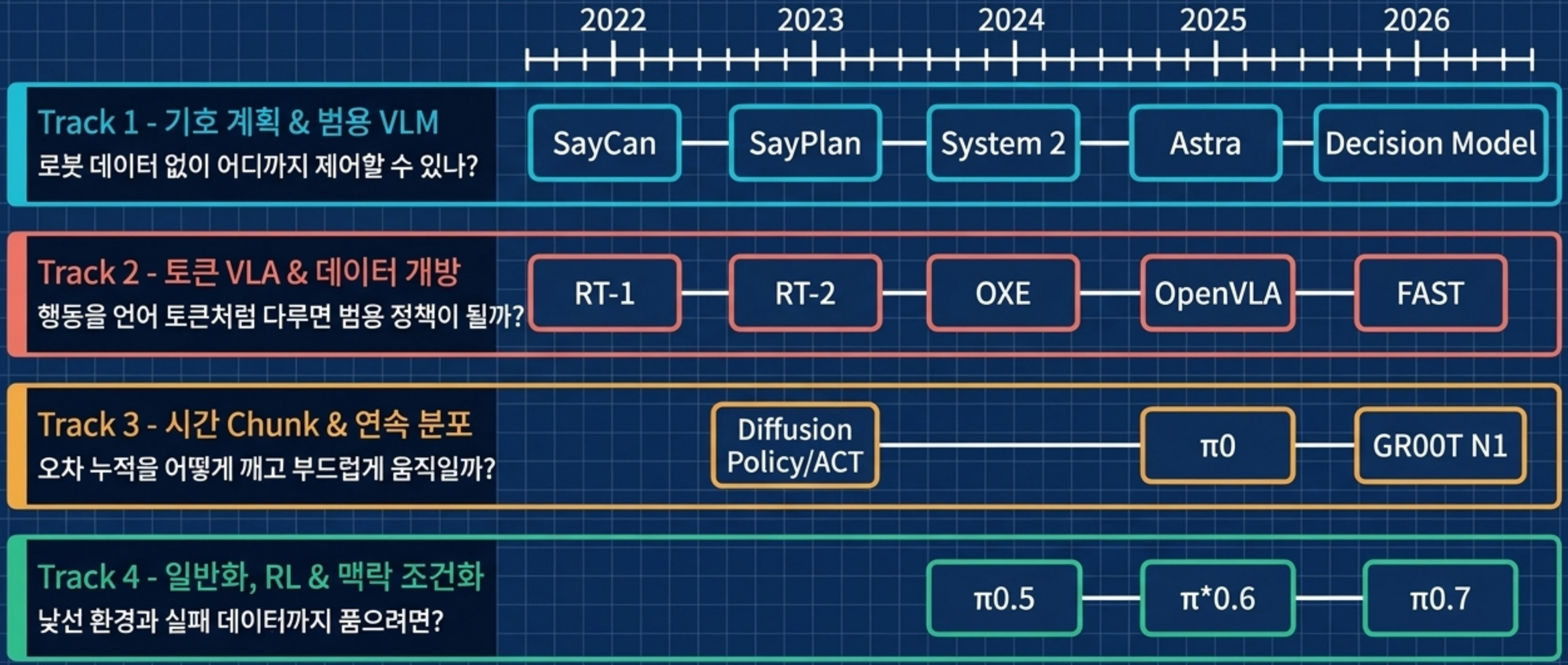
말과 모터 사이의 거대한 간극 (The Gap)



우리는 이 길고 복잡한 사슬의 어디까지를 딥러닝 모델에게 맡길 것인가?

Gemini Notebook

진화 계보도: 로봇 학습을 이끈 네 갈래의 길

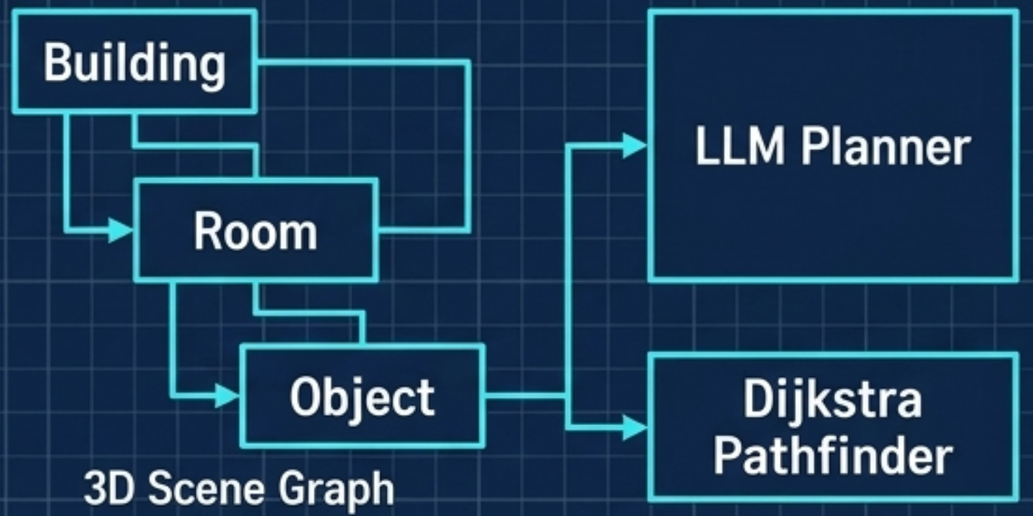


이 네 갈래의 길이 어떻게 서로의 병목을 해결하며 2026년의 통합 아키텍처로 수렴하는지 추적합니다.

