World Simulation with Video Foundation Models for Physical AI

NVIDIA[1]

2026-2-26

Paper : https://arxiv.org/abs/2402.06196

Artificial Intelligence

Creating the Future

Dong-A University

Division of Computer Engineering & Artificial Intelligence

Summary

- 본 논문은 Text2World, Image2World, Video2World 생성을 통합하고 향상된 ***text grounding*** 및 시뮬레이션 제어를 제공하는 [**Cosmos-Predict2.5**]와 Sim2Real 및 Real2Real 월드 변환을 위한 **control-net** 스타일 프레임워크인 [**Cosmos-Transfer2.5**]를 Physical AI를 위한 최신 World Foundation Models로 소개함

- 🚀 이 모델들은 2억 개 이상의 큐레이션된 비디오 클립으로 학습되고 강화 학습(RL) 기반의 post-training으로 fine-tuning되어 이전 모델인 [Cosmos-Predict1] 및 [Cosmos-Transfer1]에 비해 비디오 품질과 지침 정렬에서 상당한 개선을 이루었으며, 더욱 작고 효율적임

- 🛠️ [Cosmos-Predict2.5] 및 [Cosmos-Transfer2.5]는 로봇 정책 학습을 위한 synthetic data augmentation, 자율 주행 시뮬레이션을 위한 multi-view 세계 생성, 카메라 제어 가능한 multi-view 생성 등 다양한 Physical AI 애플리케이션에서 향상된 성능과 실용적인 이점을 제공함

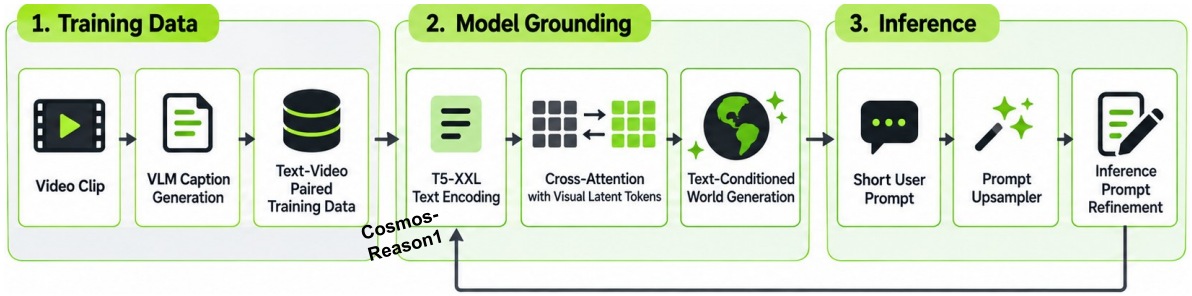
A.1. Core Contributors

- Data Curation
  Jacob Huffman, Francesco Ferroni, Alice Luo, Niket Agarwal, Hao Wang, Jing Zhang, David Page, Vasanth Rao Naik Sabavat, Sriharsha Niverty, Erik Barker, Lindsey Pavao, Stella Shi, Prithvijit Chattopadhyay, Shitao Tang, Yin Cui, Yunhao Ge, Qianli Ma, Yifan Ding, Seungjun Nah, Siddharth Gururani, Jiashu Xu, Grace Lam, Tiffany Cai, Jibin Varghese, Pooya Jannaty, Jay Zhangjie Wu, Yuxuan Zhang, Huan Ling, Hanzi Mao, Heng Wang
- Tokenizer
  Jinwei Gu, Xian Liu, Songwei Ge, Ting-Chun Wang, Haoxiang Wang, Fitsum Reda
- Diffusion-based World Foundation Model Pre-training
  Qinsheng Zhang, Lin Yen-Chen, Xiaohui Zeng, Huan Ling, Shitao Tang, Maciej Bala, Ting-Chun Wang, Yu Zeng, Seungjun Nah, Qianli Ma, Hanzi Mao
- Autoregressive-based World Foundation Model Pre-training
  Haoxiang Wang, Yifan Ding, Xian Liu, Jiaojiao Fan, Xiaohui Zeng, Yogesh Balaji
- Prompt Upsampler
  Yunhao Ge, Haoxiang Wang, Jiashu Xu, Yin Cui
- Diffusion Decoder
  Huan Ling, Jiaojiao Fan, Fitsum Reda, Yogesh Balaji, Hanzi Mao, Qinsheng Zhang
- 3D Consistency Pre-training Evaluation
  Jiahui Huang, Chen-Hsuan Lin
- Physics Alignment Pre-training Evaluation
  Francesco Ferroni, Prithvijit Chattopadhyay, Xinyue Wei, Qianli Ma, Gergely Klár, Chen-Hsuan Lin
- Camera Control Post-training Evaluation
  Xiaohui Zeng, Tsung-Yi Lin, Jingyi Jin, Chen-Hsuan Lin
- Robotics Post-training Evaluation
  Lin Yen-Chen, Wei-Cheng Tseng, Yunhao Ge, Xian Liu, Shitao Tang, Fangyin Wei, Lyne Tchapmi, Yu Zeng, Qingqing Zhao, Yin Cui, Zhaoshuo Li, Jinwei Gu
- Autonomous Driving Post-training Evaluation
  Seung Wook Kim, Jay Zhangjie Wu, Jiahui Huang, Francesco Ferroni, Michele Fenzi, Daniel Dworakowski, Despoina Paschalidou, Ed Schmerling, Shiyi Lan, Laura Leal-Taixe, Sanja Fidler, Huan Ling
- Guardrail
  Jibin Varghese, Arslan Ali, Grace Lam, Pooya Jannaty
- Platform Architect
  Ming-Yu Liu

참고

Text Grounding

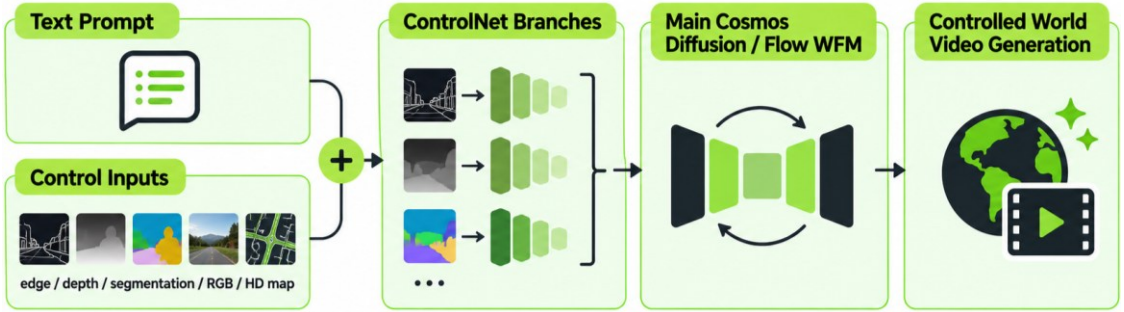
- **Text Grounding**은 AI가 텍스트 프롬프트에 포함된 단어나 문장의 의미를 **시각적, 물리적 요소(객체, 공간적 위치, 행동, 속성 등)**로 정확하게 **매핑하고 연결**하는 능력을 말함. 즉, 텍스트가 시각적 현실(시뮬레이션 세계)에 '뿌리를 내리게(Grounding)'하는 과정임
- **Cosmos의 text grounding**은 "**텍스트를 비디오 latent token 생성 과정에 cross-attention으로 결합하여, 생성되는 world video의 장면·객체·행동·동역학을 의미적으로 제어**하는 방식"임



- **Cosmos-Predict 2.5에서 Text Grounding 역할**:
  - **물리적 정밀도 향상**: 기존의 언어 모델(T5 등)은 텍스트의 언어적 맥락은 잘 이해하지만, 물리적 세계와의 연결은 약했음. NVIDIA는 이를 해결하기 위해 VLM인 **Cosmos-Reason1**을 **텍스트 인코더**로 도입함
  - **환각(Hallucination) 억제**: 예를 들어 *"빨간색 로봇 팔이 파란색 상자를 들어 올려 컨베이어 벨트 위에 놓는다"*라는 텍스트를 입력했을 때, Grounding 능력이 뛰어나면 객체의 색상, 행동의 순서, 객체 간의 물리적 상호작용이 영상 내에 정확히 구현됨.
  - **LWM** 관점: 프롬프트의 지시사항(Instruction)을 시뮬레이션 환경에서 오류 없이 그대로 실행(Alignment)하게 만드는 가장 중요한 기초 체력임

Control-Net 프레임워크

- **Control-Net**은 사전 학습된 대형 생성 모델(Diffusion 등)의 원래 Weight들을 Freeze한 상태에서, 추가적인 **공간적/구조적 조건(Condition)을 주입**하여 생성 결과물의 형태를 완벽하게 제어할 수 있도록 덧붙이는 신경망 아키텍처임
- "텍스트만으로 비디오를 생성하는 WFM"을 "**구조·깊이·의미분할·시각 조건까지 반영하는 controllable world generator**"로 확장



- **Cosmos-Transfer 2.5에서의 Control-Net 역할**:
  - **공간적 제어권 확보**: **Text Grounding이 '의미와 행동'을 제어**한다면, **Control-Net은 '형태와 구조'를 강제**함. **Edge, Depth Map, Segmentation Mask 등을 조건으로 입력**하여, 모델이 마음대로 배경이나 구조를 바꾸지 못하게 기하학적 틀을 고정함
  - **Sim2Real 및 Real2Real 변환**:
    - ✓ **Sim2Real**: 물리 엔진이나 3D 툴에서 렌더링된 Synthetic 레이아웃이나 세그멘테이션 이미지를 Control-Net에 넣으면, 그 구조를 완벽히 유지한 채 실사(Photorealistic) 비디오로 변환함
    - ✓ **Real2Real**: 실제 촬영된 블랙박스 영상에서 날씨(눈, 비)나 시간대(낮, 밤)만 바꾸면서도, 원래 있던 자동차들의 위치와 도로의 곡률은 그대로 유지하게 만듦

❖ Cosmos-Predict2.5 상세 설명

➢ Flow matching 기반의 World Foundation Model로, [Cosmos-Predict1]의 확산 (diffusion) 비디오 월드 모델을 크게 개선함

➢ 주요 개선 사항

- **데이터 처리 개선**: **2억 개 이상의 큐레이션된 비디오 클립**을 사용하여 훈련되었으며, **더 강력한 데이터 필터링 파이프라인과 물리적 AI에 특화된 정교한 Post-Training 데이터**를 사용함. **초기 raw 비디오의 약 4%**만이 필터를 통과하여 고품질의 데이터를 확보함
- **통합 모델 아키텍처**: Text2World, Image2World, Video2World 생성 기능을 단일 모델로 통합함
- **훈련 레시피 개선**: 모델 병합(model merging)과 새로운 강화 학습(RL) 알고리즘을 Post-Training에 활용하여 생성 품질과 지시어 정렬(instruction alignment)을 향상시킴
- **텍스트 인코더 교체**: [Cosmos-Predict1]에서 사용된 T5 텍스트 인코더 대신 물리적 AI에 특화된 디코더 전용(decoder-only) VLM인 [**Cosmos-Reason1**]을 활용하여 더 풍부한 Text Grounding과 세밀한 세계 시뮬레이션 제어를 제공함. [Cosmos-Reason1]은 여러 블록의 activations를 연결하여 1024차원 공간으로 투영함으로써 지역적 및 전역적 언어적 문맥을 더 충실히 포착함
- 모델 스케일: 2B 및 14B 스케일로 모델이 출시

➢ Cosmos-Reason1과 Cosmos-Reason2

구분	Cosmos-Reason1	Cosmos-Reason2
기본 역할	Physical AI용 reasoning VLM, **Cosmos-Predict2.5의 text encoder / text grounding 모듈로 활용**	**Reason1보다 강화된 Physical AI reasoning VLM**
주요 용도	**물리 상식, embodied reasoning, video captioning, prompt 이해, Predict2.5의 텍스트 조건 정렬**	**비디오 분석, 로봇 planning, 자율주행·스마트공간 reasoning, 데이터 annotation**
모델 계열	공개 repo 기준 Cosmos-Reason1-7B 중심, 논문 초록에서는 8B/56B도 언급	Cosmos-Reason2-2B, 8B, 32B 계열
기반 아키텍처	Qwen2.5-VL 기반으로 설명됨	Qwen3-VL 기반으로 설명됨
강화점	**physical common sense + embodied reasoning SFT/RL**	**spatio-temporal reasoning, timestamp precision, long-context, 2D/3D localization 강화**
Context 길이	Reason2 대비 짧음	**최대 256K input tokens 지원으로 설명**
위치/객체 grounding	**주로 장면 이해·물리 상식·행동 추론 중심**	**2D/3D point localization, bounding-box coordinates, reasoning explanation 지원**

Cosmos github

https://github.com/NVIDIA/Cosmos

Pinned

cosmos-predict2.5 Public

Cosmos-Predict2.5, the latest version of the Cosmos World Foundation Models (WFMs) family, specialized for simulating and predicting the future state of the world in the form of video.

Python 1.3k 175

cosmos-transfer2.5 Public

Cosmos-Transfer2.5, built on top of Cosmos-Predict2.5, produces high-quality world simulations conditioned on multiple spatial control inputs.

Python 690 117

cosmos-reason2 Public

Cosmos-Reason2 models understand the physical common sense and generate appropriate embodied decisions in natural language through long chain-of-thought reasoning processes.

Python 423 89

cosmos-cookbook Public

Post-training scripts and samples for NVIDIA Cosmos ecosystem

Python 440 103

- **https://github.com/nvidia-cosmos/cosmos-predict2.5**
- **https://github.com/nvidia-cosmos/cosmos-transfer2.5**
- **https://github.com/nvidia-cosmos/cosmos-reason2**
- **https://github.com/nvidia-cosmos/cosmos-cookbook**

NVIDIA Cosmos 공개코드 4종 구조 요약

Physical AI를 위한 World Foundation Model (WFM) 생태계: Predict 2.5 – Transfer 2.5 – Reason 2 – Cookbook

Cosmos WFM은 cosmos-predict, cosmos-transfer, cosmos-reason의 세 model type으로 구성되며, 모두 post-training으로 커스터마이즈 가능합니다.

Cosmos WFM

Physical AI 데이터
Video / Image / Sensor / Simulation / Action

1 cosmos-predict2.5
Future World Prediction
Text2World + Image2World + Video2World 통합

text | image 또는 video | action | multiview input

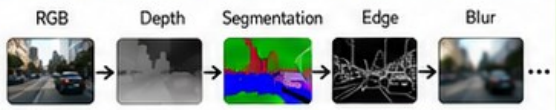


미래 video world

WFM의 기본 예측 엔진

2 cosmos-transfer2.5
Spatial Control Transfer
RGB / Depth / Segmentation / Edge / Blur 기반 변환

RGB | Depth | Segmentation | Edge | Blur



synthetic / transformed video

Sim2Real · Real2Real 데이터 증강

3 cosmos-reason2
Physical Reasoning VLM
상황 이해 · 물리 상식 · 행동 판단

상황 이해 | 물리 상식 | Planning / Decision

image | video | prompt

자연어 reasoning, planning, decision

Physical AI의 판단 두뇌

4 cosmos-cookbook
Recipes / Post-training / Evaluation / Deployment

Recipes | Post-training | Evaluation | Deployment

Cosmos 모델 + 도메인 데이터

workflow, scripts, case studies

실증 적용 매뉴얼

Robotics / AV / Smart Spaces / Industrial Safety / K-Port

- **Robotics** — 지능형 조작·작업
- **AV** — 위험 판단·경로 계획
- **Smart Spaces** — 공간 이해·운영 최적화
- **Industrial Safety** — 위험 감지·안전 조치
- **K-Port** — 자율항만 Physical AI 응용

한 장 요약 표

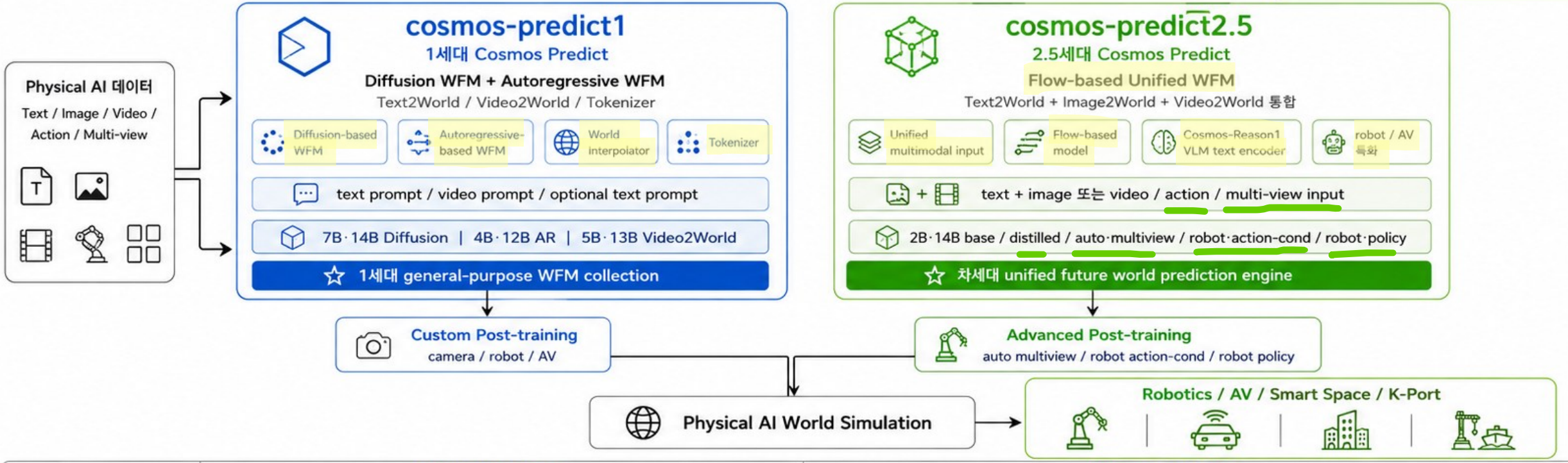
구분	저장소	핵심 역할	주요 입력	주요 출력	Physical AI 의미
1	cosmos-predict2.5	미래 세계 상태 예측·생성	text + image 또는 video, action, multiview input	미래 video world	WFM의 기본 예측 엔진
2	cosmos-transfer2.5	구조적 제어조건 기반 world transfer	RGB, depth, segmentation, edge, blur 등	제어조건을 반영한 synthetic / transformed video	Sim2Real · Real2Real 데이터 증강
3	cosmos-reason2	물리 상식·시공간 추론·행동 판단	image, video, prompt	자연어 reasoning, planning, decision	Physical AI의 판단 두뇌
4	cosmos-cookbook	후학습·평가·배포 레시피	Cosmos 모델 + 도메인 데이터	workflow, scripts, case studies	실증 적용 매뉴얼

✳ **핵심 관계:** Predict2.5는 미래 세계를 예측·생성하는 WFM, Transfer2.5는 구조적 제어조건으로 세계를 변환·증강하는 WFM, Reason2는 물리 상황을 이해하고 판단하는 reasoning VLM, Cookbook은 세 모델을 도메인 후학습·평가·배포 workflow로 연결하는 레시피 저장소입니다.

NVIDIA Cosmos Predict

Physical AI를 위한 World Foundation Model (WFM) 예측 엔진의 진화

현재 상태: Predict1과 Predict2.5는 모두 Cosmos 3로 이전 권고, GitHub README 기준 제한적 유지보수 상태



구분	cosmos-predict1	cosmos-predict2.5
세대	Cosmos Predict 1세대	Cosmos Predict 2.5세대
핵심 정의	Physical AI용 general-purpose WFM collection	video 형태로 미래 세계 상태를 시뮬레이션·예측하는 최신 Predict 계열 WFM
모델 구조	Diffusion-based WFM + Autoregressive-based WFM 분리	Flow-based unified model
주요 task	Text2World, Video2World, WorldInterpolator	Text2World, Image2World, Video2World 통합
텍스트 인코더	diffusion/AR 계열별 별도 구성	Cosmos-Reason1 VLM을 text encoder로 활용
입력	text prompt, video prompt, optional text prompt	text + image 또는 video, action, multi-view input
주요 모델군	7B/14B Diffusion, 4B/12B AR, 5B/13B Video2World, Tokenizers	2B/14B base, distilled, auto/multiview, robot/action-cond, robot/multiview-agibot, robot/policy
핵심 강점	WFM 구조를 diffusion·AR 두 계열로 명확히 제시	품질·prompt alignment 개선, multimodal input 통합, robot/AV 특화 확대
현재 상태	Cosmos 3로 이전 권고, 제한적 유지보수	Cosmos 3로 이전 권고, 제한적 유지보수

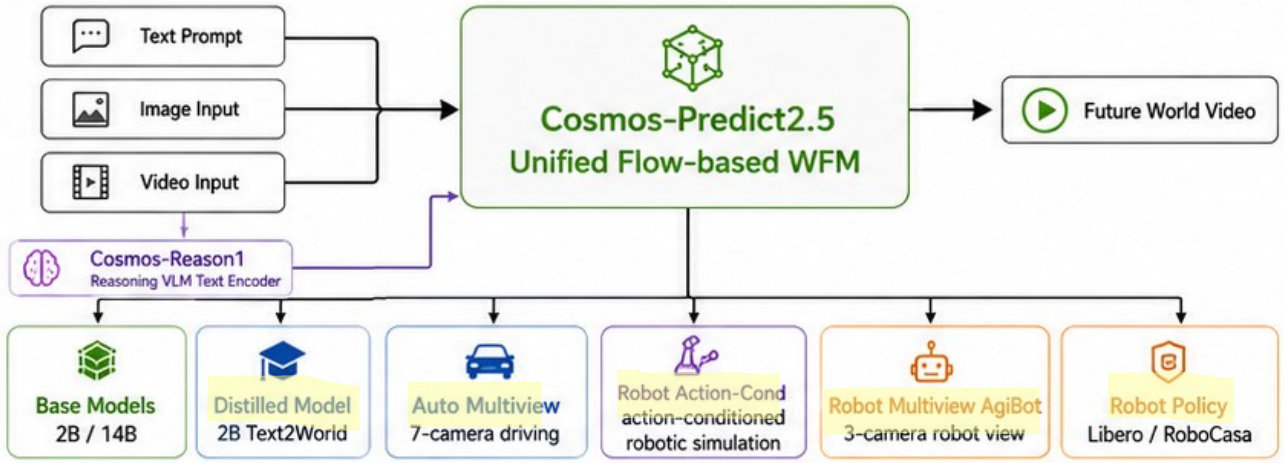
핵심 메시지: Predict1은 Cosmos WFM의 1세대 예측 엔진이고, Predict2.5는 Text2World·Image2World·Video2World를 하나로 통합한 flow-based 차세대 예측 엔진입니다.

NVIDIA Cosmos-Predict2.5 구조 및 기능 확장 요약

Physical AI용 차세대 World Foundation Model (WFM) 예측 엔진

README 기준: Predict2.5는 Predict1의 후속 버전이며, Text2World · Image2World · Video2World를 하나의 flow-based 모델로 통합합니다.

1. cosmos-predict2.5 핵심 구조



Flow-based unified model

Text2World + Image2World + Video2World 통합

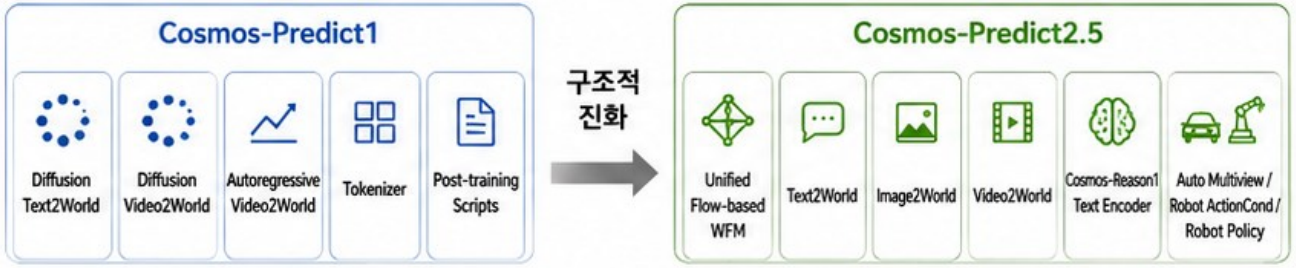
Cosmos-Reason1 text encoder 활용

Quality · Prompt Alignment 개선

2. Predict2.5 모델군

	모델군	Capability	Input
①	Cosmos-Predict2.5-2B/pre-trained	pre-trained base	text + image or video
②	Cosmos-Predict2.5-2B/post-trained	post-trained base	text + image or video
③	Cosmos-Predict2.5-2B/distilled	distilled base	text
④	Cosmos-Predict2.5-14B/pre-trained	pre-trained base	text + image or video
⑤	Cosmos-Predict2.5-14B/post-trained	post-trained base	text + image or video
⑥	Cosmos-Predict2.5-2B/auto/multiview	driving, 7-camera view	text + image or video
⑦	Cosmos-Predict2.5-2B/robot/action-cond	robotic, action-conditioned	action
⑧	Cosmos-Predict2.5-2B/robot/multiview-agibot	robotic, AgiBot data, 3-camera view	text + image
⑨	Cosmos-Predict2.5-2B/robot/policy	Libero · RoboCasa post-trained policy	action + image

3. Predict1 → Predict2.5 구조적 진화



Predict1은 diffusion · autoregressive 기반 1세대 구조, Predict2.5는 멀티모달 입력과 robot·AV 응용을 통합한 차세대 구조

4. 기능별 비교

	기능 질문	Predict1	Predict2.5
	텍스트로 세계를 생성하는가?	가능: Diffusion Text2World	가능: unified Text2World
	이미지를 넣고 미래 video를 생성하는가?	제한적 또는 post-training 중심	핵심 기능: Image2World 통합
	비디오 기반 미래 예측을 하는가?	가능: Diffusion/AR Video2World	가능: unified Video2World
	autoregressive 구조를 별도 제공하는가?	4B/12B base, 5B/13B Video2World	unified flow-based 중심, sliding window generation 등 기능 확장
	tokenizer 연구·활용에 적합한가?	매우 적합: continuous/discrete tokenizer 공개	통합 inference·post-training 활용 중심
	robot action-conditioned simulation이 있는가?	post-training guide로 제공	전용 robot/action-cond 모델군 제공
	autonomous driving multiview가 있는가?	Single2MultiView AV 기능 추가	auto/multiview 7-camera view 특화
	robot policy까지 포함하는가?	명시적 robot policy 모델군은 약함	Libero/RoboCasa post-trained robot/policy 제공
	최신 실무 적용 기준인가?	구조 이해와 legacy 활용	2.5 계열 기준 실무 적용에 더 적합

한 줄 요약

cosmos-predict1은 diffusion · autoregressive 기반 WFM의 원형을 제공한 1세대 미래 세계 예측 엔진이고,
cosmos-predict2.5는 Text2World · Image2World · Video2World를 flow-based 모델로 통합하면서 robot action-conditioned simulation, AV multiview, robot policy까지 확장한 Physical AI용 차세대 world prediction 코드베이스입니다.

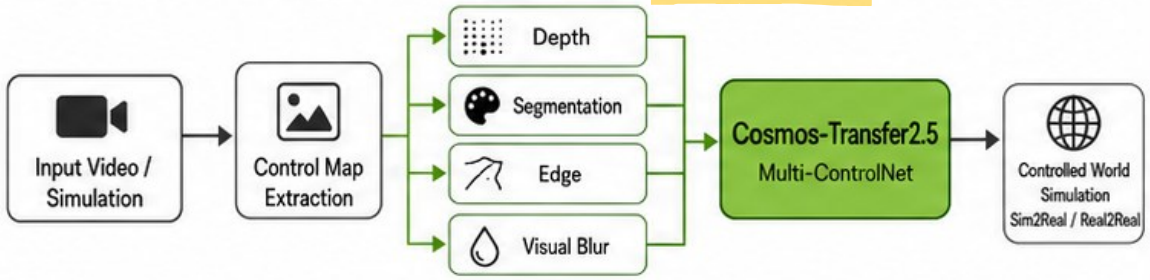
NVIDIA Cosmos-Transfer2.5 & Cosmos-Reason2 구조 요약

Physical AI를 위한 구조 제어 기반 세계 변환·증강 엔진과 물리 상식 기반 Reasoning VLM

Cosmos 생태계에서 Transfer2.5는 제어 기반 world transfer / augmentation, Reason2는 physical reasoning / planning 역할을 담당

1. cosmos-transfer2.5: 구조 제어 기반 세계 변환·증강 엔진

Cosmos-Predict2.5 위에서 동작하는 Multi-ControlNet



Transfer2.5의 주요 기능

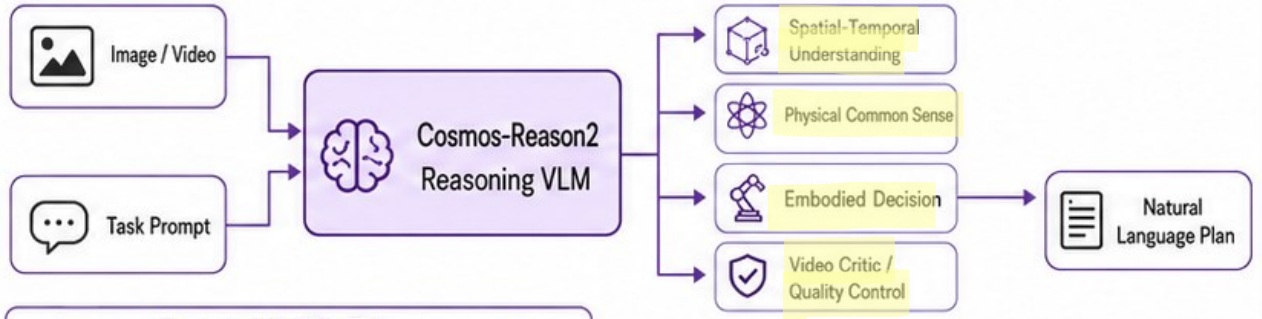
1. Multi-control input — RGB, depth, segmentation, edge, blur 등 구조 조건 입력
2. Sim2Real augmentation — 시뮬레이터 영상을 현실감 있는 영상으로 변환
3. Real2Real augmentation — 실제 영상의 날씨·조명·배경·조건 다양화
4. Auto Multiview — 자율주행 multi-camera 환경 지원
5. Robot multiview-control — robot multiview 영상에 depth, edge, visual blur, segmentation 제어 작용
6. Distilled Edge — 저지연 edge inference용 distilled 모델 공개

Transfer2.5 모델 패밀리

- General 2B checkpoint
- AV용 auto checkpoint
- Robot multiview-control checkpoint — 지원 control type: depth / edge / visual blur / segmentation

2. cosmos-reason2: 물리 상식 기반 Reasoning VLM

Physical AI와 Robotics를 위한 Open Customizable Reasoning VLM



Reason2의 주요 특징

1. 모델 성격 — Physical AI용 reasoning VLM
2. 핵심 능력 — spatial-temporal understanding, physics understanding, common sense
3. 학습 방식 — physical common sense 및 embodied reasoning data로 post-training, SFT + RL
4. reasoning 방식 — long chain-of-thought reasoning으로 world dynamics 이해
5. 모델군 — 2B / 8B / 32B
6. 실행 환경 — H100, GB200, DGX Spark, Jetson AGX Thor 등에서 예시 검증

Reason2 모델군 및 실행 힌트

2B | 8B | 32B

최소 GPU 메모리 안내: 2B는 24GB, 8B는 32GB

Physical AI 관점의 역할 분담

Transfer2.5 = 세계 변환·증강 엔진
structured control, sim2real, real2real, multiview, controllable simulation

Reason2 = 물리 상황 이해·판단 두뇌
space-time understanding, physics common sense, embodied reasoning, planning, quality control

주요 응용

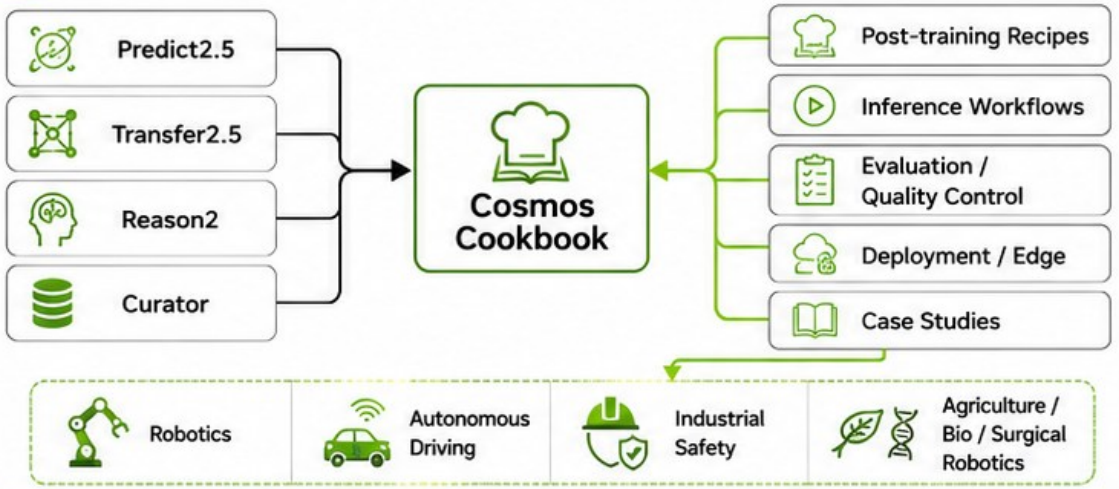
Robotics | Autonomous Vehicles | Smart Spaces | Industrial Safety | K-Port

한 줄 요약: Transfer2.5는 구조 조건으로 세계를 변환·증강하는 controllable world simulation 엔진이고, Reason2는 물리 상식과 시공간 이해를 바탕으로 다음 행동을 판단하는 Physical AI용 reasoning VLM입니다.

NVIDIA Cosmos-Cookbook 및 공개코드 기능 차이 요약

Physical AI 실무 적용 레시피 저장소와 Predict2.5 기준 기능 분담

1 cosmos-cookbook: 실무 적용 레시피 저장소



cosmos-cookbook은 모델 자체보다, Cosmos 생태계를 실제 도메인에 적용하기 위한 step-by-step workflow, case study, technical recipe 저장소입니다. Robotics, simulation, autonomous systems, physical scene understanding을 위한 real-world domain-specific application 가이드를 제공합니다.

2 Cookbook의 최신 레시피 흐름

A. 대표 사례

- Transfer2.5 기반 농업 로봇 perception용 photorealistic synthetic image 생성
- Reason2의 Jetson Thor edge VLM deployment
- Curator용 Beamr CABR video compression
- Predict2.5 기반 surgical simulator post-training
- GR00T-Dreams synthetic robot trajectory generation

B. Predict 계열 레시피

- future state prediction / generation
- robotic manipulation fine-tuning
- traffic anomaly generation
- humanoid robot trajectory generation
- sports video generation
- surgical simulator fine-tuning
- Predict2.5 distillation

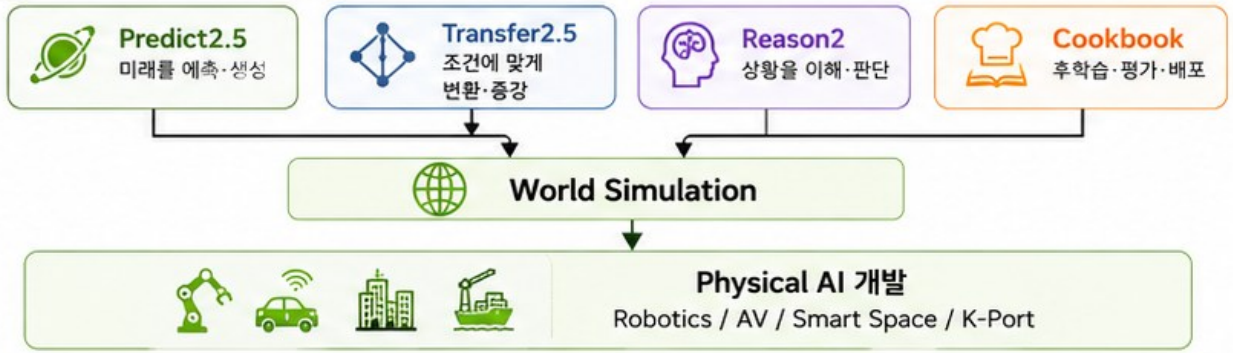
C. Transfer 계열 레시피

- edge / depth / segmentation / visual control
- CARLA Sim2Real
- weather augmentation
- warehouse simulation
- multiview AV generation
- agriculture Sim2Real

D. Reason 계열 레시피

- prompt guide
- video search / summarization
- industrial safety
- egocentric social reasoning
- Jetson Thor edge deployment
- AV captioning / VQA

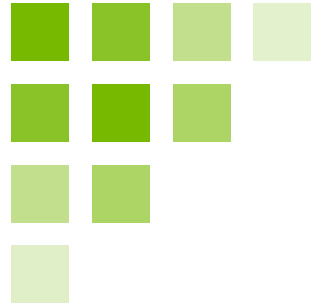
3 Predict2.5 기준으로 본 4개 코드의 기능 차이



질문별 가장 적합한 코드

	질문	가장 적합한 코드
1	현재 영상·이미지·텍스트로 미래 상황을 생성하고 싶은가?	cosmos-predict2.5
2	depth, segmentation, edge, blur 등 구조 조건으로 현실감 있는 영상을 만들고 싶은가?	cosmos-transfer2.5
3	영상·이미지를 보고 물리적 상황을 이해하고 다음 행동을 판단하고 싶은가?	cosmos-reason2
4	모델을 내 도메인에 후학습하고 평가·배포 workflow가 필요한가?	cosmos-cookbook

핵심 메시지: Predict2.5는 미래 세계를 예측·생성하고, Transfer2.5는 구조 조건에 맞게 세계를 변환·증강하며, Reason2는 물리 상황을 이해·판단하고, Cookbook은 이를 실제 도메인에 후학습·평가·배포하는 실무 레시피 저장소입니다.



NVIDIA COSMOS · WORLD FOUNDATION MODEL PLATFORM

Chapter 1

Introduction

Why Physical AI must learn in a digital world first — and how a large-scale, pre-trained World Foundation Model becomes the safe, general substrate for that learning.

Cosmos Introduction

1. Introduction 핵심 정리

NVIDIA

Physical AI용 World Foundation Model Platform의 출발점

1 핵심 메시지

- Physical AI는 센서와 액추에이터를 통해 현실 세계를 관측하고 변화시키는 AI 시스템
- 문제는 현실 세계에서 안전하게 시행착오하며 학습하기 어렵다는 점
- 이를 해결하기 위한 물리 세계의 디지털트윈, 즉 World Foundation Model(WFM)
- Cosmos는 Physical AI 개발자를 위한 WFM 플랫폼을 제안

핵심: Physical AI는 현실에 투입되기 전에 디지털 세계에서 먼저 학습·검증되어야 한다.

2 왜 WFM이 필요한가?

Observation → Action → Perturbation → Future State

현실 세계 학습의 병목

- interleaved observation-action 데이터 필요
- 초기 exploratory action은 위험·비용이 큼
- 현실 데이터 스케일링이 매우 어려움

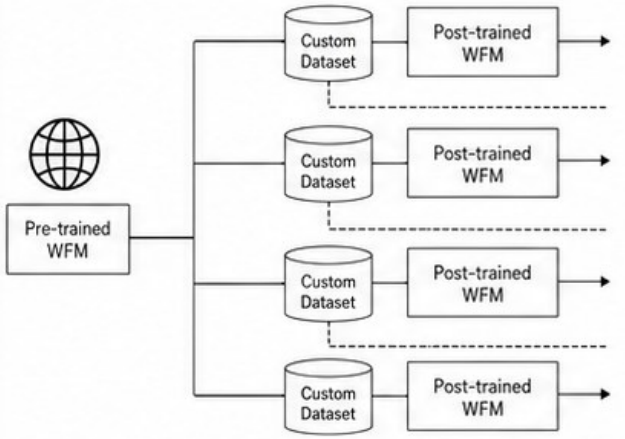
WFM = Physical AI가 안전하게 상호작용할 수 있는 물리 세계의 디지털 트윈

3 논문 Fig. 1: WFM 예시 결과



Pre-trained WFM과 post-trained WFM의 대표 예시:
diffusion / autoregressive / camera control / robotic manipulation

4 Pre-training → Post-training 패러다임



Pre-trained WFM은 generalist, post-trained WFM은 target Physical AI setup에 맞춘 specialist

- 대규모 비디오 데이터로 범용 WFM 학습
- 소규모 custom dataset으로 특정 환경에 맞게 fine-tuning
- action, trajectory, instruction 등이 prompt 역할

5 Introduction이 제시하는 플랫폼 로드맵

Data Curation	Tokenizer	Pre-trained WFM	Post-trained WFM	Guardrail
고품질·고동역학 비디오 확보	비디오를 compact token으로 압축	diffusion + autoregressive 기반 generalist	camera / robot / driving 특화	harmful input/output 차단

6 핵심 기술 포인트

A. 데이터	B. 모델	C. 응용
• 2,000만 시간 규모 raw video 기반 • rich dynamics + high visual quality 강조	• Diffusion WFM: continuous token • Autoregressive WFM: discrete token	• Camera Control • Robotic Manipulation • Autonomous Driving

7 Physical AI 관점의 의미

안전한 학습	효율적 확장	개발 플랫폼화
현실 대신 디지털 세계에서 rollout	generalist WFM + specialist post-training	data, model, tokenizer, safety를 통합

8 K-Port 시사점



- 자율항만 장비도 현실에서 직접 시행착오 학습이 어려움
- 항만 WFM에서 QC / YC / AGV action을 먼저 시뮬레이션 가능
- "범용 WFM -> 항만 custom dataset -> 장비별 post-trained WM" 구조로 확장 가능

한 줄 결론: 1장 Introduction은 Physical AI의 핵심 병목을 현실 데이터 스케일링과 안전한 시행착오의 어려움으로 정의하고, 이를 해결하기 위한 Cosmos World Foundation Model Platform의 전체 철학과 개발 로드맵을 제시한다.