Gemini Notebook

두 갈래의 접근법: 하향식 지휘봉 vs 상향식 픽셀 매핑

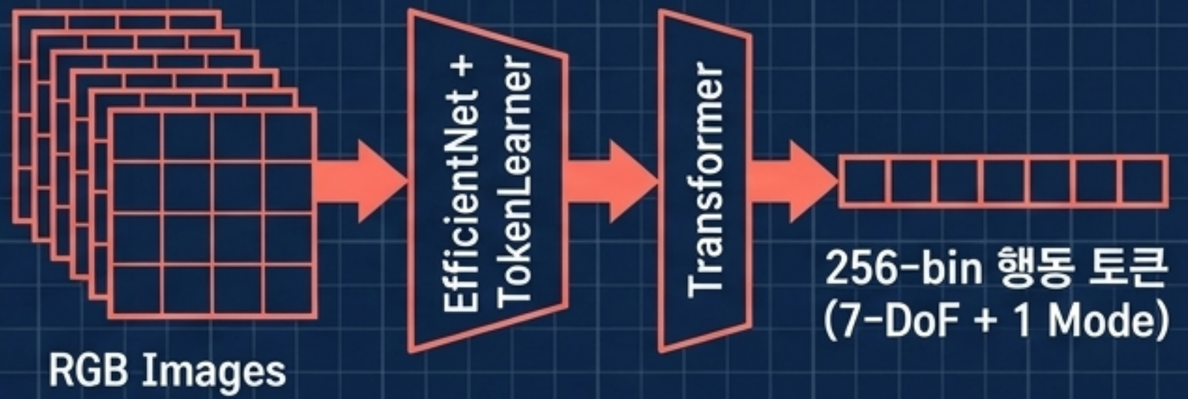
SayPlan (2023.07)



언어 모델에 지휘봉을 쥐여주다.

한계: 완벽한 상징적 계획도 컵이 미끄러지는 물리적 접촉의 순간에는 무력함 (Graph validation ≠ Physical success).

RT-1 (2022.12)



스킬을 부수고 숫자를 직접 배우다. (130k episodes, 3Hz Closed-loop).

한계: 로봇 제어는 완벽하지만 세상의 상식을 모름 (Cannot generalize to 'Bring a drink for a sleepy person').

Gemini Notebook

인터넷의 기억과 개방의 시대 (RT-2 & OXE)

RT-2 (2023.07)

Web VQA Data



Robot Demos



PaLI-X / PaLM-E
(Co-fine-tuning)

‘의미 일반화’ (Semantic Generalization)의 달성.
처음 보는 ‘멸종한 동물 장난감’을 웹의 지식으로 추론하여 집어냄.

OXE & OpenVLA (2023/2024)