Train in the digital world before the real one

Physical AI = an AI system with sensors and actuators: it observes the world and acts on it. But acting in the real world is costly and dangerous to learn in directly.

Real-world trial-and-error

Early exploratory actions can damage equipment, the environment, and people

데이터 스케일링이 가장 큰 문제임; Physical AI는 관측과 행동이 얽힌 sequence data가 필요함. 특히 Physical AI의 행동은 실제 물리 세계를 perturbation하며, 초기 탐색 행동은 장비와 환경에 심각한 손상을 줄 수 있음

WFM as a digital twin

a safe world model the agent interacts with instead of reality

Physical AI가 안전하게 상호작용할 수 있는 물리 세계의 디지털 트윈, 즉 World Foundation Model, WFM이 데이터 부족과 안전 문제를 해결할 핵심 수단으로 제시

Deploy only what's verified

learn, evaluate, and plan in the WFM; put validated policy to the field

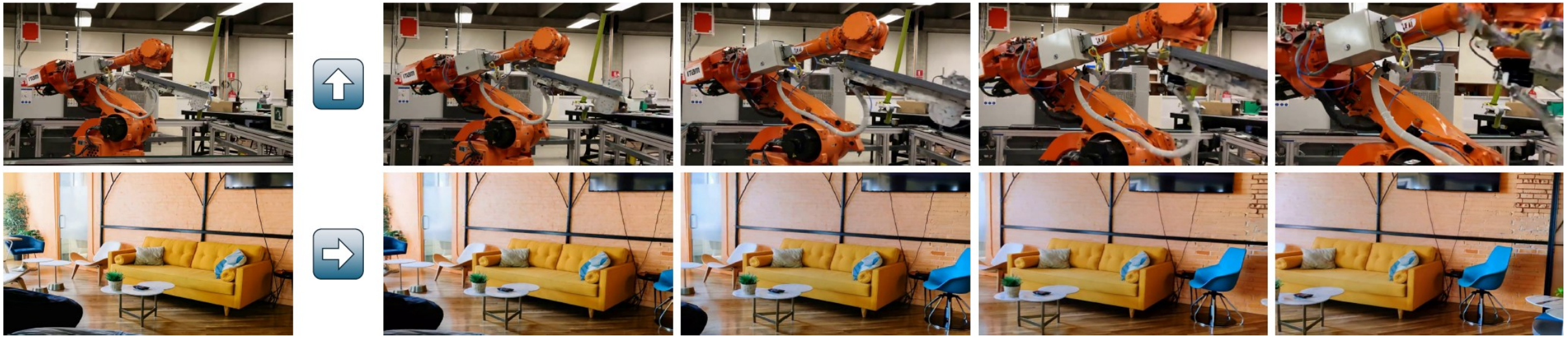
One-line thesis: Cosmos presents a general-purpose World Foundation Model platform so Physical AI can be **safely trained, validated, and planned in a digital world** before real-world deployment.



Pre-training: Diffusion WFM



Pre-training: Autoregressive WFM



Post-training: Camera Control



Post-training: Robotic Manipulation



Post-training: Autonomous Driving

Physical AI의 근본 문제: 현실 세계 데이터 스케일링의 어려움

Unlike text or images, Physical-AI training data is **time-coupled observation–action sequences** that must be collected from real robots, vehicles, and equipment.

1
Observation
센서가 관측한 현실 세계 상태

→

2
Action
AI 또는 로봇 (에이전트)이 수행한 행동

→

3
Perturbation
행동이 물리 세계에 미치는 변화

→

4
Future state
행동 이후 변화된 세계 상태

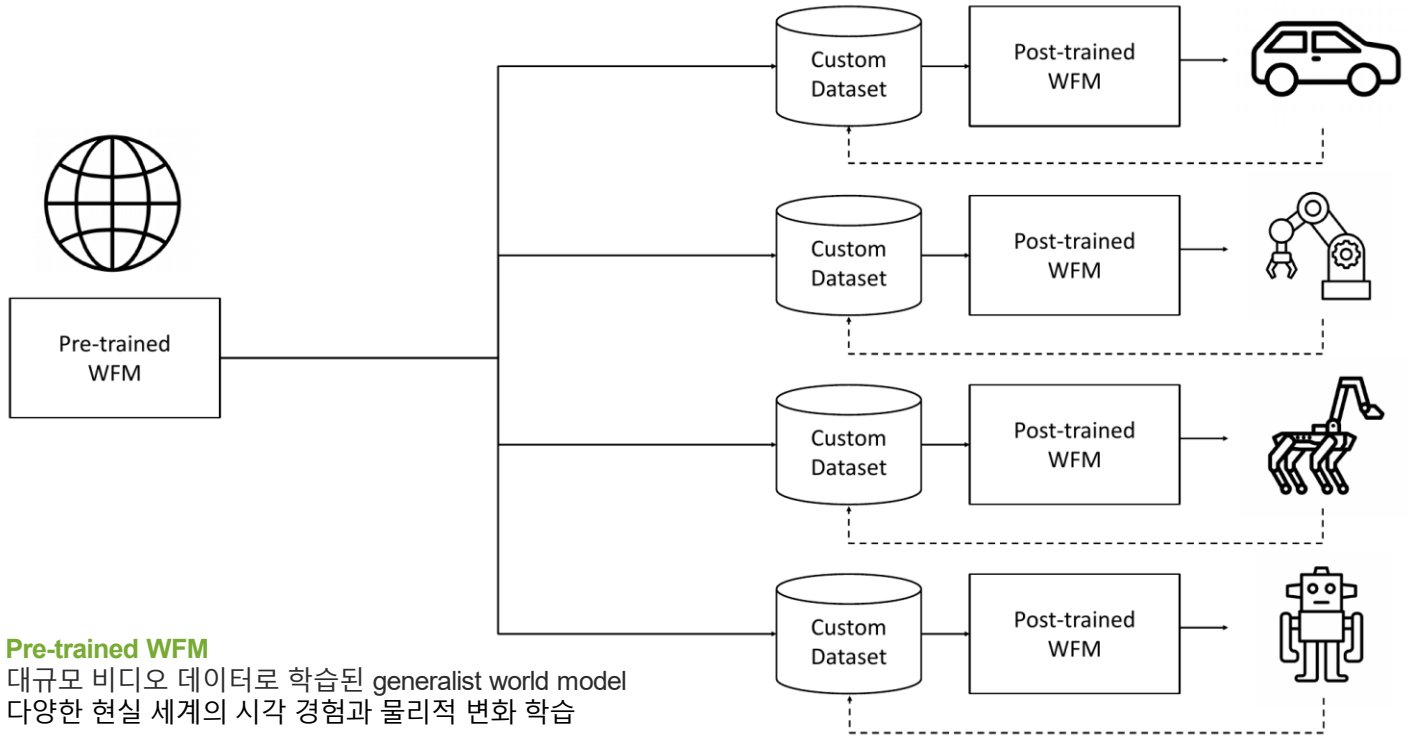
training tuple

$$(x_t,\ a_t,\ x_{t+1})$$

Why it's hard: Collection is expensive and risky; exploratory actions can cause collisions, equipment damage, and safety incidents — repeated experimentation is impractical.

Cosmos' answer: a WFM acts as the safe world the agent interacts with in place of reality — turning a data-collection-and-safety bottleneck into a model-rollout problem.

Cosmos WFM Platform의 기본 방향 : “Pre-train a generalist, post-train a specialist”



Pre-trained WFM
대규모 비디오 데이터로 학습된 generalist world model
다양한 현실 세계의 시각 경험과 물리적 변화 학습

Custom Dataset
목표 Physical AI setup에서 수집된 prompt-video pair
action, trajectory, instruction 등을 조건 입력으로 사용

Post-trained WFM
특정 Physical AI 환경 데이터로 fine-tuning된 specialist world model
camera, robot, autonomous driving 등 도메인 특화

Fig. 2 — one pre-trained WFM → many post-trained specialists; dashed lines are the data loop.

WFM, formally

$$\hat{x}_{t+1} = \mathcal{W}(x_{0:t},\ c_t)$$

predict the next state from past observations and the current perturbation (action / trajectory / instruction / text).

$x_{0:t}$ be a sequence of visual observations of the real world from time 0 to t.
c_t be the perturbation to the world

Why this works

- **Pre-trained WFM** — generalist from large-scale video
- **Post-trained WFM** — specialized per Physical-AI setup
- **Custom dataset** — small prompt–video pairs suffice

Data Curation : WFM 성능의 상한을 결정하는 요소

A good WFM needs not just many videos, but clips **rich in dynamics** and **high in visual quality** — built by a **Physical-AI "data factory."**

~20M hrs	~100M	2–60 s	per 256 frames
raw video collection	Video clips extracted	per clip length	VLM caption granularity

raw video → **high-dynamics clips** → **high visual quality** → **VLM captions**

- WFM generalist 학습의 원천
- 물리 세계의 변화와 상호작용 학습
- 생성 품질과 세계표현 품질 확보
- Text2World / Video2World 조건 학습 기반

Data factory: to make this tractable at scale, Cosmos uses GPU H.264 hardware encode/decode and a Ray-based orchestration pipeline — already implying a Physical-AI Data Factory in Chapter 1.

Two scalable WFM families - Cosmos builds both

Both decompose hard video generation into easier sub-problems and use scalable transformer architectures.

Diffusion WFM

denoising · continuous token

Generates video by gradually removing noise from a Gaussian-noise video.

Strength: highest visual quality.

- 두 접근 모두 복잡한 video generation 문제를 더 작은 sub-problem으로 분해함; Diffusion 모델은 noise 제거 문제로, autoregressive 모델은 next-token prediction 문제로 변환함
- 두 계열 모두 **scalable transformer architecture**를 사용

Autoregressive WFM

next-token · discrete token

Generates video piece-by-piece, conditioned on past generations in a preset order.

Strength: reuses LLM-style scaling & inference optimization.

- Cosmos가 diffusion과 autoregressive 중 하나만 선택하지 않는다는 것임. Diffusion WFM은 고품질 생성에 강하고, autoregressive WFM은 LLM식 scaling과 inference 최적화에 강함
- 따라서 Cosmos는 두 계열을 모두 플랫폼 안에 포함함

Cosmos picks neither exclusively — **it includes both families in one platform**, pairing **diffusion's quality** with **autoregression's scalability**.

Tokenizer : Compress video into a learnable representation

Raw pixels are high-dimensional and redundant — training cost grows with token count. Tokenization compresses video into compact tokens, much like building a video codec.

Continuous tokenizer

Vector-valued tokens → **Diffusion WFM.**

$$z_t \in \mathbb{R}^d$$

Discrete tokenizer

Integer-valued tokens → **Autoregressive WFM.**

$$q_t \in \{1, \ldots, K\}$$

More than preprocessing: the tokenizer defines the WFM's world-representation space. Cosmos builds both continuous and discrete tokenizers to match its two WFM families.

Post-training: Physical AI 도메인 특화 (From generator to conditional simulator)

Fine-tuning the pre-trained WFM on domain prompts turns it into an action-conditioned simulator for three flagship applications.

Camera Control

조건 입력 **camera pose**

사용자가 virtual viewpoint를 움직이며 탐색 가능한 세계 생성

Robotic Manipulation

조건 입력 **robot action / instruction**

로봇 행동 이후 미래 상태 예측

Autonomous Driving

조건 입력 **trajectory / multi-view**

self-driving agents 위한 world simulation

Camera pose conditional WFM을 통해 사용자가 생성한 세계를 자유롭게 탐색할 수 있고, 로봇 task에서는 action에 따른 미래 세계 상태를 더 잘 예측할 수 있으며, autonomous driving task에도 fine-tuning할 수 있음

→ **The shift:** **Conditioning** makes the **WFM** a **simulator**, not just a video generator — the basis for **policy evaluation**, **planning**, and **synthetic data** (detailed in §6).

Guardrails and open models

Guardrail system

- **pre-Guard** — blocks harmful inputs before generation
- **post-Guard** — blocks harmful outputs after generation

Detailed in §7.

Open model release

Pre-trained WFMs and tokenizers are released under the **NVIDIA Open Model License** to support Physical-AI builders.

Honest limitation: the paper makes several design improvements but is explicit that the world-foundation-model problem is **not solved** — substantial further research is required. A useful anchor for our own critical reading.

Chapter Organization

Each later chapter expands a piece of the development flow introduced here.

Chapter	Role introduced in §1
2 · WFM Platform	Define the WFM concept and Cosmos platform structure
3 · Data Curation	Build the large-scale, high-quality video data factory
4 · Tokenizer	Compress video into compact continuous / discrete tokens
5 · Pre-training	Train the diffusion and autoregressive WFMs
6 · Post-training	Specialize for camera, robot, and driving
7 · Guardrails	Secure input and output safety
8 · Related Work	Position Cosmos against prior research

Physical AI 관점의 해석

패러다임의 전환 : Physical AI는 현실에서 바로 학습하는 것이 아니라, 먼저 WFM 기반 디지털 세계에서 학습·검증·계획되어야 한다.

Conventional approach	Cosmos WFM approach
Trial-and-error in the real environment	**Safe rollout in a digital world model**
Hand-built per-domain simulators	**Generalist WFM from large-scale video pre-training**
Task-specific single models	**Post-training to many Physical-AI setups**
A single generative model	**Curation + tokenizer + WFM + post-training + guardrail platform**

Read it as: not a video-generation paper but a platform paper that aims to change how Physical AI is developed.

Why §1 maps directly to autonomous ports

항만장비 Physical AI도 실제 현장에서 시행착오를 통해 학습하기 어렵습니다. QC, YC, AGV, 야드트럭, 작업자, 컨테이너, 선박이 동시에 움직이는 환경에서 잘못된 action은 충돌, 작업 중단, 인명 위험, 장비 파손으로 이어질 수 있습니다.

Cosmos §1 concept	K-Port extension
Physical AI	QC, YC, AGV, 항만장비 에이전트
Observation	CCTV, 장비 카메라, LiDAR, RTK-GPS, TOS 로그
Perturbation	장비 action, 작업지시, trajectory, disturbance
Pre-trained WFM	Port general physical/operations-world generalist
Post-trained WFM	QC WM, YC WM, AGV WM 등 항만 장비별 world model
Guardrail	작업자 안전, 충돌 위험, 위험물, 보안구역 통제

Core logic: 실제 항만에서 위험한 trial-and-error를 줄이고, 항만 WFM에서 장비 action과 작업 시나리오를 먼저 rollout한 뒤, 안전하고 검증된 policy만 현장에 적용한다..

ONE-SENTENCE SUMMARY

Chapter 1 frames Physical AI's core bottleneck as **real-world data scaling a nd the danger of trial-and-error**, and proposes the Cosmos platform — **larg e-scale pre-trained WFMs, then post-training per domain** — as the safe di gital substrate for Physical-AI development.

Problem
unsafe real-world learning

Solution
WFM digital twin

Platform
pre-train → post-train

NVIDIA COSMOS · WORLD FOUNDATION MODEL PLATFORM

Chapter 2

World Foundation Model Platform

WFM의 형식적 정의 · 의도된 용도(Future Cosmos) · 실제 구성요소(Current Cosmos)

arXiv:2501.03575v3 · §2 · Study & presentation deck

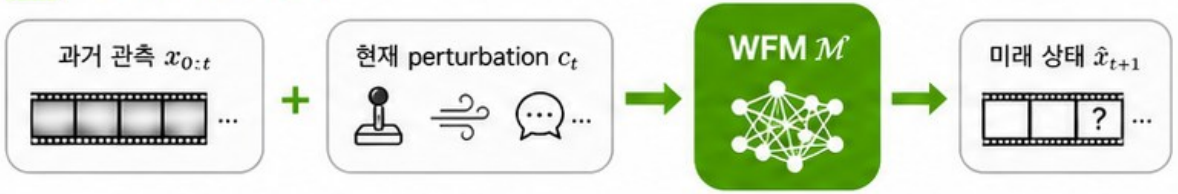
Cosmos World Foundation Model Platform

2. Cosmos WFM Platform 핵심정리

NVIDIA

Physical AI용 COSMOS WFM 개념, 활용, 플랫폼 구조

1 1. WFM의 정의



- 입력: 과거 RGB 비디오 관측과 현재 개입 조건
- perturbation: action, 외란, 택스트 명령 등
- 출력: 다음 시점의 세계 상태 예측

WFM은 현재 세계와 개입 조건을 바탕으로 미래 세계를 예측하는 범용 세계모델

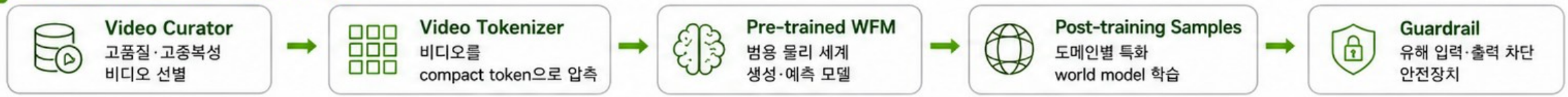
2 2. Future Cosmos: WFM의 활용 가능성

Policy Evaluation:	→	실제 정책 대신 가상 세계에서 정책 평가
Policy Initialization:	→	정책모델의 초기화 기반 제공
Policy Training:	→	reward model과 결합한 강화학습 환경
Planning / MPC:	→	여러 행동 시나리오를 예측해 최적 행동 선택
Synthetic Data Generation:	→	Sim2Real을 합성데이터 생성

MPC: Model Predictive Control (모델 예측 제어)

핵심 의미: 위험한 현실 실험을 줄이고, 더 빠르고 안전한 학습·검증 가능

3 3. Current Cosmos Platform 구성



Cosmos는 단일 모델이 아니라 데이터·토크나이저·사전학습·후학습·안전장치가 결합된 플랫폼

4 4. 개발 패러다임



Camera Control | Robotic Manipulation | Autonomous Driving

Pre-training + Post-training 구조를 통해 범용 모델을 특정 환경에 빠르게 특화

5 5. 2장의 핵심 결론

- Cosmos는 Physical AI를 World Foundation Model 플랫폼이다
- WFM은 과거 관측과 개입 조건으로 미래 세계를 예측한다
- 정책평가·학습·계획·합성데이터 생성의 기반이 된다
- 핵심 구성요소는 Curator, Tokenizer, Pre-trained WFM, Post-training, Guardrail 이다

한 줄 요약: Cosmos는 Physical AI가 안전하게 학습·검증·계획할 수 있도록 물리 세계를 예측하는 플랫폼이다.

WFM이란 무엇인가

World Foundation Model $\mathcal{W}$ 는 과거 시각 관측 $x_{0:t}$ 와 현재 perturbation c_t (행동·명령·외란 등)로부터 다음 관측 $\hat{x}_{t+1}$ 을 예측·생성하는 모델



c_t 는 다양한 형태가 될 수 있음

- Physical AI의 action
- Random perturbation
- Test Description

Figure 3 — WFM은 과거 관측과 현재 perturbation으로 다음 상태를 생성한다;

- Cosmos WFM은 단순히 "동영상을 만드는 모델"이 아니라,현재 세계 상태와 개입 조건을 바탕으로, 물리 세계의 다음 상태를 예측하는 범용 월드모델임
- c_t는 action, random perturbation, text description 등으로 확장 가능함. 따라서 WFM은 행동 조건부 미래 예측 모델, 텍스트 조건부 세계 생성 모델, 물리 환경 시뮬레이션 모델의 성격을 동시에 가짐

비판적 메모 c_t 를 action·random·text로 한데 묶으면, action-conditioned MDP(Markov Decision Process) 시뮬레이터와 text-conditioned 비디오 생성을 동일시하게 된다. 텍스트 프롬프트는 제어 신호가 아니다 — planning/RL 인터페이스로는 곧장 성립하지 않음.

Future Cosmos : Cosmos WFM이 Physical AI 개발자에 유용한 5가지 용도

저자들이 제시하는 "의도된" 효용 — 본 논문에는 이에 대한 실증 결과가 없음 (모두 future work)

Policy Evaluation

실세계 배포 대신 디지털 복제본을 WFM과 상호작용시켜 **정책 품질 평가**. **저비용·고속**, **미경험 환경**에서도 부적합 정책을 빠르게 선별

Policy Initialization

세계의 동역학 패턴을 학습한 WFM을 Policy Model의 좋은 초기화로 사용. Physical AI의 데이터 희소성 문제 완화

- 처음부터 policy를 학습하는 것이 아니라,세계의 물리적 변화 패턴을 이미 학습한 WFM을 기반으로 policy를 출발시킴

Policy Training

WFM + reward model을 물리 세계의 proxy로 삼아 RL 피드백 제공. Physical AI 에이전트가 WFM과 상호작용하며 과제 숙련.

- 현실에서 위험한 trial-and-error를 수행하지 않고, WFM 내부에서 대규모 반복 학습을 수행

Planning / Model-Predictive Control (MDP)

여러 action sequence의 다양한 미래 상태를 시뮬레이션, **cost/reward**로 평가 후 최적 시퀀스 실행 (receding horizon).

- WFM의 성능이 높을수록 Physical AI의 planning, control, decision-making 성능도 높아짐
- WFM은 단순 생성모델이 아니라 **의사결정 시스템의 기반 모델**

Synthetic Data Gen.

학습용 합성 데이터 생성. depth·semantic map 등 rendering metadata 조건화되도록 Finetuning함. 이를 Sim2Real에 활용.

- WFM은 실제 데이터를 **보완하는 Physical AI Data Factory의 핵심 생성 엔진** 역할함

① 상한(Upper-bound) 명제

"The accuracy of the world model upper-bounds the performance of these decision-making strategies."

함의: WFM의 동역학 정확도가 모든 하류 의사결정 성능의 천장이다. 그러나 2장은 그 '**동역학 정확도**'를 무엇으로 측정할지 일절 언급하지 않는다 — 뒤 장의 PSNR/SSIM/rFVD 같은 픽셀 메트릭은 동역학 정확도를 보장하지 못한다.

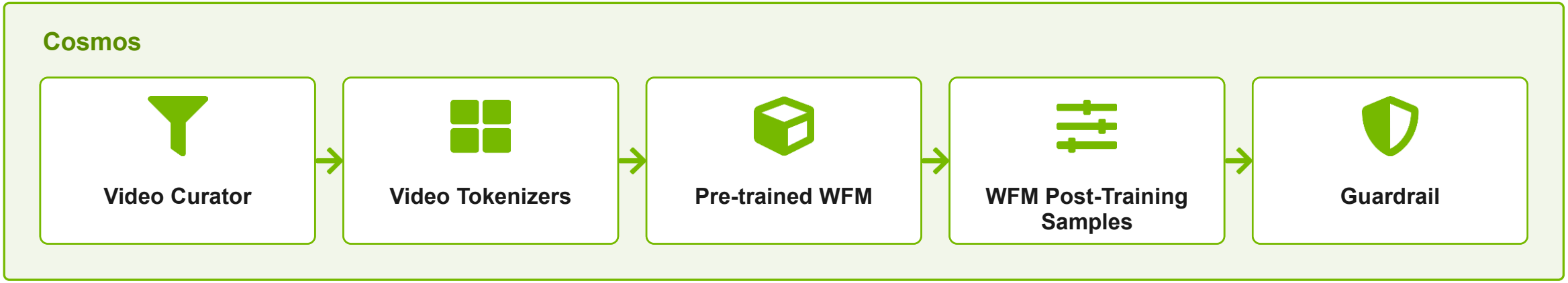
② Future-work 선언 (전부 미검증)

"This paper does not include empirical results in applying Cosmos WFMs to them.
We are eager to verify the claims in future work."

함의: policy eval·init·training·planning·synthetic data — 5개 용도 전부 실증 결과 없음. 2장은 capability 주장 장이 아니라 roadmap 선언 장이다. downstream transfer 증거(적절한 baseline 포함)는 이 논문에 존재하지 않는다.

→ **따라서 2장의 검증 대상은 'WFM이 유용한가'가 아니라, Current Cosmos(데이터·tokenizer·학습 레시피)가 견고한가로 한정해 읽어야 한다.**

Current Cosmos : 실제로 제공되는 플랫폼 구성요소



논문 Figure 4 — Cosmos WFM 플랫폼의 주요 구성요소

- **Video Curator** : 샷 분할 → 필터링 → VLM 주석 → semantic de-dup 으로 다양하지만 compac한 데이터셋 구성
- **Video Tokenizers** : 다양한 압축비의 causal video tokenizer 군 — 인코더가 raw data를 compact token으로 변환
- **Pre-trained WFM** : Diffusion / Autoregressive 두 갈래, 모두 Transformer 기반의 Video2World 생성 모델
- **Post-Training Samples** : camera-pose·humanoid·autonomous driving 등 Downstream 과제로 fine-tuning 데모
- **Guardrail** : 유해 입력·출력을 차단하는 pre-Guard / post-Guard 안전 시스템

① 데이터 Curation & ② Causal Tokenizers

① Video Curator

scalable 비디오 데이터 큐레이션 파이프라인

고품질·고동역학·정보량이 높은 클립만 선별

1. 장면 전환(scene change) 없는 개별 샷(shot)으로 분할
2. 필터링: **고품질·동적** 정보가 풍부한 부분집합 추출
3. VLM으로 고품질 샷 자동 주석(annotation)
4. Semantic de-duplication → 다양하지만 compact한 데이터셋

② Video Tokenizer (Causal)

인과적(causal) Tokenizer 설계 → 현재 프레임의 토큰이 미래 관측에 의존하지 않음 (현재 frame의 token을 계산할 때 미래 frame을 사용하지 않기 때문에, 실제 Physical AI가 살아가는 causal world와 부합)

Causal 설계의 두 가지 이점

학습 측면 — 단일 이미지 입력 시 image tokenizer가 되어 image+video joint training 가능. 풍부하고 다양한 이미지 데이터셋 활용.

응용 측면 — 인과적 세계에 사는 (미래 정보를 사용할 수 없는) Physical AI 시스템과 더 잘 정렬됨.

단, 목적은 '의미 있는 동역학'이 아니라 '픽셀 복원' — reconstruction commitment.

- WFM의 성능 상한은 데이터 품질이 결정하므로, Cosmos는 모델보다 먼저 데이터 큐레이션 파이프라인을 플랫폼의 핵심 구성요소로 둠
- Cosmos의 tokenizer는 단순 압축기가 아니라, Physical AI용 시계열 세계표현 인코더로 볼 수 있음

③ WFM Pre-training — 두 가지 확장 가능한 접근

공통: Transformer 아키텍처(확장성). 둘 다 '현재 관측(과거 비디오) + control input(prompt)'으로 미래를 생성하는 Video2World 모델.

Diffusion-based WFM 사전학습

STEP 1 · Text2World

입력 텍스트 Prompt부터 Video World 생성하는 Text2World 학습

STEP 2 · Video2World

과거 Video + 텍스트 Prompt → 미래 Video World 생성하는 Video2World 모델로 fine-tune

Autoregressive-based WFM 사전학습

STEP 1 · Next-token (foresight) generation

과거 Video만으로 미래 Video 예측하는 Foresight Generation 학습

STEP 2 · Text-conditioned Video2World

과거 Video + 텍스트 Prompt 조건 추가하여 Video2World Generation으로 fine-tune

보조 모듈:

Diffusion WFM을 fine-tune한 **diffusion decoder** 로 AR 모델의 생성 품질 보강

LLM 기반 **prompt upsampler** 로 WFM 제어성 향상. 각 갈래마다 capacity가 다른 모델 군(family) 구축

④ Post-training 응용 & ⑤ Guardrail

④ World Model Post-Training

사전학습된 WFM을 여러 Downstream Physical AI 과제로 fine-tuning함

Camera-pose 조건화

카메라 포즈를 입력 프롬프트로 → 생성된 세계 안을 자유롭게 navigate

Humanoid

휴머노이드 과제용 fine-tuning 시연

Autonomous Driving

자율주행 과제용 fine-tuning

⑤ Guardrail

안전한 사용을 위해 유해한 입력·출력을 차단하는 가드레일 시스템. **pre-Guard**(유해 입력 차단) + **post-Guard**(유해 출력 차단) 구조.

2장에서 미리 봉합되지 않은 약점들

인터페이스 불일치

ct가 행동인지, 의미 조건인지, 확률적 sampling 변수인지 모호

$c_t \in$ {action, random, text} 로 묶어 텍스트 프롬프트를 'control input'이라 부른다. 그러나 ***텍스트는 MDP의 제어 신호가 아니다*** → 2.1의 RL/MPC 용도는 이 인터페이스로 성립하지 않음.

측정 불가능한 상한

좋은 WFM이라는 것을 무엇으로 측정할 것인가?

'WFM 정확도가 성능을 upper-bound한다'고 선언하지만, 그 **동역학 정확도**를 무엇으로 측정할지 부재. 뒤 장의 픽셀 메트릭(PSNR/SSIM/FVD)은 동역학 정확성을 보장하지 못함.

Reconstruction commitment

세계를 이해하는 것인가, 픽셀을 잘 복원하는 것인가?

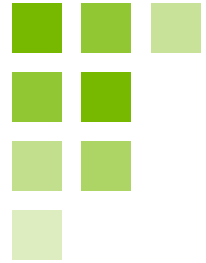
모든 경로가 tokenizer→픽셀 복원을 통과. JEPA류 latent-objective 대안은 설계 공간에서 배제 — '**의미 있는 동역학**'이 아니라 '복원'을 최적화.

Long-horizon / Drift 무방비

짧은 미래는 맞는데, 계속 돌리면 왜 무너지는가?

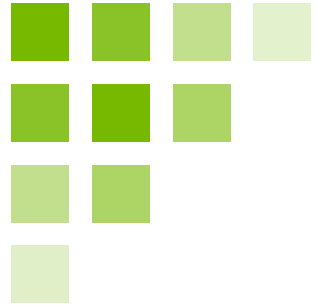
Video2World를 receding-horizon으로 반복 호출할 때의 compounding error 미언급. 특히 AR 경로는 teacher-forcing↔inference 괴리에 직격.

공정한 평가: future-work deferral을 한 단락으로 명시한 점은 신뢰할 만하다. 2장은 capability 장이 아니라 roadmap 장 — 검증은 데이터 파이프라인·causal tokenizer·open-weight 레시피에 한정해 3~6장에서.



ONE-SENTENCE SUMMARY

Cosmos 2장은 WFM을 **‘Physical AI가 안전하게 상호작용할 수 있는 물리 세계의 디지털 예측모델’**로 정의하고, 이를 구현하기 위한 Data Curation, Video Tokenizer, Diffusion/Autoregressive 기반 Pre-training WFM, Post-training Samples, guardrail로 구성된 플랫폼 구조를 제시한다.



NVIDIA COSMOS · WORLD FOUNDATION MODEL PLATFORM

Chapter 3

Data Curation

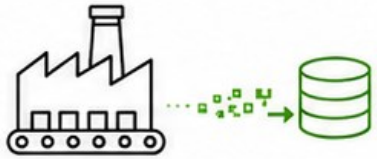
Cosmos Video Curator: Splitting · Filtering · Annotation · Deduplication · Sharding

Cosmos Data Curation

3. Data Curation 핵심정리

Physical AI를 위한 Cosmos WFM 데이터 큐레이션 파이프라인

1. 핵심 메시지



- WFM 성능의 상한은 데이터 품질이 결정
- 대규모 raw video를 물리 세계 학습용 clip으로 정제
- Cosmos Video Curator는 Physical AI Data Factory 역할 수행

핵심: Cosmos는 고품질·고동역학·고다양성 비디오 데이터를 선별해 WFM 학습용 데이터셋으로 변환한다.

2. End-to-End Data Curation Pipeline

Splitting	Filtering	Annotation	Deduplication	Sharding
Shot detection	품질·동역학 선별	VLM caption	Semantic dedup	WebDataset packaging
물리적으로 일관된 clip 분할	저품질·부적절 clip 제거	비디오 설명 생성	중복 제거·다양성 확보	학습 가능한 shard 구성

Splitting → Filtering → Annotation → Deduplication → Sharding

3. 주요 처리 단계

A Splitting & Transcoding

- 2초 미만 clip 제거
- 60초 초과 clip 재분할
- GPU 기반 transcoding
- 약 6.5× throughput 향상

B Filtering

- Motion
- Visual Quality
- Text Overlay
- Video Type

정적 영상, blur/noise, 과도한 텍스트, 비현실적 유형 제거

C Annotation

- VLM 기반 고품질 caption 생성
- 8개 sampled frame 사용
- 평균 약 97 words
- TensorRT-LLM으로 약 10× throughput

D Deduplication

- InternVideo2 embedding 활용
- GPU k-means clustering
- 중복 데이터 약 30% 제거
- 해상도 높은 clip 우선 보존

4. Dataset & Infrastructure

Dataset Composition

Nature dynamics	20%
Hand motion & manipulation	16%
Spatial awareness & navigation	16%
Driving	11%
Human motion & activity	10%
First-person POV	8%
Dynamic camera movements	8%
Synthetic rendered	4%
Others	7%

Raw video 약 2,000만 시간 → pre-training용 약 10^8 clip, fine-tuning용 약 10^7 clip

Infrastructure

분산 클러스터 → Streaming Pipeline (AnyScale Ray) → GPU 가속 처리

- AnyScale Ray 기반 streaming pipeline
- 분산 클러스터 자원 효율화
- Remote storage 환경에서 안정적 처리
- Network · NVDEC · GPU compute 동시 활용

5. 3장의 핵심 결론

- Cosmos Data Curation은 WFM 학습을 위한 **데이터 공장**이다
- 핵심 단계는 **Splitting**, **Filtering**, **Annotation**, **Deduplication**, **Sharding**이다
- 목표는 물리적으로 **일관되고**, **품질이 높으며** 의미적으로 **다양한** 비디오 확보이다
- 결과적으로 Curator는 **tokenizer**와 **WFM 학습의 기반**이 된다

한 줄 요약: Cosmos의 Data Curation은 대규모 raw video를 고품질 Physical AI 학습 데이터로 바꾸는 NVIDIA식 데이터 팩토리이다.

Video Data Curation Pipeline : WFM의 성능 상한은 '데이터'가 결정한다

Physical AI Data Factory

WFM은 비디오를 '많이' 학습한다고 좋아지지 않는다 — **물리 세계의 동역학·시각 품질·장면 일관성·의미적 다양성**을 갖춘 데이터를 선별해야 함. Cosmos는 이를 위해 **대규모 GPU 기반 큐레이션 파이프라인을 구축**함

5단계 Cosmos Video Curator 파이프라인 (Fig. 5)



→ tokenizer와 WFM 모두의 학습용 고품질 데이터셋을 만드는 기반

Video Data Curation Pipeline 전체 구조

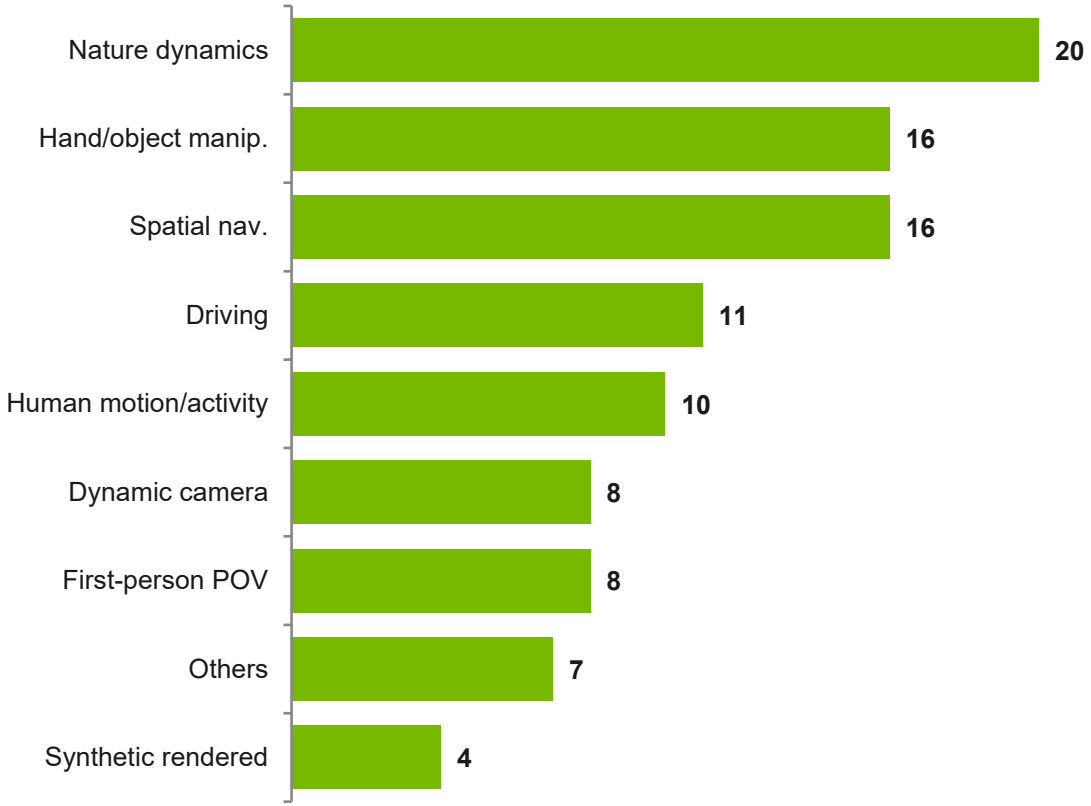
tokenizers와 WFMs의 학습용 고품질 데이터셋을 만들기 위한 기반

단계	목적	핵심 기능
1. Splitting	긴 비디오를 물리적으로 일관된 clip으로 분할	Shot Detection · GPU 기반 Transcoding
2. Filtering	WFM 학습 가치가 낮은 clip 제거	Motion · Quality · Overlay Text · Video Type Filtering
3. Annotation	clip별 고품질 설명(caption) 생성	VLM 기반 video caption
4. Deduplication	중복·유사 데이터 제거로 다양성 확보	Semantic deduplication (k-means)
5. Sharding	학습 가능한 webdataset 형태로 패키징	해상도·종횡비·길이 기준 분할

각 단계는 데이터 품질을 높이고 모델 학습 요구사항(해상도 curriculum, aspect ratio, clip length)에 맞도록 설계됨.

Dataset : 원천 데이터와 도메인 구성

Physical AI 응용을 포괄하는 비디오 카테고리 (%)



규모 & 처리 결과

$\sim 2 \times 10^7$ 시간

시간 규모의 raw video
해상도 720p ~ 4K

$\sim 10^8$ 개

pre-training용
video clips

데이터 출처

proprietary + open-domain
Internet video

$\sim 10^7$ 개

fine-tuning용
video clips

일반 웹 비디오 컬렉션이 아니다. Physical AI가 세계의 물리적 변화, 객체-행동 상호작용, 이동, 조작, 시점 변화, 자연 동역학을 학습할 수 있도록 구성된 데이터셋임. 의미적 중복·저가치 비디오는 이후 단계에서 제거

Splitting : 물리적으로 일관된 clip 생성

Splitting은 긴 비디오를 shot change가 없는 물리적으로 일관된 clip으로 분할함
편집된 장면 전환(shot change)을 학습하지 않도록, 장면 전환 기준으로 비디오를 분할한다.

Shot Detection

- raw video → shot 단위로 분할
- < 2초 clip 제거 (transition·visual effect 가능성)
- > 60초 clip은 최대 60초로 재분할
- 알고리즘 비교: PySceneDetect · Panda70M · TransNetV2 · AutoShot
- ✓ **end-to-end 학습 방법** (TransNetV2·AutoShot)이 hand-crafted feature나 heuristic 기반 방법보다 전반적으로 우수함. 특히 **TransNetV2**가 복잡한 shot change에 강하고 GPU 가속에 유리함

Transcoding

- 다양한 codec·해상도·종횡비 → 일관된 고품질 mp4로 재인코딩
- 이후 처리 안정성 + dataloader 효율 향상
- NVIDIA GPU의 NVENC/NVDEC 하드웨어 인코딩·디코딩 활용
- PyNvideoCodec + ffmpeg 결합이 최대 throughput

약 6.5× throughput 증가 (여러 개선 결합 시)

Table 1: Comparison of splitting algorithms on different datasets.