RLDS Standard

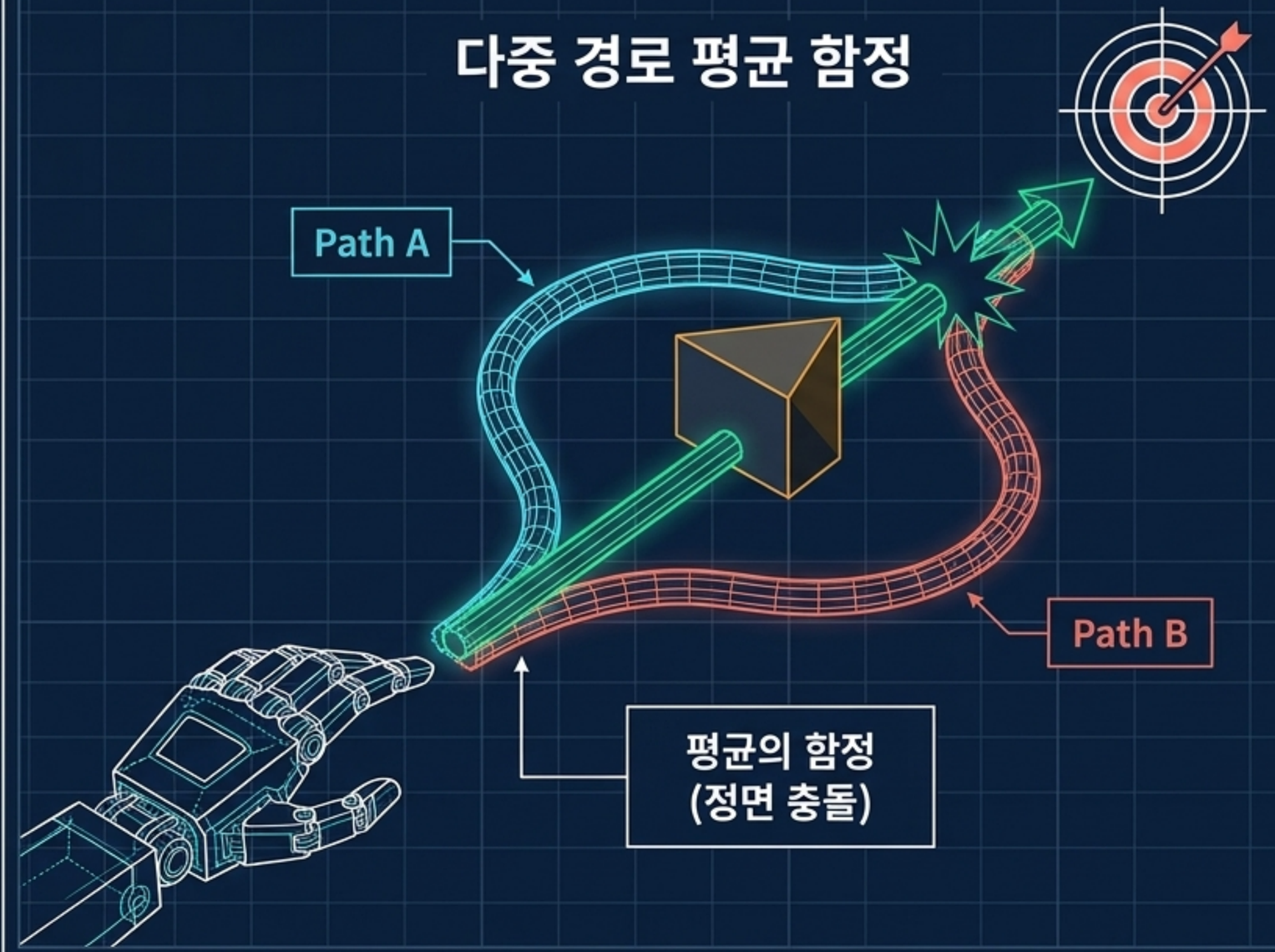
OpenVLA
(7B Llama-2 based open weight)

Cross-embodiment 전이 입증.
개별 연구실에서 LoRA 파인튜닝으로 7B 모델을 해부할 수 있는 오픈소스 시대 개막.

Gemini Notebook

"1-스텝" 예측의 치명적 한계 (The Bottleneck)

다중 경로 평균 함정



문제 1: 오차 누적 (Compounding Errors)

정지 사진 1장 -> 1-스텝 7차원 행동 예측.
1mm만 어긋나도 다음 프레임이 틀어지며
눈덩이처럼 오차가 불어남.

문제 2: 평균의 함정

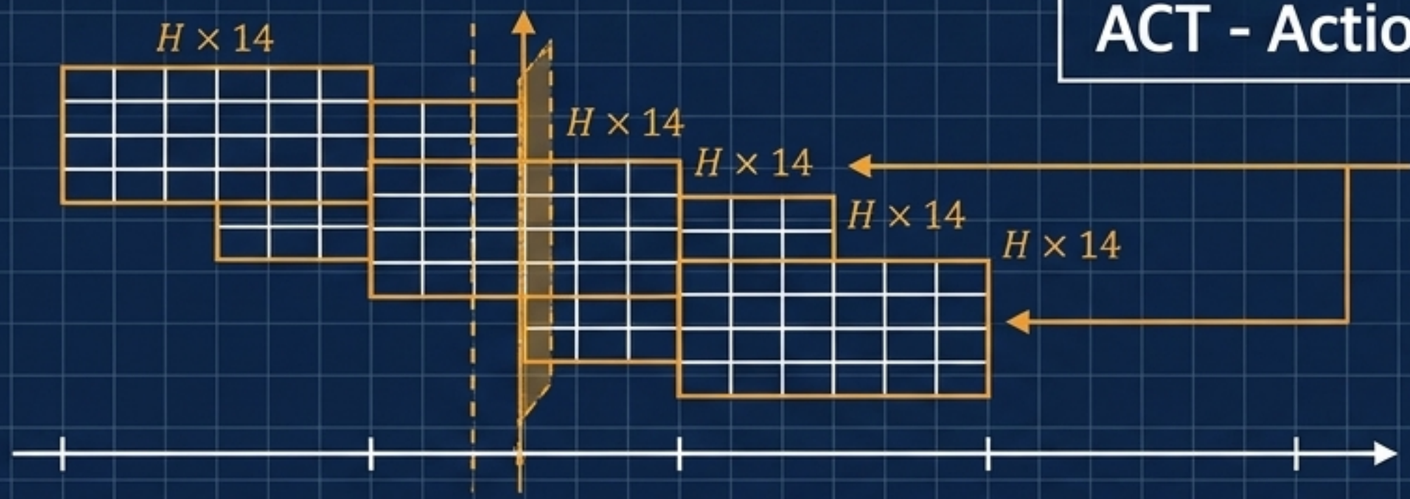
양갈래 회피 경로가 있을 때, 단일 스텝
회귀 모델은 두 경로의 평균값을 출력하여
장애물 정면 충돌.

왜 행동을 한 번에 하나씩만 예측하면 안 되는가?