Dataset	Metric	PySceneDetect	Panda70M	TransNetV2 ✓	AutoShot ✓
BBC	Precision ↑	0.894	0.959	0.983	**0.984**
	Recall ↑	0.884	0.653	**0.951**	0.922
	F1 ↑	0.889	0.777	**0.967**	0.952
RAI	Precision ↑	0.856	**0.933**	0.918	0.889
	Recall ↑	0.807	0.746	0.921	**0.923**
	F1 ↑	0.831	0.829	**0.919**	0.906
SHOT	Precision ↑	0.769	**0.949**	0.883	0.866
	Recall ↑	0.673	0.462	0.767	**0.804**
	F1 ↑	0.718	0.622	0.821	**0.834**
ClipShots	Precision ↑	0.395	0.649	**0.685**	0.653
	Recall ↑	0.602	0.424	0.772	**0.781**
	F1 ↑	0.477	0.513	**0.726**	0.711

Table 2: Transcoding performance with different software settings.

Method	GPU	CPU	Codec	Batch	NVDEC (#accelerator)	NVENC (#accelerator)	Throughput (videos/s)
ffmpeg	H100	28	libx264	1	7	0	0.0574
ffmpeg	L40S	1	h264_nvenc	1	3	3	0.0674
ffmpeg	L40S	1	h264_nvenc	16	3	3	0.1026
✓ pynvc+ffmpeg	L40S	1	h264_nvenc	1	3	3	✓ **0.3702**

Filtering : WFM 학습에 적합한 데이터 선별

Filtering은 splitting 결과로 생성된 clip 중에서 품질이 낮거나, WFM 학습에 부적절하거나, 데이터 분포를 왜곡할 수 있는 clip을 제거하는 단계임
목적: ① 최소 품질 미달 clip 제거 ② fine-tuning용 고품질 clip 선별 ③ WFM 구축에 맞게 데이터 분포 조정

1. Motion Filtering

- 1) **정적 비디오 및 급격한 카메라 흔들림 제거**, 2) **pan/zoom/tilt 카메라 움직임 유형을 태깅**하여 모델 학습에 활용
- 이를 위해 motion vector/optical flow 입력으로 하는 **ViT lightweight classifier** 사용 — TensorRT 가속 optical flow 추정 기반 Classifier가 최고 성능

2. Visual Quality Filtering

- Artifact·noise·blur·노출 문제 제거하기 위해 **DOVER 기반 VQA**(video quality assessment model)로 **perceptual score 하위 15% 제거**
- Sampled frame에 **Image Aesthetic Model(IAA, Image Aesthetic Assessment)**로 **appearance quality가 낮은 clip 제거**

3. Text Overlay Filtering

- 후처리로 삽입된 과도한 텍스트(흔히 VFX 동반)는 물리 세계 학습에 방해
- **InternVideo2**로 **Semantic embedding** 생성 후 이를 작은 **MLP binary classifier**로 이러한 Clip를 탐지

❖ **InternVideo2**: 대규모 video representation을 학습한 video foundation model

4. Video Type Filtering (Filtering + Dataset Rebalancing)

- 각 Clip을 Content type/Visual style에 기반한 taxonomy로 분류하는 단계임
- 추상적 시각 패턴, game footage, animated content 등은 poor generation quality 또는 unrealistic dynamics를 유발할 수 있어 제외
- **human action, human-object interaction**처럼 WFM에 중요한 category는 **upsampling**하고, **nature 또는 landscape**처럼 상대적으로 덜 중요한 category는 **downsampling**하여 데이터 분포를 조정

Annotation : VLM 기반 비디오 설명(caption) 생성

VLM Video Captioning

- 각 video clip에 대해 고품질 text description을 생성
- Alternative text에 의존하지 않고 VLM으로 caption 생성 — **비디오의 material facts·details에 집중**
- 다양한 text style/format에 모델의 적응 부담 감소
- 비교: VFC · Qwen2-VL · VILA → 소규모 human evaluation에서 **VILA**가 더 정확
- 내부 VILA 13B video captioning 모델 사용, 8개 uniformly sampled frame 입력으로 넣어 caption 생성

VFC	**Video Foundation model 기반 Captioning 모델**	영상의 시공간 정보를 분석하여 기본적인 video caption 생성. 비교 기준선 역할
Qwen2-VL	Alibaba의 범용 **Vision-Language Model**	이미지·영상과 텍스트를 함께 이해하며 객체, 장면, 행동 등에 대한 설명 및 질의응답에 강점
VILA	NVIDIA/MIT 계열의 **Vision-Language Model**	이미지뿐만 아니라 **video의 시간적 변화와 행동을 상세하게 이해**하는 데 강하며, Cosmos 데이터 captioning에서 높은 정확도를 보여 최종적으로 활용

≈ 559자

평균 caption 길이
(약 97 words)

≈ 10× throughput

FP8 TensorRT-LLM
vs PyTorch FP16

핵심

비디오를 단순 frame sequence로 두지 않고 **언어적 조건·의미 정보**를 부여하여, **Text2World** 및 **Video2World** 학습에 쓰는 **multimodal training sample**로 변환한다.

Table 3: Inference throughput comparison of VILA on a single H100 GPU.

Engine	Precision	Batch Size	Throughput (clips/s)	Throughput (tokens/s)
PyTorch	FP16	1	0.21	49.6
TRT-LLM	FP16	1	0.40	95.6
TRT-LLM	FP16	16	1.09	260.9
TRT-LLM	FP8	16	1.96	470.6

Deduplication : 의미적 중복 제거와 다양성 확보

*대규모 웹 비디오의 중복은 학습 효율을 낮추고 특정 샘플 memorization 위험을 높인다 → **semantic deduplication**으로 균형 있는 분포 구성.*

Semantic Deduplication 절차

Video Clips → InternVideo2 Embedding → K-means Clustering → Pairwise Similarity → Duplicate Removal → Representative Clip

1 SemDeDup · DataComp 접근

InternVideo2로 모든 clip을 Embedding (filtering 단계 InternVideo2 embedding 재사용)

2 Multi-node GPU k-means

Embedding을 클러스터링 (여러 GPU/노드에 embedding과 clustering 계산을 분산시켜 대규모 데이터셋을 처리)

3 Cluster 내 pairwise distance

Embedding 간 거리가 매우 가까운 영상들을 semantic duplicate 후보로 판단 (중복 데이터 탐지. 중복이 발견되면 가장 높은 해상도의 clip을 남겨 품질 저하를 방지)

4 최고 해상도 clip 유지

의미적으로 동일한 영상들을 묶은 후 그 중 가장 좋은 대표 영상을 보존함 (품질 저하 방지)

약 30% 의 training data 제거

중복 데이터 정리로 더 다양하고 균형 잡힌 학습 분포 확보.

Visual Search Engine

InternVideo2 embedding과 clustering 결과 활용하여 **free-form text/video query** 지원하는 Visual Search Engine 구축 → 데이터 디버깅과 pre-training dataset과 downstream application 간 gap 분석

Sharding : 학습 가능한 데이터셋 패키징

```
‹/› 일반 텍스트

shard-000001.tar
 ├ 000001.mp4
 ├ 000001.txt/json
 ├ 000002.mp4
 ├ 000002.txt/json
 └ ...

shard-000002.tar
 ├ ...
```

caption, metadata, filtering 결과 등 학습에 필요한 정보

Webdataset Packaging

$$Shard_1 \rightarrow GPU_1, \quad Shard_2 \rightarrow GPU_2, \quad \cdots$$

Sharding은 **학습 조건에 따라 데이터를 그룹화**하고, trainer가 효율적으로 streaming할 수 있는 **WebDataset shard로 패키징**하는 과정
Cosmos는 **해상도·aspect ratio·길이**를 기준으로 shard로 나누어 training curriculum과 정렬함
Pre-training dataset뿐 아니라 엄격한 filter를 적용한 high-quality fine-tuning dataset도 생성함

모델 학습 스케줄

Curriculum learning :
Easy/Coarse → Complex/Fine

초기에 낮은 해상도나 비교적 짧은 clip 사용하여 기본적인 visual/dynamics representation 학습하고, 이후 더 높은 해상도와 고품질 데이터 사용

해상도 curriculum

해상도 Curriculum 예시 :
256p →480p → 720p

낮은 해상도로 먼저 scene structure, object, motion, coarse dynamics등을 학습하고, 이후 높은 해상도에서 fine detail, texture, geometry를 정교하게 학습

aspect ratio bucket

Aspect ratio 예시 :
16:9, 9:16, 4:3, 1:1

Aspect Ratio Bucketing → Less Crop/Pad + Efficient Batch + Better Scene Preservation

clip length 조건

clip의 길이 또는 frame 수에 따라 grouping
예시 :
T = 16, 32, 64, 128 frames

WFM의 temporal horizon과 관련
짧은 clip에서 short-term motion을 주로 학습하고, 긴 clip에서 long-term dynamics / temporal consistency을 학습 가능

단순 저장이 아니다. **모델 학습 스케줄**, **해상도 curriculum**, **aspect ratio bucket**, **clip length** 조건에 맞춰 데이터를 학습 가능한 형태로 구조화하는 단계임

Infrastructure : Ray 기반 대규모 스트리밍 파이프라인

Ray : Python 기반의 오픈소스 분산 컴퓨팅 프레임워크
Anyscale : Ray 기반의 클라우드/클러스터 플랫폼

Anyscale Ray Streaming Pipeline :
Ray 분산 실행 엔진 + 대규모 Streaming Data Pipeline

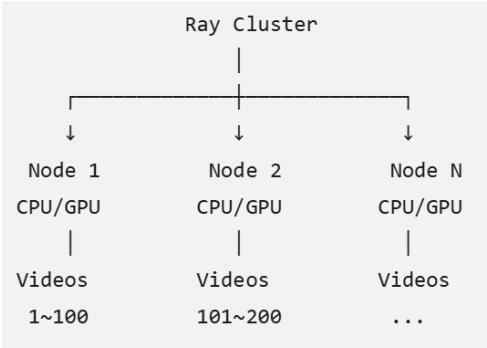
AnyScale Ray Streaming Pipeline

Cosmos의 데이터 큐레이션은 단일 서버 전처리가 아니라, 지리적으로 분산된 클러스터에서 동작하는 **대규모 ML data processing system**임
→ **AnyScale Ray**를 사용하여 streaming pipeline을 구현

두 핵심 문제 해결	**homogeneous node 전반의 효율적 resource utilization** **고지연 remote data source 환경의 robust operation**
전송·계산 분리	remote storage를 쓰면서도 효율적 동작. memory 요구량이 dataset size가 아니라 pipeline complexity에 비례하도록 설계
자원 동시 활용	network bandwidth, NVDEC video decoding unit, GPU compute를 pipeline stage 별로 동시 활용

Cosmos Data Curation = 알고리즘 파이프라인이자 GPU 가속 데이터 공장 (Physical AI Data Factory)

여러 노드에 작업을 자동 분산



Pipeline 작업	주로 사용하는 자원
Remote video read	Network bandwidth
Video decoding	**NVDEC**
Optical Flow	GPU
InternVideo2 embedding	GPU
VLM Captioning	GPU
K-means/Dedup	CPU/GPU
WebDataset write	CPU + Storage I/O

Data Curation Algorithms

$$Split \rightarrow Filter \rightarrow Caption \rightarrow Dedup \rightarrow Shard$$

$$\Downarrow$$

Ray Streaming Infrastructure

$$Distributed\ CPU/GPU + NVDEC + Network + Storage$$

WFM 학습 데이터 조건 ↔ Cosmos의 대응

WFM 학습에 필요한 데이터 조건	Cosmos의 대응 방식
물리적으로 일관된 장면 전개	Shot detection & clip splitting
충분한 동역학 정보	Motion filtering
높은 시각 품질	Visual quality filtering
후처리 artifact 최소화	Text overlay filtering
현실적인 물리 동작	Video type filtering
텍스트 조건 학습 가능성	VLM annotation
데이터 다양성	Semantic deduplication
대규모 학습 효율	Sharding & Ray 기반 infrastructure

2장이 플랫폼의 개념도라면, 3장은 그 첫 번째 핵심 엔진 — 고품질 Physical AI 학습 데이터 생산 체계

Cosmos v1에서 "물리 동역학 학습"은 raw video에 힘·속도·가속도 등의 물리량을 명시적으로 라벨링하는 방식이 아니라, 실제 세계의 움직임과 상호작용이 풍부한 영상을 대규모로 선별하고 그 연속적인 시공간 state transition을 생성모델이 학습하게 하는 방식

$$\textbf{Physical Dynamics} \begin{cases} \text{① Ego/Camera Motion} & \text{이동, 회전, pan/tilt} \\ \text{② Object Motion} & \text{이동, 회전, 속도 변화} \\ \text{③ Human/Agent Action} & \text{걷기, 잡기, 밀기, 운전} \\ \text{④ Object Interaction} & \text{접촉, 충돌, 밀기, 잡기} \\ \text{⑤ Environment Dynamics} & \text{물, 연기, 날씨, 조명 변화} \\ \text{⑥ Causal/Action Dynamics} & a_t \rightarrow s_{t+1} \end{cases}$$

단계	Dynamics 관점의 역할
Splitting	긴 영상을 dynamics가 비교적 일관된 짧은 clip으로 분할
Motion Filtering	정적 영상 및 과도한 camera motion 제거 → 의미 있는 temporal change 확보
Visual Quality Filtering	blur/noise/artifact 제거 → 관찰 가능한 dynamics 확보
Text Overlay Filtering	실제 물리세계와 무관한 overlay 제거
Video Type Filtering	animation/game 등 비현실적 dynamics를 줄이고 human action/interaction 등을 강화
VLM Captioning	clip의 객체·행동·상호작용을 자연어 semantic description으로 부여
Semantic Deduplication	반복되는 동일 dynamics를 줄여 다양성 확보
Sharding	해상도·길이·aspect ratio 등에 맞춰 실제 학습 데이터로 구성

Cosmos : 대규모 실세계 비디오로부터 물리 세계의 동역학을 학습하는 World Foundation Model

2,000만 시간 규모의 Raw Video에서 고품질·고동역학·의미적으로 다양한 데이터를 선별하고, VLM caption • semantic deduplication • sharding • Ray 기반 인프라를 통해 물리 세계의 동역학을 학습하는 Physical AI Data Factory



1 물리 세계의 동역학 (Dynamics)이란?

현재 세계 상태가 시간에 따라 다음 상태로 변하는 과정



Dynamics = Temporal Change + Object Motion + Interaction + Physical Consequence

움직임 (Motion) → 상호작용 (Interaction) → 물리적 결과 (Physical Consequence)

2 어떤 Dynamics가 중요한가? (분류 예시)

WFM 학습에 유용한 다양한 물리적 변화와 상호작용 장면을 포함

1 카메라/자기운동



이동, 회전, pan/tilt, zoom

2 객체 운동



이동, 회전, 가속/감속

3 인간/에이전트 행동



걷기, 달리기, 조작, 운전 등

4 객체 상호작용



잡기, 밀기, 충돌, 접촉, 쌓기

5 환경 동역학



물, 연기, 날씨, 조명 변화

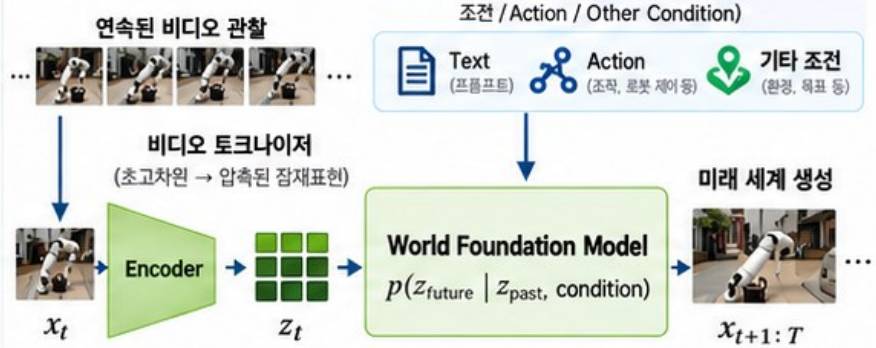
6 인과적 동역학



행동 → 미래 상태 변화

3 학습 방식 : 명시적 물리식이 아닌, 압축적 동역학 학습

대규모 비디오에서 관찰되는 물리적 규칙성을 잠재공간에서 학습하여, 이후 세계가 어떻게 변할지 예측/생성



4 Data Curation : 2,000만 시간의 Raw Video를 WFM 학습용 데이터로 변환 (Physical AI Data Factory)

물리적 동역학이 풍부하고 의미적으로 다양한 고품질 데이터셋 구축

4.1 Raw Video Collection (2,000만 시간)

- 인터넷
- 공공 데이터
- 로보틱스 데이터

다양한 소스에서 수집



4.2 Splitting (클립 분할)

- 긴 영상을 일정한 dynamics를 갖는
- 짧은 clip으로 분할



4.3 Filtering (품질 및 동역학 필터링)

- Motion Filtering
- Visual Quality Filtering
- Text Overlay Filtering
- Video Type Filtering



4.4 VLM Captioning (비디오 설명 생성)

- VILA (가장 정확)
- Qwen2-VL, VFC 등



A robotic arm picks up a cup and places it on the table.

4.5 Semantic Deduplication (의미적 중복 제거)

- InternVideo2 embedding
- Clustering + 유사도 비교
- 최고 해상도 clip만 유지



중복된 / 유사한 의미의 영상



대표 clip 유지 (최고 해상도)

4.6 Sharding (학습용 패키징)

- 해상도·가로세로비·길이별 WebDataset으로 패키징
- 학습 스케줄과 정렬

shard-00001.tar
- 000001.mp4 + caption.json
- 000002.mp4 + caption.json
- 000003.mp4 + caption.json
- ...

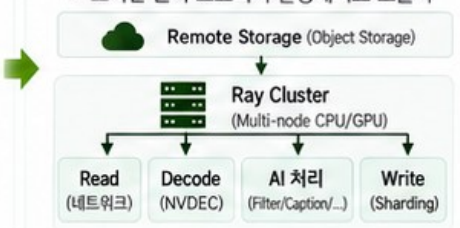
shard-00002.tar

shard-00003.tar

...

4.7 Ray Streaming Pipeline (대규모 분산 처리 인프라)

- 전송과 계산의 분리 (Streaming)
- CPU/GPU, NVDEC, Network 동시 활용
- 고지연 원격 스토리지 환경에서도 효율적



5 학습 데이터셋의 특징

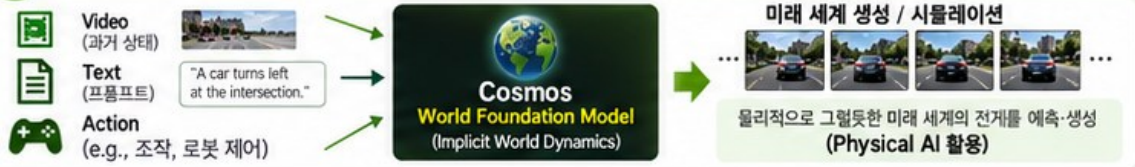
고품질
- 저화질·노이즈 제거
- 실제 세계 중심 데이터

고동역학
- 움직임·상호작용이 있는 clip
- human action, interaction 강화

의미적 다양성
- VLM caption으로 객체·행동·장면 구조화
- semantic deduplication

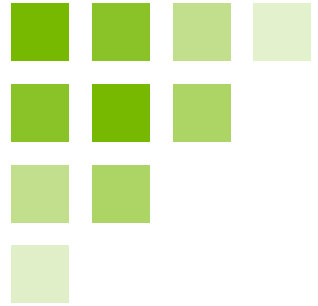
대규모 · 학습 효율적
- 2,000만 시간 규모
- WebDataset sharding
- Ray 기반 분산 파이프라인

6 결과 : World Foundation Model (Cosmos) – Physical AI에서의 활용



ONE-SENTENCE SUMMARY

3장「**Data Curation**」은 Cosmos WFM이 물리 세계의 동역학을 학습할 수 있도록, **2,000만 시간 규모의 raw video**에서 고품질·고동역학·의미적으로 다양한 clip을 선별하고, VLM caption · semantic deduplication · sharding · Ray 기반 GPU 인프라를 통해 WFM 학습용 데이터셋으로 변환하는 **Physical AI Data Factory**를 설명한다.



NVIDIA COSMOS · WORLD FOUNDATION MODEL PLATFORM

Chapter 4

Tokenizer

Cosmos Tokenizer: 고해상도 image/video → compact token (continuous / discrete)

Cosmos Tokenizer

4. Tokenizer 핵심정리

Physical AI를 위한 Cosmos WFM Tokenizer 구조와 성능

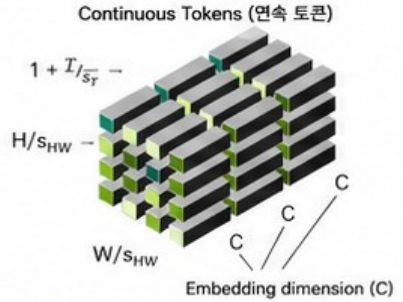
1. 핵심 메시지



- 고해상도 image/video를 WFM 학습 가능한 compact token으로 변환
- 시각 정보와 시간적 동역학을 최대한 보존
- 높은 압축률, 재구성 품질, 빠른 runtime을 동시에 추구

Tokenizer는 WFM의 입력 표현 공간을 결정하는 핵심 모듈이다.

2. Token Types: Continuous vs Discrete

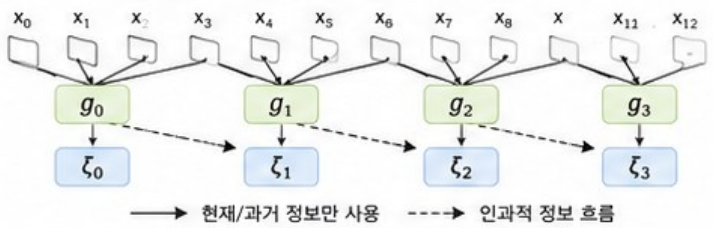


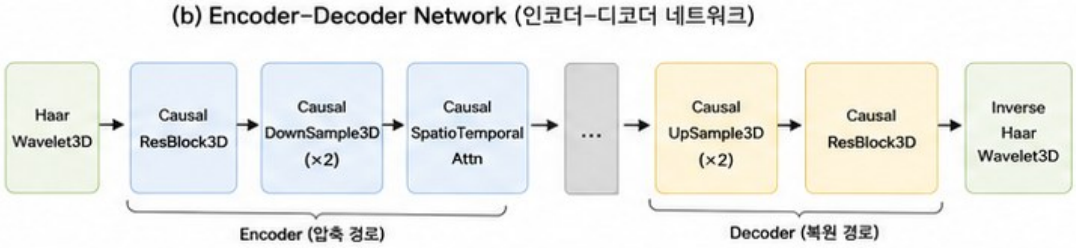


- Continuous Tokenizer → Diffusion WFM
- Discrete Tokenizer → Autoregressive WFM
- Joint Image-Video Shared Latent Space
- Spatial compression s_{HW}, temporal compression s_T

첫 temporal token은 첫 입력 frame을 나타내며 image와 video를 통합적으로 처리

3. Cosmos Tokenizer Architecture





- Temporally causal encoder-decoder
- 2-level wavelet transform 기반 압축
- Spatio-temporal factorized 3D convolution
- Causal self-attention으로 장거리 의존성 처리
- Continuous는 AE, Discrete는 FSQ 사용
- 다양한 aspect ratio와 긴 temporal length 지원

4. Training Strategy



Two-stage training

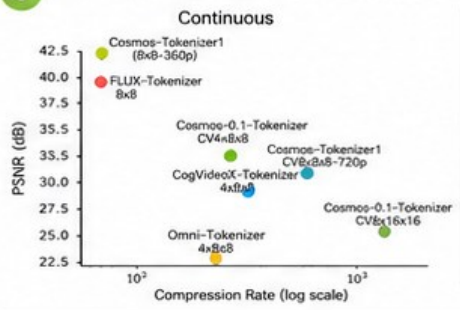
Stage 1		Stage 2		
L1 loss	Perceptual loss	Optical Flow loss	Gram-matrix loss	Adversarial fine-tuning

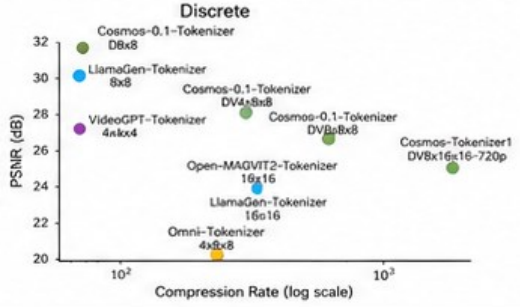
Auxiliary latent loss 없이
final decoder output만 supervision

Compression rates
Image 8x8, 16x16 / Video 4x8x8, 8x8x8, 8x16x16

Cosmos-0.1-Tokenizer /
Cosmos-Tokenize1 family

5. Results & Benchmarks





- DAVIS, TokenBench, MS-COCO, ImageNet-1K 평가
- 압축률 대비 재구성 품질 trade-off 우수
- DAVIS에서 기존 대비 **약 +4 dB PSNR 개선**
- 단일 A100 80GB에서 **최대 12배 빠른 속도**
- 1080p 8초, 720p 10초 **비디오 원샷** encoding 가능

6. 6. 4장의 핵심 결론

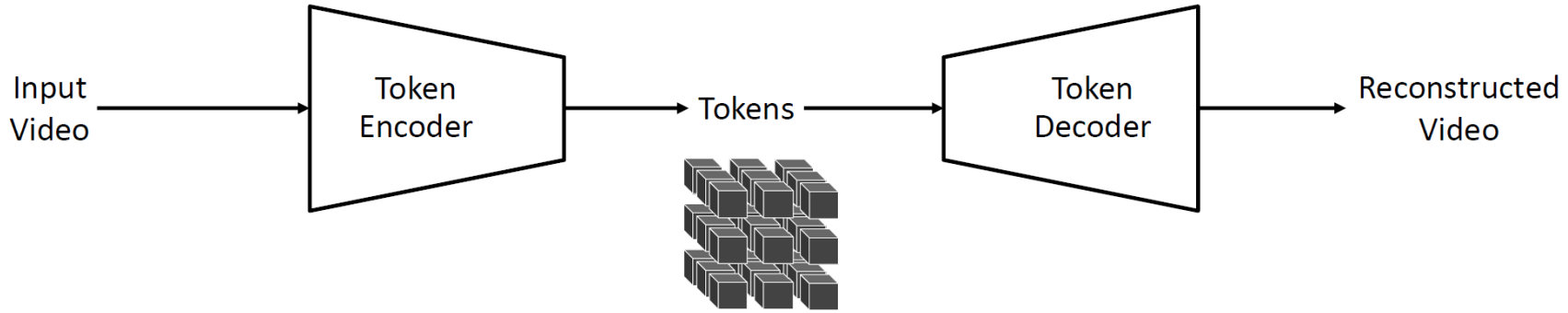
1. "Cosmos Tokenizer"는 Physical AI용 시각-시간 표현 엔진이다.
2. "Causal" 구조로 실제 세계의 시간 흐름과 정합적이다.
3. "Continuous + Discrete" 지원으로 diffusion과 autoregressive WFM을 모두 뒷받침한다.
4. "Fast + High Quality" 특성으로 WFM 학습의 기반 인프라가 된다.

★ 한 줄 요약: Cosmos Tokenizer는 image/video를 compact latent token으로 바꾸어 Physical AI가 세계를 이해하고 예측할 수 있게 하는 WFM의 핵심 표현 계층이다.

비디오를 WFM 학습용 compact token으로 변환

비디오는 고차원·고중복 데이터 — raw pixel로 Transformer WFM에 넣으면 token 수·연산량이 폭증. 시각·시간 정보를 보존하면서 token을 줄이는 고품질 압축 표현(visual tokenizer)이 필수.

Video Tokenization Pipeline



논문 Figure 6 — 입력 비디오를 token으로 encoding하고, decoder가 다시 복원. 학습 목표는 token이 시각 정보를 최대한 보존하도록 encoder·decoder를 학습하는 것.

Cosmos Tokenizer는 고해상도·장시간 비디오를 WFM 학습에 적합한 **compact token**으로 압축하면서, **물리 세계의 시각 정보와 시간적 동역학을 최대한 보존**하는 핵심 모듈이다.

Tokenizer가 필요한 이유 : WFM의 입력 표현 공간을 결정하는 기반 인프라

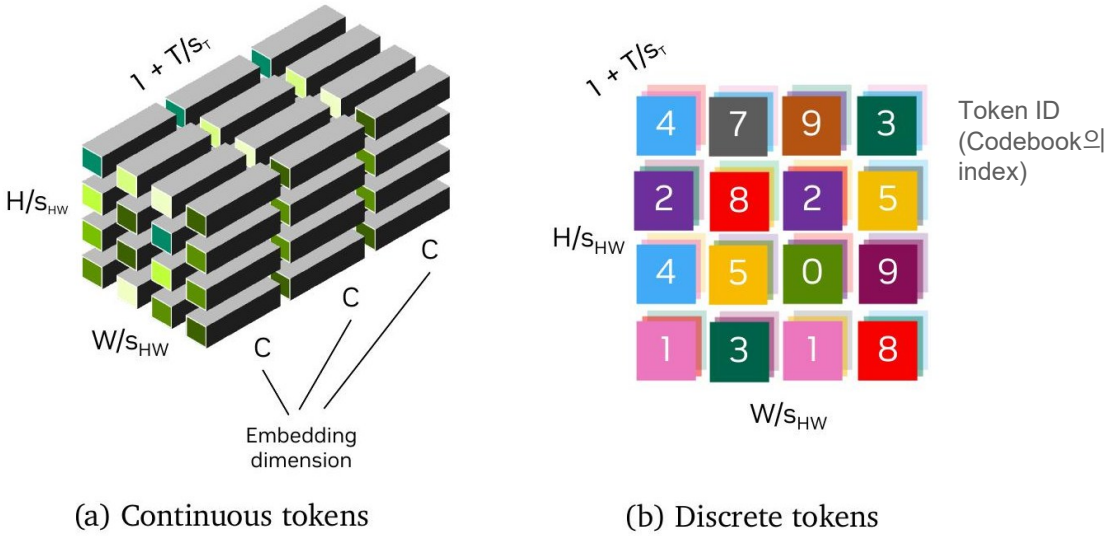
WFM은 비디오 기반으로 물리 세계의 미래를 예측해야 함. 그러나 비디오는 공간 해상도, 시간 길이, 색상 채널, 움직임 정보를 모두 포함하기 때문에 raw pixel 수준에서 직접 학습하면 token 수와 연산량이 폭증함

비디오 tokenization은 video codec과 유사한 문제 — 원본을 최대한 보존하면서 학습 가능한 수준으로 압축. WFM 학습 복잡도는 token 수에 비례하므로 compact token sequence가 필수.

구분	의미
입력	고차원 image / video
처리	encoder가 compact latent token으로 압축
출력	WFM 학습·생성에 사용할 token sequence
복원	decoder가 token에서 image / video 재구성
핵심 목표	압축률(compression rate)과 재구성 품질의 균형

→ Cosmos Tokenizer는 단순 전처리가 아니라, WFM의 입력 표현 공간을 결정하는 기반 인프라.

Continuous Tokenizer vs Discrete Tokenizer



논문 Figure 7 — (a) 연속 latent embedding(차원 C) / (b) 양자화된 discrete index

Tokens along spatial ($\frac{H}{s_{HW}} \times \frac{W}{s_{HW}}$) and temporal ($1 + \frac{T}{s_T}$) dimensions, with a spatial compression factor of s_{HW} and a temporal compression factor of s_T.

The first temporal token represents the first input frame, enabling joint image ($T = 0$) and video ($T > 0$) tokenization in a shared latent space.

(a) Continuous latent embeddings with an embedding size of C. (b) Quantized indices, each color representing a discrete latent code.

구분	표현 / 연결 WFM
Continuous tokenizer	**영상 특징을 실수 벡터 자체로 표현** 연속 latent embedding vector → Diffusion WFM (Stable Diffusion류 latent diffusion model처럼 연속 latent embedding을 사용)
Discrete tokenizer	**영상 특징을 codebook의 정수 ID로 표현** 양자화된 integer code index → Autoregressive WFM (GPT류 autoregressive transformer가 next-token prediction을 수행할 수 있도록 discrete code index로 비디오를 표현)

Cosmos WFM 전체 구조에서의 의미

Diffusion WFM → Continuous Video Tokenizer
Autoregressive WFM → Discrete Video Tokenizer

4장은 5장의 두 WFM 계열을 잇는 공통 표현 계층.

Cosmos Tokenizer의 핵심 특징

Cosmos Tokenizer는 image와 video 모두를 처리하고, continuous와 discrete token 모두를 지원하는 visual tokenizer suite

Causal
현재 frame token 계산 시 미래 frame 미사용

Joint Image-Video
image·video를 통합 architecture로 처리

Continuous + Discrete
diffusion·autoregressive WFM 모두 지원

Multi-aspect
1:1, 3:4, 4:3, 9:16, 16:9 종횡비 지원

Length-agnostic
학습보다 긴 temporal length도 inference 가능

High Efficiency
기존 SOTA 대비 빠른 encoding/decoding

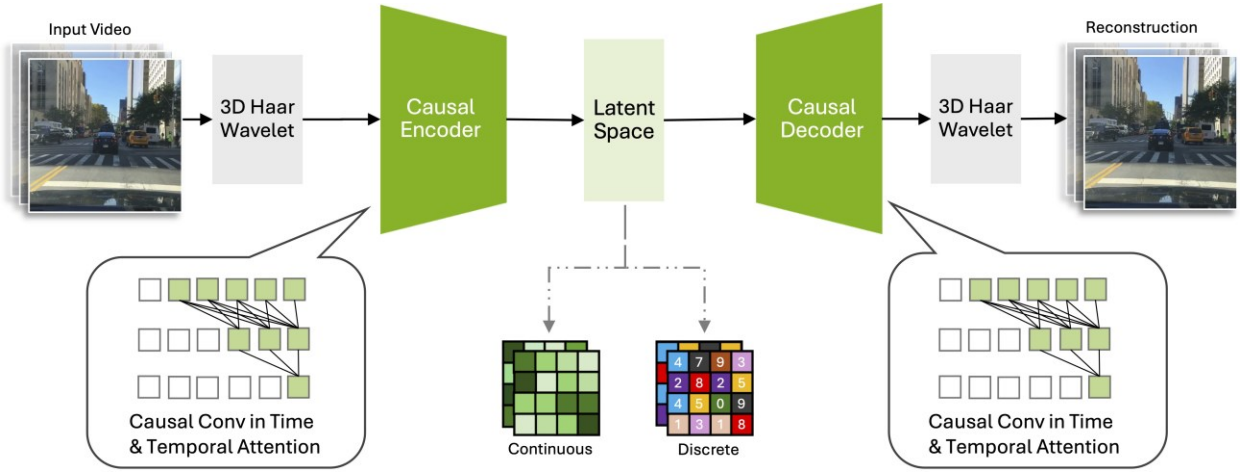
Table 4 — 기능 비교 (Cosmos-Tokenize1은 6개 항목 모두 지원)

Model	Causal	Image	Video	Joint	Discrete	Continuous
FLUX-Tokenizer	–	✓	✗	✗	✗	✓
Open-MAGVIT2-Tokenizer	–	✓	✗	✗	✓	✗
LlamaGen-Tokenizer	–	✓	✗	✗	✓	✗
VideoGPT-Tokenizer	✗	✗	✓	✗	✓	✗
Omni-Tokenizer	✗	✓	✓	✓	✓	✓
CogVideoX-Tokenizer	✓	✓	✓	✓	✗	✓
Cosmos-Tokenize1	✓	✓	✓	✓	✓	✓

Architecture : Temporally Causal Encoder–Decoder

[Cosmos Tokenizer: A suite of image and video neural tokenizers.]

$$\hat{x}_{0:T} = \mathcal{D}(\mathcal{E}(x_{0:T}))$$

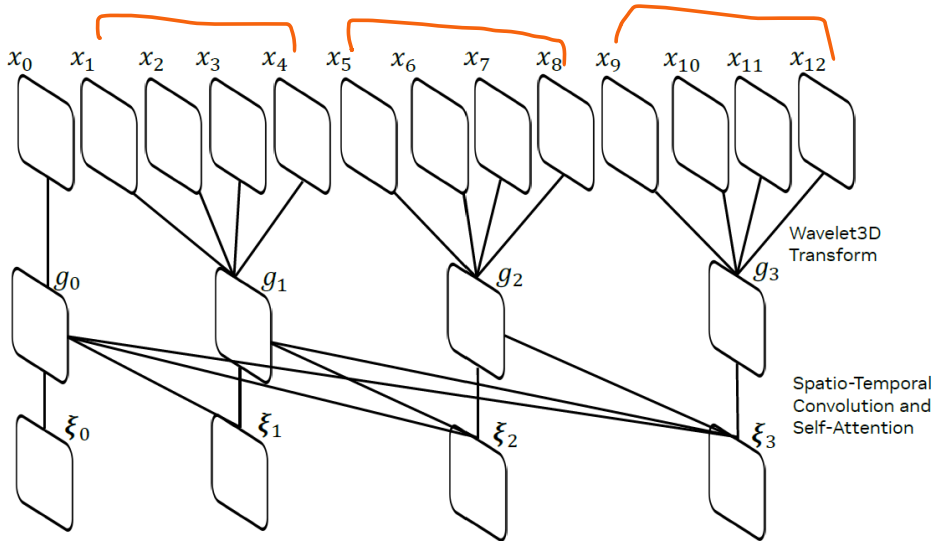




Architecture : Temporally Causal Encoder–Decoder

Temporal Causal Design

Cosmos Tokenizer는 각 단계가 **현재와 과거 frame만 사용**하도록 설계됨.
(미래 frame에 의존하지 않음)



$$\hat{x}_{0:T} = \mathcal{D}(\mathcal{E}(x_{0:T}))$$

입력 비디오 $x_{0:T}$를 encoder $\mathcal{E}$가 token video $z_{0:T'}$로 압축하고, decoder $\mathcal{D}$가 이를 다시 reconstructed video $\hat{x}_{0:T}$ 로 복원함

논문 Figure 9(a) — Temporal causality (과거·현재 frame만 사용); Illustration of the temporal causality mechanism, where inputs $x_0, x_1 \ldots, x_{12}$ are processed through grouped intermediate outputs $g_0, g_1, \ldots,$ and further refined by spatio-temporal convolution and attention operations.

Wavelet Space

입력 비디오를 2-level wavelet transform으로 변환. 이는 x,y,t 방향에서 중복성을 줄이고 더 compact한 표현으로 변환한 뒤, 이후 neural network layer들이 더 semantic한 압축에 집중

$$s_{HW} = H/H' = W/W' \qquad s_T = T/T'$$

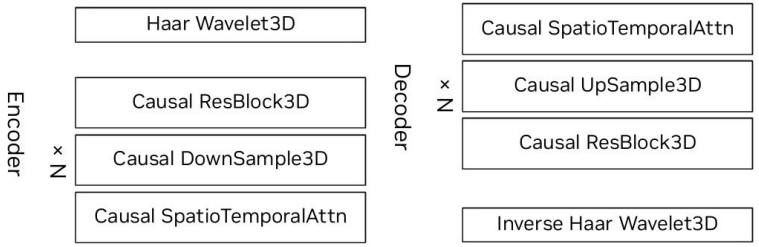




Figure 9(b) — Encoder–Decoder 블록 (Haar Wavelet3D · Causal ResBlock/DownSample/Attn)

Temporally Causal

현재·과거 frame만 사용

Wavelet Space

2-level wavelet로 x,y,t 중복 감소

Spatio-temporal Factorized 3D Conv

3D Conv : 2D conv → temporal conv, left padding : Causality 보장

Causal Spatio-temporal Self -Attn

장거리 temporal 의존성 처리; 장면 내 장기 의존성, 물체 움직임, 동작 연속성 등 표현

NVIDIA Cosmos Tokenizer1: Temporally Causal Encoder-Decoder Architecture

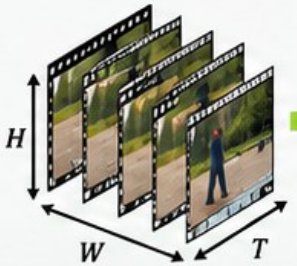
Cosmos Tokenizer1은 비디오를 시공간적으로 압축하여 **World Foundation Model(WFM)**이 처리할 수 있는 latent token으로 변환하는 **Temporally Causal Video Autoencoder**입니다. 3D Wavelet 변환, Causal 3D Convolution, Spatio-Temporal Self-Attention을 사용하여 미래 프레임을 보지 않고 현재와 과거 프레임만으로 compact하고 의미 있는 시공간 latent를 학습합니다.