Gemini Notebook

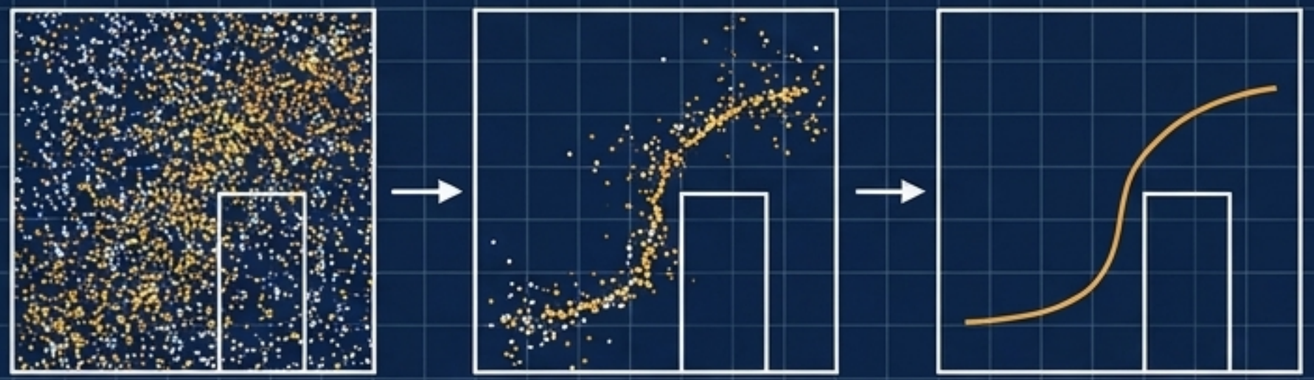
시간 Chunk와 다중 모달 분포의 발견 (2023의 혁명)

ACT - Action Chunking with Transformers



Temporal Ensembling.
매 시점마다 미래 H스텝 구간을 병렬 예측하고, 겹치는 시간축의 예측값들을 지수 가중 평균으로 부드럽게 융합.

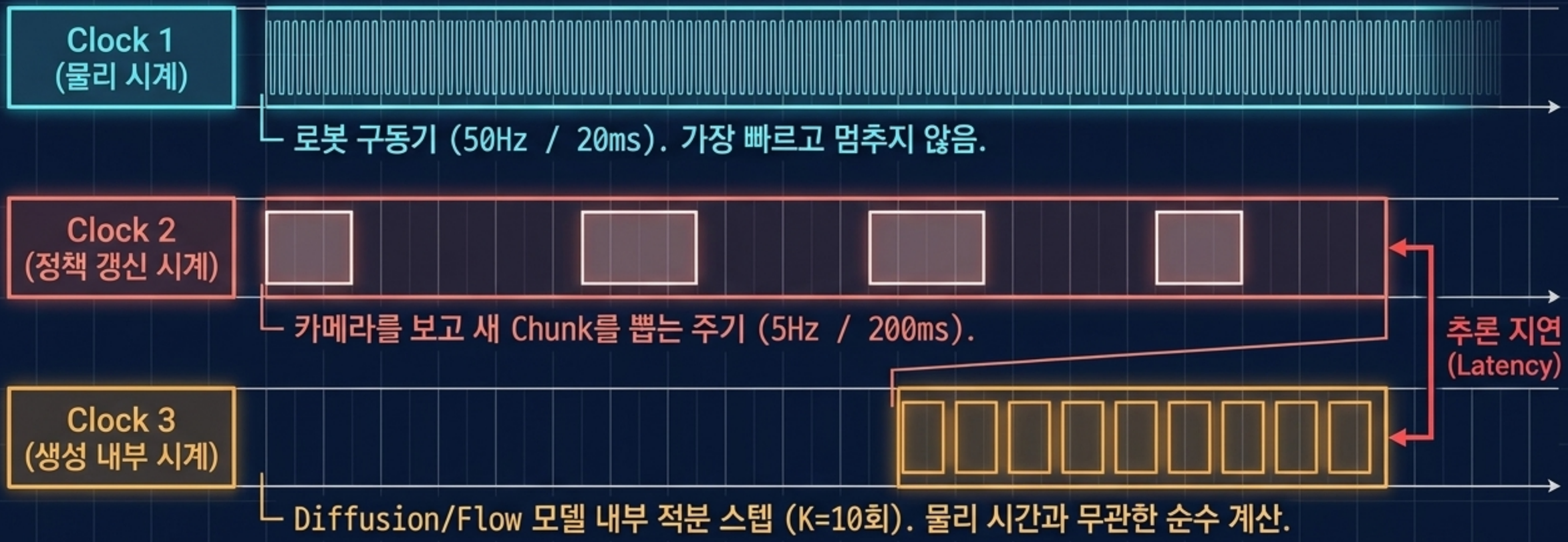
Diffusion Policy



Conditional DDPM.
평균을 내는 대신 '다중 모달 행동 분포' 자체를 학습하여 여러 유효 경로 중 하나를 온전히 샘플링.
Receding-horizon 제어.

Gemini Notebook

딜레마: '세 개의 시계'가 엇갈리다 (The 3 Clocks)





추론 지연(Latency)이 0.3초 발생하면, 로봇은 0.3초 전의 과거 풍경을 보고 생성한 궤적을 맹목적으로 실행하게 됨. **계산과 구동의 비동기화 숙제.**

Gemini Notebook

아키텍처의 분화: System 1+2 vs Flow Matching

System 1+2 (Helix / GR00T N1 - 2025)

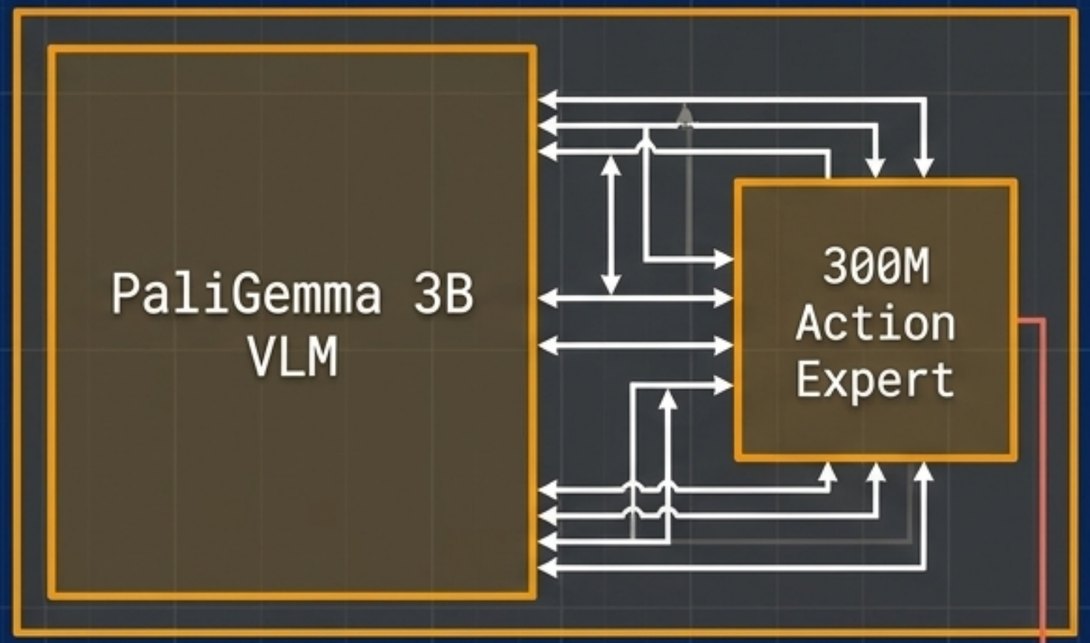
System 2 VLM @ 7-9Hz
(Brain)

Latent Goals

System 1 DiT @ 200Hz
(Spine/Cerebellum)

무거운 의미 추론과 실시간 고주파 제어의 물리적 분리.

Flow Matching (π0 - 2024.10)



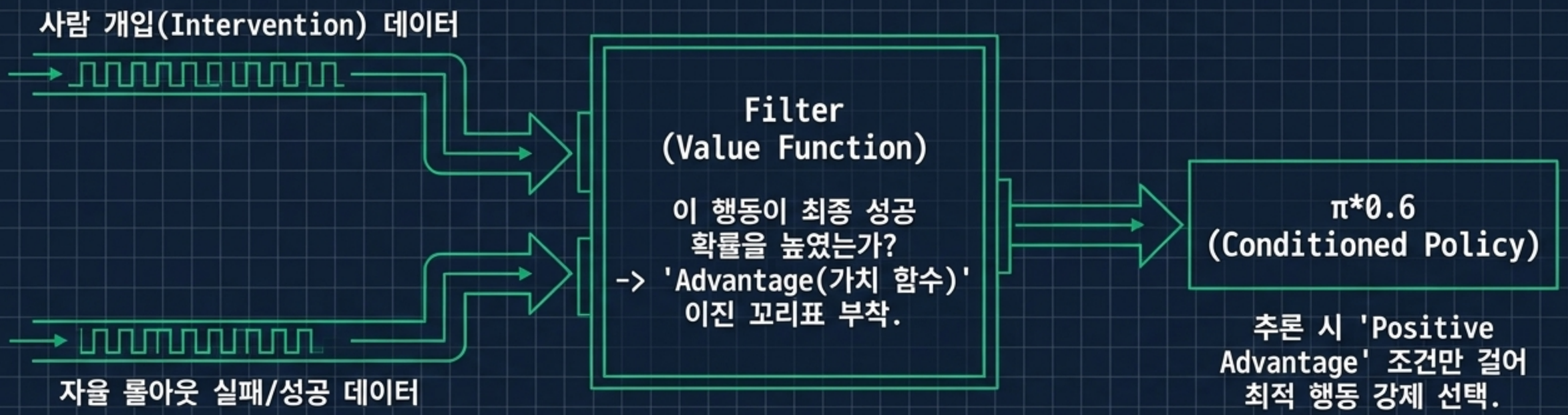
단일 모델 내부 결합.
수십 번의 Diffusion 스텝 대신,
10회의 Euler 적분만으로
노이즈를 H=50 연속 행동
궤적으로 변환.