Key Features

1. **Temporal Causality** : 현재와 과거 프레임만 사용
2. **Wavelet Space** : 3D Haar Wavelet로 시공간 중복성 제거
3. **Spatio-Temporal Modeling** : 3D Conv + Self-Attention
4. **High Compression** : WFM에 최적화된 compact latent token

1 입력 비디오 (Video)

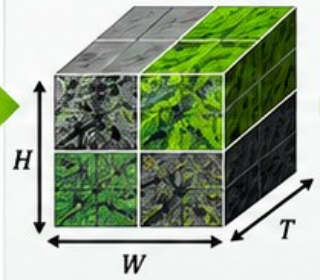
$X \in \mathbb{R}^{T \times H \times W \times 3}$



T : 시간 프레임 수
H, W : 공간 해상도
3 : RGB 채널

2 Haar Wavelet3D

시공간 중복성 분해
(가역 변환, invertible)



x, y, t 방향으로 2-level Haar Wavelet 변환하여 저주파/고주파 성분으로 분해

3 Causal Encoder $\mathcal{E}$: 과거와 현재 프레임만 사용하여 시공간 표현을 압축

× N (반복)

Causal ResBlock3D	Causal DownSample3D	Causal Spatio-Temporal Self-Attention
Causal 3D Conv + Residual	시공간 압축 $(H', W', T'$로 축소)	장거리 시공간 의존성 (인과적 마스크 적용)

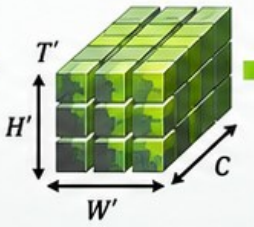


공간·시간 특징 추출 (미래 정보 차단)

시공간 해상도 감소 $(H, W, T) \rightarrow (H', W', T')$

4 Latent Tokens z

WFM이 처리할 compact한 시공간 표현
$z \in \mathbb{R}^{T' \times H' \times W' \times C}$



$s_{HW} = H/H' = W/W'$
$s_T = T/T'$
C : latent channel 수

5 Causal Decoder $\mathcal{D}$: latent token에서 원본 비디오를 복원

× N (반복)

Causal Spatio-Temporal Self-Attention	Causal UpSample3D	Causal ResBlock3D
장거리 시공간 의존성 (인과적 마스크 적용)	시공간 확장 $(H', W', T' \rightarrow H, W, T)$	Causal 3D Conv + Residual



세부 영상 복원

6 복원 비디오

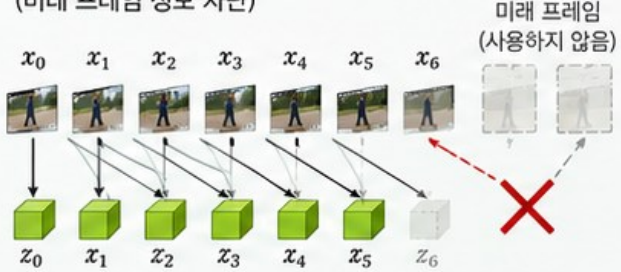
$\hat{x}_{0:T} = \mathcal{D}(\mathcal{E}(x_{0:T}))$



학습 목표
$\hat{x}_{0:T} \approx x_{0:T}$
(재구성 손실 최소화)

1 Temporal Causality (인과성)

각 시점의 표현은 현재와 과거 프레임만 사용
(미래 프레임 정보 차단)

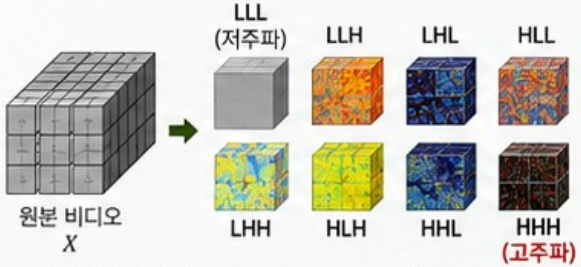


$z_t = f(x_0, x_1, \ldots, x_t)$ $(x_{t+1}, x_{t+2}, \ldots$ 사용하지 않음)

→ World Model이 미래를 예측해야 하므로 미래 정보 누수를 방지

2 Haar Wavelet3D 변환

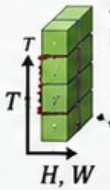
x, y, t 방향으로 2-level Haar Wavelet 변환



- 공간/시간 중복성 제거로 더 compact한 표현
- 주요 구조(저주파)와 변화(고주파) 성분 분리
- 가역 변환이므로 Decoder에서 원본 영상 복원 가능

3 Causal 3D Conv + Self-Attention

Causal 3D Convolution



과거와 현재만 사용하는 left padding

Past & Current (미래 사용 X)

- 3D Conv로 공간 외에도 시간 방향의 local motion 특징 학습
- 인과적 패딩으로 미래 프레임 차단

Spatio-Temporal Self-Attention

멀리 떨어진 시공간 위치 간 장거리 의존성 학습
(인과적 마스크 적용)



- 멀리 떨어진 객체/동작 관계 모델링
- causal mask로 미래 정보 차단

4 토크나이저 출력과 WFM 연결

Continuous Tokenizer (연속 토큰)



연속 latent embedding
$z \in \mathbb{R}^{T' \times H' \times W' \times C}$

Diffusion WFM (Stable Diffusion류)

연속 latent 공간에서 denoising / generation

Discrete Tokenizer (이산 토큰)



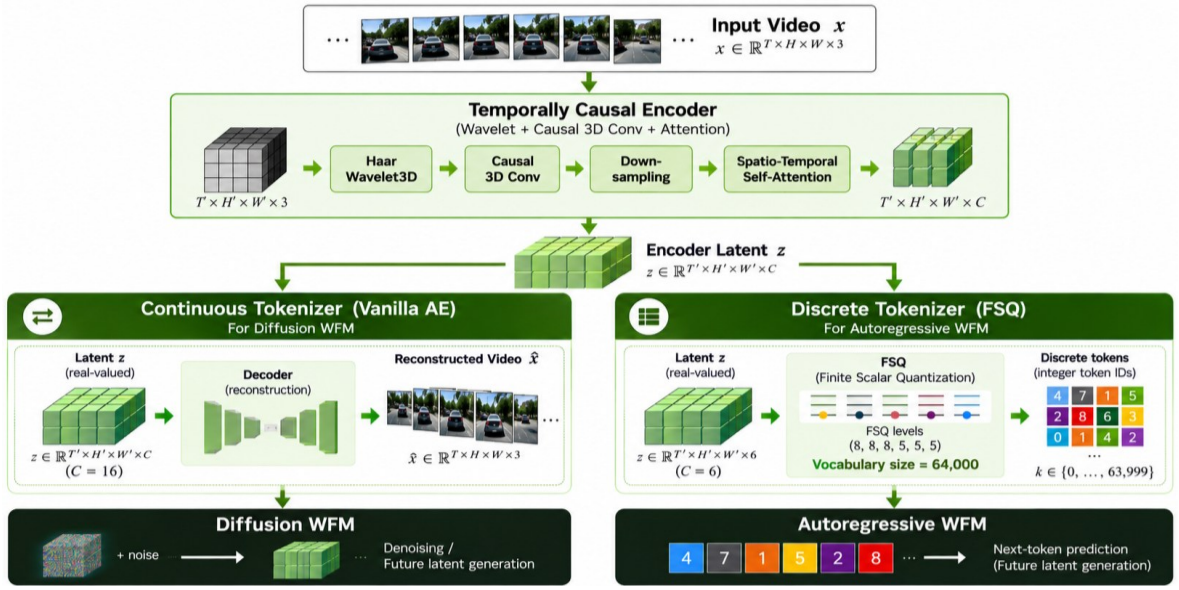
양자화된 code index
$z \in \{0, \ldots, K-1\}$

Autoregressive WFM (GPT류)

다음 토큰 예측 (Next-token prediction)

Continuous AE & Discrete FSQ

$$\text{공통 Encoder} \rightarrow \begin{cases} \text{Continuous: latent를 그대로 사용} & \rightarrow \text{Diffusion WFM} \\ \text{Discrete: latent를 FSQ로 양자화} & \rightarrow \text{Autoregressive WFM} \end{cases}$$



Continuous tokenizer — Vanilla AE

- Vanilla autoencoder(AE) formulation 사용
- latent dimension = 16
- diffusion WFM의 연속 latent로 직접 연결

Discrete tokenizer — FSQ Codebooks

- Finite-Scalar Quantization(FSQ)를 latent quantizer로 사용
- latent dimension = 6 (= FSQ level 수)

FSQ levels $(8, 8, 8, 5, 5, 5)$ 하나의 visual token을 구성하는 6차원 latent 각각의 양자화 해상도

→ vocabulary size = **64,000**

Auxiliary latent loss 미사용: decoder 최종 출력만 supervise. VAE의 KL prior loss나 VQ-VAE의 commitment loss 없이 학습 — latent space에 직접 붙는 보조 손실 제거.

Training Strategy : Joint Image-Video + 2-stage Training

Joint Image-Video Training: image·video mini-batch를 교대 사용 → image dataset의 풍부한 appearance 정보를 video tokenizer 학습에 활용

Stage 1 — Reconstruction

- 원본($x_{0:T}$)과 복원 결과($\hat{x}_{0:T}$)의 **pixel-wise RGB L1 loss**
- VGG-19 feature 기반 **perceptual loss**
- 목적: 기본 시각 재구성 품질 확보

Stage 2 — Temporal Smoothness & Sharpness

- **optical flow loss** → Reconstructed video의 시간적 일관성 개선
- **Gram-matrix loss** → 복원 이미지 선명도 향상
- Fine-tuning 단계에서 **adversarial loss** → 큰 압축률에서 reconstruction detail 강화

Stage 1 loss:

pixel-wise RGB L1 loss $$\mathcal{L}_1 = \|\hat{x}_{0:T} - x_{0:T}\|_1$$

perceptual loss $$\mathcal{L}_{\text{Perceptual}} = \frac{1}{L}\sum_{l=1}^{L}\sum_{t}\alpha_l \left\|\texttt{VGG}_l(\hat{x}_t) - \texttt{VGG}_l(x_t)\right\|_1$$

Stage 2 loss:

Optical flow loss $$\mathcal{L}_{\text{Flow}} = \frac{1}{T}\sum_{t=1}^{T}\|\texttt{OF}(\hat{x}_t, \hat{x}_{t-1}) - \texttt{OF}(x_t, x_{t-1})\|_1 + \frac{1}{T}\sum_{t=0}^{T-1}\|\texttt{OF}(\hat{x}_t, \hat{x}_{t+1}) - \texttt{OF}(x_t, x_{t+1})\|_1$$

Gram-matrix loss $$\mathcal{L}_{\text{Gram}} = \frac{1}{L}\sum_{l=1}^{L}\sum_{t}\alpha_l \left\|\texttt{GM}_l(\hat{x}_t) - \texttt{GM}_l(x_t)\right\|_1$$

압축률 구성 & 두 tokenizer 계열

image tokenizer와 video tokenizer를 다양한 compression rate로 학습

Tokenizer 구분	Compression rate
Image tokenizer (CI / DI)	8×8, 16×16
Video tokenizer (CV / DV)	4×8×8, 8×8×8, 8×16×16

Video tokenizer의 두 계열

Cosmos-0.1-Tokenizer

상대적으로 적은 수의 720p video frame으로 학습

Cosmos-Tokenize1

더 많은 수의 360p 또는 720p video frame으로 학습

→ *다양한 temporal/spatial resolution에서 유연하게 동작. 학습 해상도뿐 아니라 더 높은 해상도에서도 품질 유지.*

평가 설정: TokenBench · DAVIS · MS-COCO · ImageNet

Benchmarks & Metrics

- Image Tokenizer : MS-COCO 2017 val 5,000 · ImageNet-1K val 50,000
- Video Tokenizer : 고해상도·장시간 표준 benchmark 부재 → TokenBench 신규 구성
- TokenBench: robotic manip, driving, egocentric, web (BDD100K·EgoExo-4D·BridgeData V2·Panda-70M), 총 500개
- 추가로 DAVIS(1080p) 평가

데이터	평가 지표
Image	PSNR · SSIM · rFID
Video	PSNR · SSIM · rFVD

주요 실험결과

+4 dB
DAVIS PSNR 개선
(기존 tokenizer 대비)

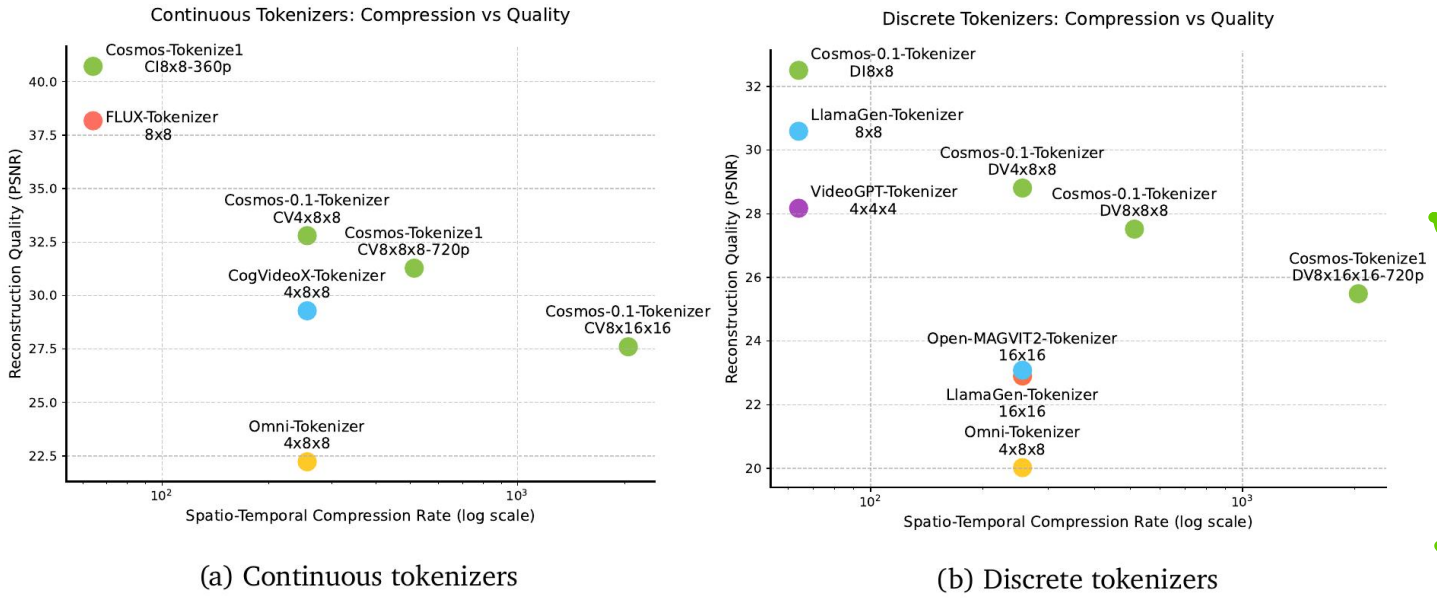
최대 12×
encoding 속도
(기존 대비)

2×~12×
image/video runtime
전반 가속

A100 80GB
1080p 8초 / 720p 10초
한 번에 encoding

평가 관점	결과
Video tokenizer	DAVIS와 TokenBench에서 SOTA 수준 성능
Compression-quality	4×8×8뿐 아니라 8×8×8, 8×16×16에서도 우수
Image tokenizer	8×8 compression에서 SOTA 수준 성능
고압축 image 품질	16×16에서도 기존 8×8 수준과 비교 가능
Runtime / Model size	기존 대비 2×~12× 빠르며 가장 작은 모델 크기 유지

평가 설정: TokenBench · DAVIS · MS-COCO · ImageNet



(a) Continuous tokenizers (b) Discrete tokenizers

Table 5: Evaluation of continuous video (CV) tokenizers on DAVIS and TokenBench.

			DAVIS			TokenBench		
Tokenizer	Frames	Formulation	PSNR ↑	SSIM ↑	rFVD ↓	PSNR ↑	SSIM ↑	rFVD ↓
CogVideoX-Tokenizer4×8×8	17	VAE	29.29	0.864	19.58	32.06	0.909	6.97
Omni-Tokenizer4×8×8	17	VAE	22.23	0.713	117.66	24.48	0.830	35.86
Cosmos-0.1-Tokenizer-CV4×8×8	49	AE	32.80	0.900	15.93	35.45	0.928	6.85
Cosmos-0.1-Tokenizer-CV8×8×8	49	AE	30.61	0.856	30.16	34.44	0.917	11.62
Cosmos-0.1-Tokenizer-CV8×16×16	49	AE	27.60	0.779	93.82	31.61	0.875	43.08
Cosmos-Tokenize1-CV4×8×8-360p	49	AE	**35.85**	**0.920**	**10.057**	**38.42**	**0.950**	**3.34**
Cosmos-Tokenize1-CV8×8×8-720p	121	AE	31.28	0.868	23.49	35.13	0.926	9.82

Table 6: Evaluation of discrete video (DV) tokenizers on DAVIS and TokenBench.

			DAVIS			TokenBench		
Tokenizer	Frames	Quantization	PSNR ↑	SSIM ↑	rFVD ↓	PSNR ↑	SSIM ↑	rFVD ↓
VideoGPT-Tokenizer4×4×4	-	VQ	28.17	**0.850**	72.33	33.66	**0.914**	**13.85**
Omni-Tokenizer4×8×8	17	VQ	20.02	0.703	188.60	25.31	0.827	53.55
Cosmos-0.1-Tokenizer-DV4×8×8	17	FSQ	28.81	0.818	**37.36**	31.97	0.888	19.67
Cosmos-0.1-Tokenizer-DV8×8×8	17	FSQ	27.51	0.789	100.15	30.95	0.873	43.86
Cosmos-0.1-Tokenizer-DV8×16×16	17	FSQ	25.09	0.714	241.52	28.91	0.829	113.48
Cosmos-Tokenize1-DV4×8×8-360p	49	FSQ	**32.97**	0.840	53.44	**35.74**	0.910	19.25
Cosmos-Tokenize1-DV8×16×16-720p	49	FSQ	25.49	0.719	259.33	29.33	0.838	107.43

논문 Figure 8 — Compression rate(log) vs Reconstruction quality(PSNR). 좌: continuous / 우: discrete
Our tokenizer demonstrates an excellent compression-quality trade-off, delivering superior quality even at higher compression rates compared to other methods. The evaluation was performed on the DAVIS dataset. We calculate the PSNR of image tokenizers on all the individual frames.

- **Cosmos-0.1-Tokenizer**: Trained using mini-batches sampling a smaller number of 720p video frames (49 frames for CV and 17 frames for DV).
- **Cosmos-Tokenize1**: Trained using mini-batches sampling a larger number of 360p or 720p video frames (121 frames for CV and 49 frames for DV).

WFM 전체 구조에서 4장의 위치

3장	4장	5장	6장
Data Curation	**Tokenizer**	WFM Pre-training	Post-training
고품질 clip 확보	compact token 변환	token 기반 학습	task 특화

Physical AI 관점의 의미

실시간성

지속 관측·판단을 위해 raw video를 효율적으로 압축

인과성

미래 frame 미참조 → 실제 시간 흐름 시스템과 정합

세계표현 품질

물체·공간·움직임·세부 보존; 재구성 품질이 WFM 예측 품질의 상한

> **Cosmos Tokenizer = Physical AI가 관측한 세계를 WFM이 이해·예측할 수 있는 latent world representation으로 바꾸는 시각-시간 표현 엔진.**

ONE-SENTENCE SUMMARY

4장「**Tokenizer**」는 Cosmos WFM 학습을 위해 고해상도 image/video를 continuous 또는 discrete compact token으로 변환하는 Cosmos Tokenizer를 설명하며, temporally causal encoder–decoder, wavelet 기반 압축, spatio-temporal convolution/attention, joint image-video training을 통해 **높은 압축률·재구성 품질·빠른 runtime**을 동시에 달성하는 WFM의 핵심 표현 계층을 제시한다.