Add-on (FAST Tokenizer)

시간축 궤적을 DCT 주파수로
변환 후 BPE 압축
-> 700개 실수를 30~60개
토큰으로 10배 압축.

Gemini Notebook

일반화와 경험의 구조화: 실패에서 배우다 (π0.5 & π*0.6)

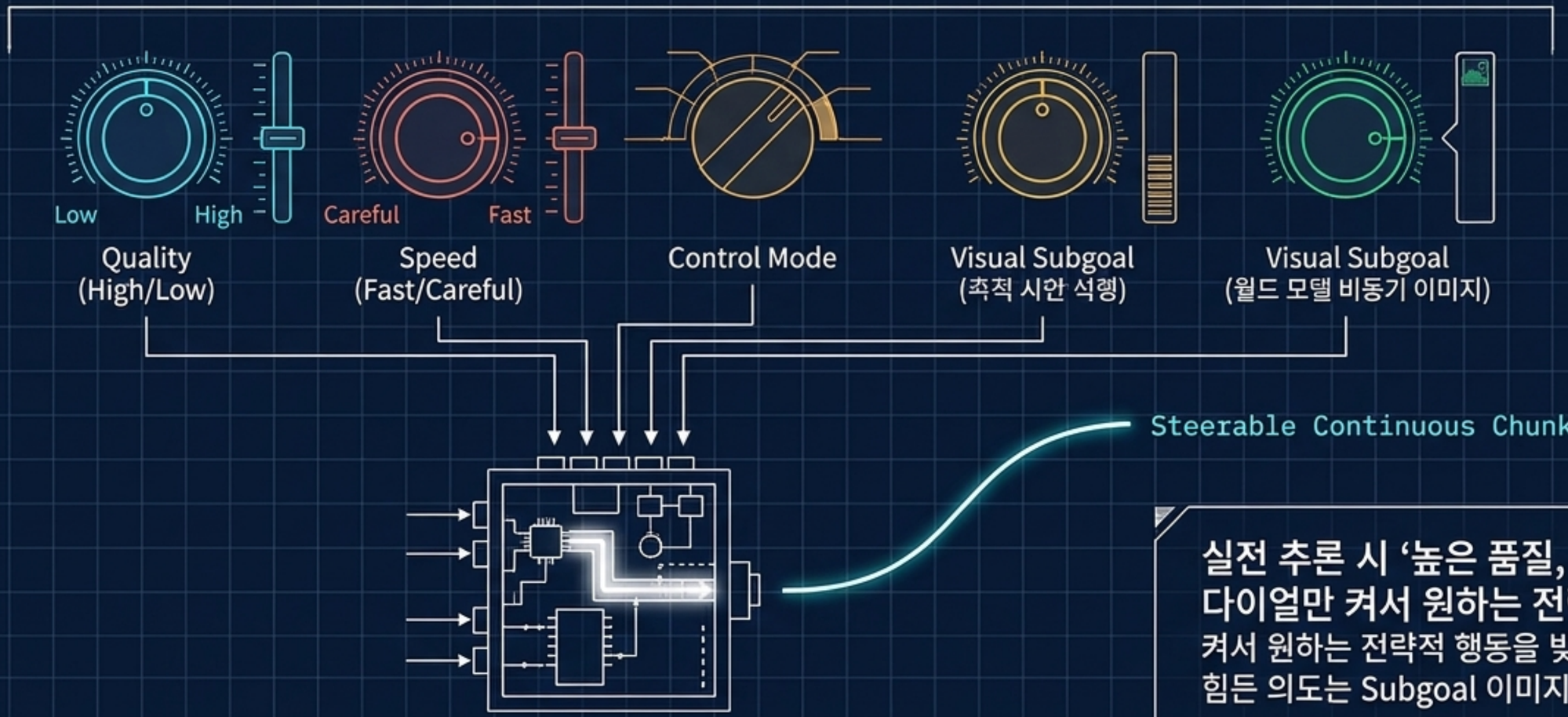


모방학습(BC)의 천장 돌파. 실패 데이터를 버리지 않고 거름으로 사용하여 처리량 2배 향상, 실패율 절반 감소. 낯선 집(Zero-shot 환경)에서의 의미-행동 동시 수행(π0.5).

Gemini Notebook

전방위적 맥락 조건화 (Context Conditioning in π0.7)

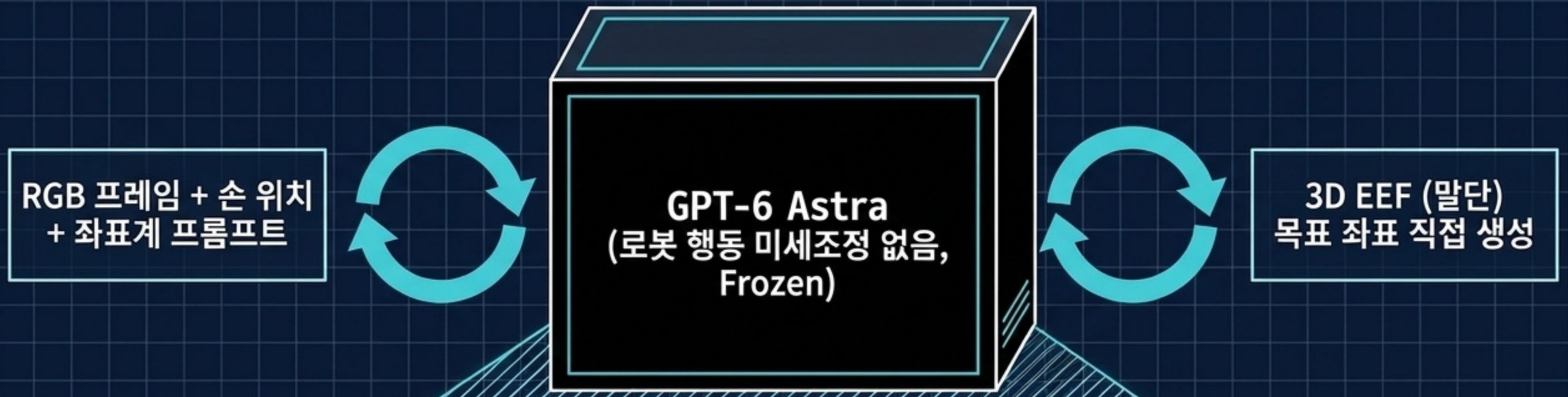
섞여 있는 잡음 데이터(고속, 저속, 실패 시연)를 맥락(Context)으로 격리하여 학습.



실전 추론 시 '높은 품질, 빠른 속도' 다이얼만 켜서 원하는 전략적 행동을 켜서 원하는 전략적 행동을 빚어냄. 말로 하기 힘든 의도는 Subgoal 이미지로 직접 주입.

Gemini Notebook

범용 VLM의 직접 침공: Astra의 충격과 그림자



지연 시간 (Latency): 한 번의 턴(성공 에피소드 기준) 호출에 약 18초 소요.
비용 (Cost): 에피소드당 최대 3,090만 토큰 소진.
환경의 정지: 모델이 생각하는 동안 시뮬레이터의 시간이 멈춰 있어야 함 (동적 환경 반응성 결여). No-CoT의 환상.

Gemini Notebook

텍스트를 버리고 속도를 얻다: Decision Model의 부상

구분	생성형 VLM (Astra)	Decision Model (Gemini D1/d1.2)	연속 Flow VLA ($\pi 0.7$)
출력 형태	자유 텍스트 / CoT 순차 생성	고정 선택지에 대한 보정 확률 (병렬 채점)	$H \times d$ 연속 궤적 행렬 (적분)
속도 (지연)	수 초 ~ 수십 초 (High Latency)	밀리초 (ms) 단위 초고속	50Hz ~ 200Hz 제어 루프
시스템 역할	깊은 추론, 코드 작성, 상위 플래너	System 2 라우터, 검증기(Verifier)	실물 손발을 움직이는 물리 제어기

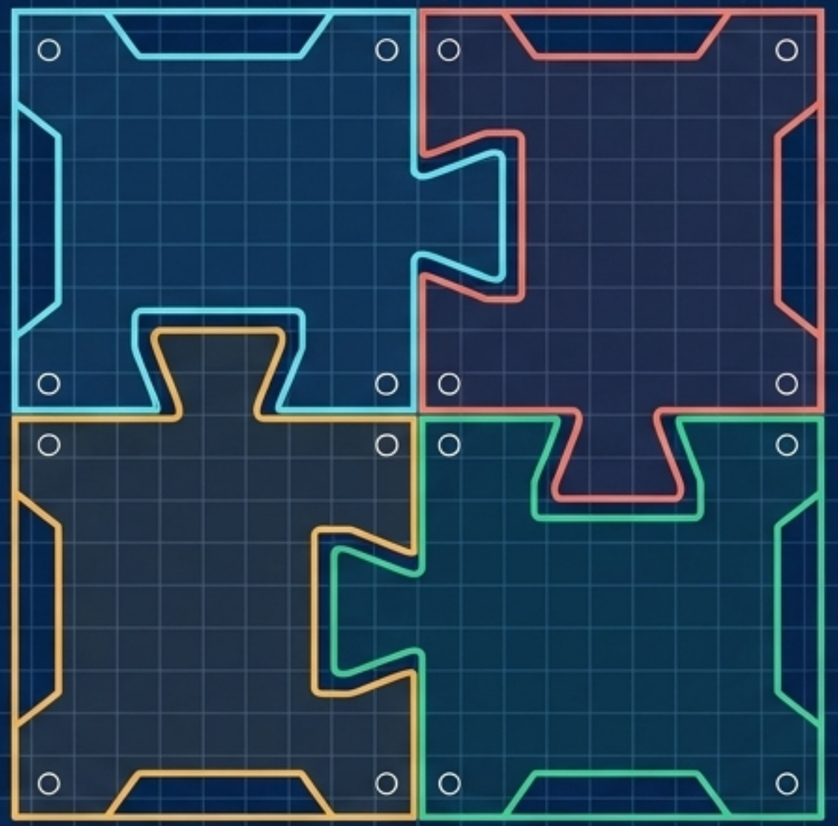
SayCan의 철학(확률 채점)을 계승. 문장 생성을 과감히 생략하여 속도 병목을 파괴함.

Gemini Notebook

공정한 비교를 위한 연구자의 눈 (Evaluation Rigor)

1. 관측 이력 (Memory)

현재 1장 vs 전체 에피소드 이력.
(기억 효과가 모델의 지능으로 착각됨).



2. 행동 인터페이스 (Interface)

관절(Joint) 각도 제어 vs EEF 좌표 + 고성능 Motion Planner.
(실행기의 공을 모델 능력으로 착각).

3. 시간 지연 (Time)

시뮬레이터 정지 유무.
(시간이 멈춘 세계의 성공률 vs 실시간 물리 반응성).

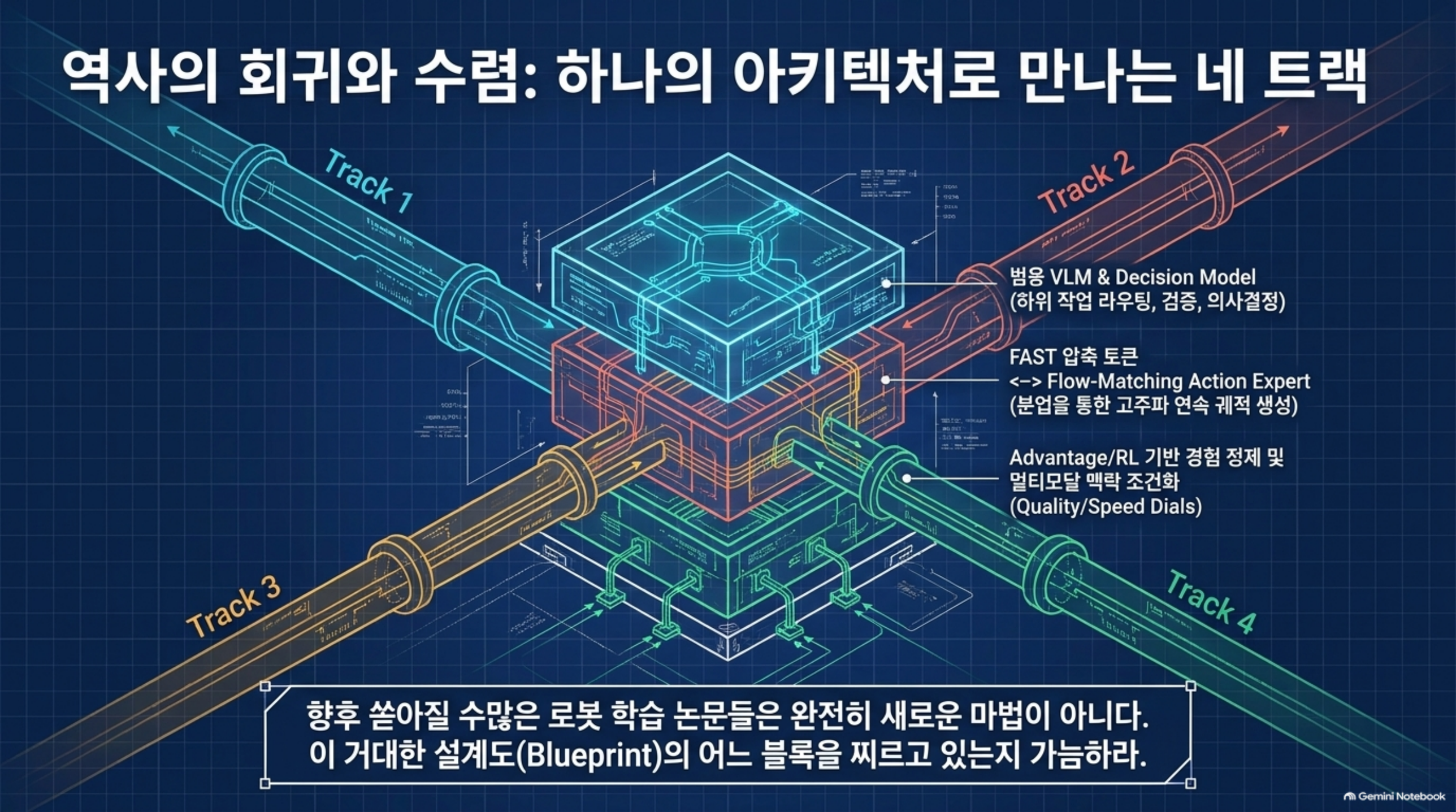
4. 적응 (Adaptation)

프롬프트 엔지니어링 튜닝 시 Test split 참조 여부.

모델 가중치(Weights) 하나가 성패를 가르지 않는다.
인터페이스와 지연 비용을 반드시 정량화하라.

Gemini Notebook

역사의 회귀와 수렴: 하나의 아키텍처로 만나는 네 트랙



향후 쏟아질 수많은 로봇 학습 논문들은 완전히 새로운 마법이 아니다.
이 거대한 설계도(Blueprint)의 어느 블록을 찌르고 있는지 가늠하라.

Gemini Notebook