NVIDIA COSMOS · WORLD FOUNDATION MODEL PLATFORM

Chapter 5 part1

WFM Pre-training: Diffusion

범용 월드모델(generalist) 구축 · EDM denoising · DiT 아키텍처 · 학습 전략 · 스케일업 · prompt upsampler

Cosmos의 WFM pre-training은 대규모 비디오 데이터를 기반으로 물리 세계의 시각적 변화, 움직임, 3D 일관성, 기초 물리 동역학을 학습하는 범용 세계모델 구축 과정이다.

arXiv:2501.03575v3 · §4 Tokenizer

Cosmos World Foundation Model Pre-training

5. World Foundation Model Pre-training 핵심정리

NVIDIA

Physical AI를 위한 Cosmos WFM 사전학습 구조와 평가

1 핵심 메시지

사전학습된 World Foundation Models(WFM)은 대규모 비디오 토큰을 통해 현실세계의물리 법칙과 자연스러운 행동을 학습하는 범용 세계 시뮬레이터이다.

- 범용 WFM은 현실 세계의 물리 동역학과 자연스러운 행동 패턴을 학습
- Diffusion WFM과 Autoregressive WFM의 두 계열로 구성
- 이후 post-training을 통해 camera, robot, driving 등 도메인으로 특화

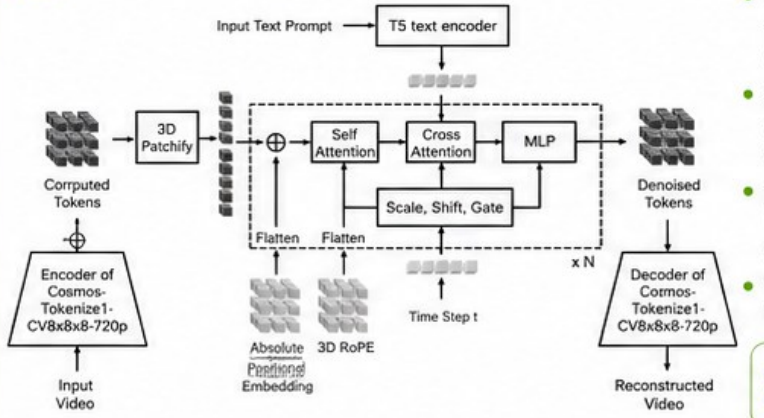
★ 핵심: Cosmos의 pre-training은 Physical AI용 범용 세계 시뮬레이터를 만드는 단계이다.

2 모델 맵: Two-family WFM

Diffusion-based WFM	Autoregressive-based WFM
Continuous tokens	Discrete tokens
• 7B / 14B Text2World	• 4B / 12B Next-token WFM
• 7B / 14B Video2World	• 5B / 13B Video2World
• 고품질 시각 생성	• 빠른 token rollout
• Text + Video conditioning	• LLM-like 확장성
• Prompt Upsampler 사용	• Diffusion Decoder 연계



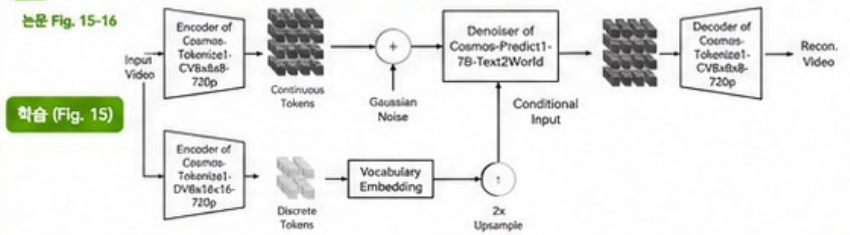
3 Diffusion WFM 구조



- Continuous latent space에서 denoising 기반 world generation
- 3D patchification + self-attention + cross-attention + MLP
- T5-XXL text encoder, FPS-aware 3D RoPE, AdaLN-LoRA
- Text2World에서 Video2World로 확장

AdaLN-LoRA로 11B → 7B 수준의 파라미터 효율화

4 Autoregressive WFM + Diffusion Decoder



- Discrete video token의 next-token prediction으로 미래 video 생성
- 3D RoPE, cross-attention, z-loss, KV caching, Medusa speculative decoding
- Diffusion decoder가 discrete output을 continuoos token으로 복원해 blur와 artifact 완화

320×512 저해상도 적용에서 8×H100 기준 약 10 FPS 실시간 생성

5 학습 전략과 스케일링

- **Joint image-video training**
  이미지와 비디오를 함께 학습하여 표현력을 강화
- **Progressive training**
  512p 57 frames → 720p 121 frames → high-quality fine-tuning
- **Multi-aspect training**
  1:1, 3:4, 4:3, 9:16, 16:9
- **대규모 학습 인프라**
  약 10,000 H100 GPU, 약 3개월 학습
- **Diffusion scale-up**
  FSDP + Context Parallelism
- **Autoregressive scale-up**
  Tensor Parallelism + Sequence Parallelism

AI Factory 규모의 WFM 학습 체계

6 평가: 3D Consistency & Physics Alignment

3D Consistency

모델	VideoLDM pose success rate
VideoLDM	4.4%
Cosmos-Predict1-7B-Text2World	62.6%
Cosmos-Predict1-7B-Video2World	68.4%

핵심 의미: 2D 생성이 아니라 기하학적으로 더 일관된 3D world simulation

Physics Alignment

논문 Fig. 20



- PhysX / Isaac Sim 기반 물리 시나리오 평가
- 1-frame보다 9-frame conditioning이 더 우수
- rigid-body physics는 아직 한계 존재

⚠ 한계: object impermanence, deformation, implausible kinematics

5장의 핵심 결론

1. Pre-training은 Physical AI용 범용 WFM을 만든다
2. Diffusion은 품질, Autoregressive는 확장성과 속도에 강하다
3. Prompt Upsampler와 Diffusion Decoder가 실사용 성능을 보강한다
4. 3D consistency와 physics alignment 평가는 WFM의 핵심 검증 축이다

★ 한 줄 요약: Cosmos의 WFM pre-training은 대규모 비디오 토큰 학습을 통해 Physical AI가 예측·계획·시뮬레이션할 수 있는 범용 월드모델을 구축하는 핵심 단계이다.

Cosmos Diffusion-based Model Foundation Model

5.1 Diffusion-based World Foundation Model 핵심정리

Physical AI를 위한 Cosmos WFM의 Diffusion 기반 사전학습 구조

1 핵심 메시지

- 사전학습된 Diffusion WFM은 대규모 비디오 토큰을 바탕으로 현실 세계의 물리 동역학과 자연스러운 시각 변화를 학습하는 범용 세계 시뮬레이터
- Continuous video token latent space에서 denoising 기반으로 미래 world를 생성
- Text2World에서 시작해 Video2World로 확장

★ 핵심: Diffusion WFM은 Cosmos의 고품질 world generation 핵심 엔진이다.

2 기본 개념과 Formulation

Input Video → Continuous Token → Gaussian Noise → Diffusion Denoiser → Generated Token → Reconstructed Video

$D_\theta(x_0 + n; \sigma) \rightarrow x_0$ 복원

EDM 기반 denoising score matching

- latent diffusion으로 학습·추론 비용 절감
- 다양한 noise level에서 안정적 복원 학습
- uncertainty-aware weighting으로 학습 균형 확보

4 학습 전략

- Joint image-video training: 이미지와 비디오를 함께 학습하여 appearance와 dynamics 동시 확보
- Progressive training: 512p 57 frames → 720p 121 frames → high-quality fine-tuning
- Multi-aspect training: 1:1, 3:4, 4:3, 9:16, 16:9
- Text conditioning: T5-XXL 사용
- Video2World 확장: conditional frames + noise augmentation + binary mask

6 Prompt Upsampler와 결과 해석

A. Prompt Upsampler

- 짧은 사용자 prompt를 학습 분포에 맞는 풍부한 설명으로 확장
- 목표: fidelity, training-distribution alignment, visual detail enhancement
- Text2World: Mistral-NeMo-12B-Instruct fine-tuning
- Video2World: Pixtral-12B 기반 upsampling

B. 결과 해석

- 7B / 14B Text2World 모두 우수한 visual quality와 motion dynamics
- 14B가 더 복잡한 scene richness와 motion stability 제공
- Video2World는 image/video conditioning과 extended generation 지원

3 Architecture (논문 Fig. 11)

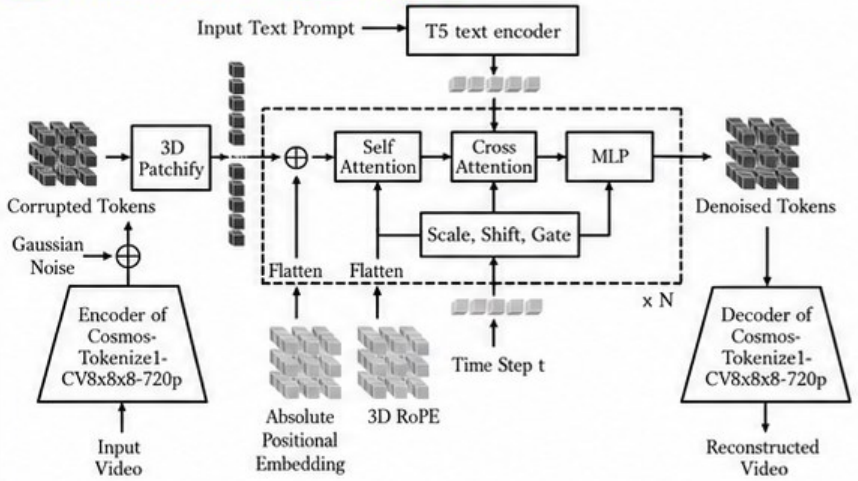


Figure 11: Overall architecture of Cosmos-Predict1 World Foundation Model.

Architecture 요약

- 3D patchification
- self-attention + cross-attention + MLP
- T5-XXL text conditioning
- FPS-aware 3D RoPE + learnable positional embedding
- AdaLN-LoRA로 파라미터 효율화

AdaLN-LoRA: 11B → 7B 수준 파라미터 절감

5 스케일링과 인프라

- 약 10,000 H100 GPU, 약 3개월 학습
- FSDP로 parameter / gradient / optimizer state 분산
- Context Parallelism으로 긴 video context 처리
- selective activation checkpointing으로 메모리 절감
- AI Factory 규모의 diffusion WFM 학습 체계

대규모 병렬화가 diffusion WFM 학습의 핵심 기반

5.1의 핵심 결론

1. Diffusion WFM은 latent world denoising 기반 생성 모델이다.
2. Fig. 11 구조가 Cosmos diffusion architecture의 핵심이다.
3. Joint training + progressive training이 성능 향상의 핵심이다.
4. Prompt upsampler와 대규모 병렬화가 실사용성을 높인다.

★ 한 줄 요약: Cosmos의 Diffusion-based WFM은 continuous video token latent space에서 물리 세계를 고품질로 생성·예측하는 범용 world generator이다.

범용 세계모델(generalist) 구축 과정

Pre-trained WFM은 특정 시스템에 맞춘 모델이 아니라, **현실 세계의 일반적 물리 동역학·자연스러운 행동 패턴을 학습한 범용 세계모델**. 대규모 비디오로 시각 변화·움직임·3D 일관성·기초 물리를 학습 (이후 6장에서 task별 post-training).

계열	핵심 학습 원리	token
Diffusion WFM	Denoising 으로 비디오 생성	Continuous video token
Autoregressive WFM	next-token prediction (다음 token 예측)	Discrete video token

AI Factory 규모: 모든 WFM을 **10,000개 NVIDIA H100 GPU** 클러스터에서 약 **3개월** 학습 — GPU 인프라·tokenization·병렬화·데이터 curriculum이 결합된 학습 체계.

Cosmos Pre-trained WFM 제품군 (Table 10)

Table 10: Maps of Cosmos World Foundation Model Map. We have two sets of WFMs. One is based on **diffusion models**, while the other is based on **autoregressive models**. For each family, we build two base models and two derivative models. To achieve the best generation quality, we also build a **prompt upsampler for the diffusion models** and a **diffusion decoder for the autoregressive models**.

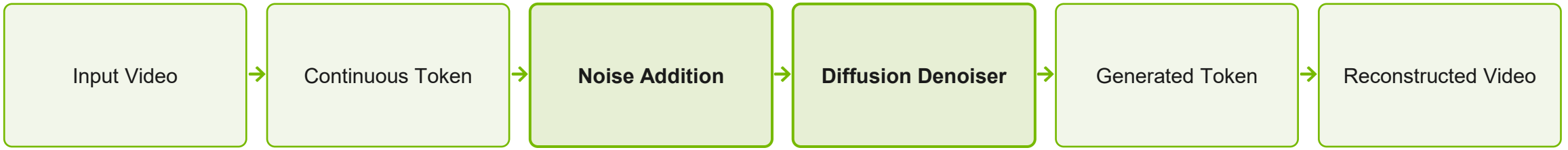
Type	Diffusion	Autoregressive
Models	1. Cosmos-Predict1-7B-Text2World → Cosmos-Predict1-7B-Video2World	1. Cosmos-Predict1-4B → Cosmos-Predict1-5B-Video2World
	2. Cosmos-Predict1-14B-Text2World → Cosmos-Predict1-14B-Video2World	2. Cosmos-Predict1-12B → Cosmos-Predict1-13B-Video2World
Tokenizer	Cosmos-Tokenize1-CV8×8×8-720p	Cosmos-Tokenize1-DV8×16×16-720p
Enhancer	Cosmos-UpsamplePrompt1-12B-Text2World	Cosmos-Predict1-7B-Decoder-DV8×16×16ToCV8×8×8-720p

Diffusion 계열 Text2World(텍스트→비디오 세계) 학습 후, 과거 video + text를 받아 미래를 생성하는 Video2World로 fine-tuning.

Autoregressive 계열 과거 video token으로 다음 token 예측을 학습 후, T5 text embedding을 cross-attention으로 결합한 text-conditioned Video2World로 확장.

기본 개념: latent diffusion model

raw video pixel이 아니라, 4장 continuous video tokenizer가 만든 compact latent 공간에서 denoising → 학습·추론 비용 절감, denoising 단순화.



해석: diffusion WFM은 '비디오를 직접 생성하는 모델'이라기보다, **비디오 latent world를 점진적으로 복원하면서 미래 세계를 생성**하는 모델이다. tokenizer encoder가 latent를 만들고, Gaussian noise를 더한 뒤 denoiser가 복원하며, 마지막에 tokenizer decoder가 비디오로 되돌린다.

Diffusion WFM은 토크나이저의 학습된 잠재 공간 내에서 작동하는 잠재 확산 모델(latent diffusion model)로, 비디오의 압축된 저차원 표현을 가능하게 함.
학습 및 추론 중 계산 비용을 절감하는 동시에 노이즈 제거 작업을 단순화함 (Hoogeboom et al., 2024; Rombach et al., 2022).
비디오를 잠재 표현으로 토큰화하기 위해 Cosmos-Tokenize1-CV8×8×8-720p를 사용함

EDM 기반 Denoising Score Matching

깨끗한 image/video latent x_0에 Gaussian noise를 더한 뒤, denoiser D_θ가 원래 clean sample을 복원하도록 학습함.
*또한 noise level별 학습 난이도를 multi-task learning처럼 다루기 위해 **uncertainty-based weighting**을 도입함.*

Denoising score matching loss (noise level σ)

$$\mathcal{L}(D_\theta, \sigma) = \mathbb{E}_{x_0, n}\left[\|D_\theta(x_0 + n; \sigma) - x_0\|_2^2 \right] \tag{5}$$

D_θ를 매개변수화하기 위해 EDM에 도입된 사전 조건화(preconditioning) 설계를 따름.
전체 학습 손실은 노이즈 수준에 걸친 $L(D_\theta; \sigma)$의 가중 기대값으로 정의됨

전체 학습 손실 (noise level에 대한 가중 기대값)

$$\mathcal{L}(D_\theta) = \mathbb{E}_\sigma \left[\frac{\lambda(\sigma)}{e^{u(\sigma)}} \mathcal{L}(D_\theta, \sigma) + u(\sigma) \right] \tag{6}$$

$$\lambda(\sigma) = \frac{\sigma^2 + \sigma_{\text{data}}^2}{(\sigma \cdot \sigma_{\text{data}})^2} \tag{7}$$

with

$$\ln(\sigma) \sim \mathcal{N}(P_{\text{mean}}, P_{\text{std}}^2)$$

노이즈 수준 σ의 분포는 하이퍼파라미터 P_{mean}, P_{std}에 제어 (8)

σ_{data}는 학습 데이터의 표준 편차이며, 가중치 함수 $\lambda(\sigma)$는 학습 시작 시 각 노이즈 수준의 동일한 기여를 보장함. 그러나 학습이 진행됨에 따라 이러한 균형은 악화될 수 있음. 이 문제를 완화하기 위해 다양한 노이즈 수준에 대한 최적화를 다중 작업 학습(multi-task learning)의 한 형태로 취급함.

핵심 요소

x_0	원본 image/video token (clean latent)
σ	noise level — 다양한 난이도의 denoising
D_θ	noise-conditioned denoiser network

Uncertainty-based weighting

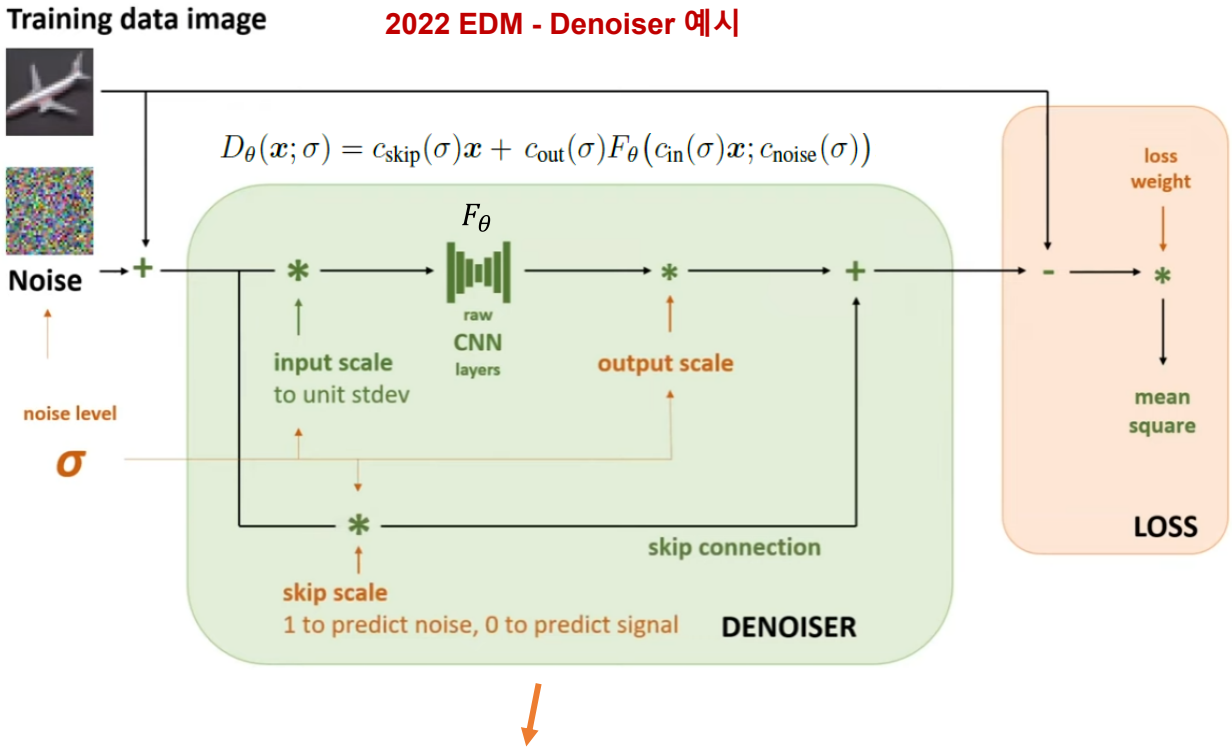
noise level별 학습을 multi-task learning으로 보고, **불확실성 함수 u(σ)**로 균형 조절. 단일 denoising이 아니라 여러 noise scale에서 안정적으로 월드 latent를 복원하도록 학습.

노이즈 수준 σ에서의 노이즈 제거 목표 $L(D_\theta, \sigma)$에 대한 불확실성을 정량화하는 연속 불확실성 함수로서 $\mu(\sigma)$를 도입하여 불확실성 기반 가중치 접근 방식을 활용함

간단한 MLP를 사용하여 $\mu(\sigma)$를 매개변수화하고 학습 중에 전체 손실 $L(D_\theta)$를 최소화함. 직관적으로, 모델이 Task에 대해 불확실한 경우, 즉 $\mu(\sigma)$가 높으면 노이즈 수준 σ에서의 손실 기여도는 낮게 가중됨. 동시에 모델은 이러한 불확실성에 대해 페널티를 받으며, $\mu(\sigma)$가 가능한 한 낮아지도록 유도됨

EDM 기반 Denoising Score Matching

EDM : Tero Karras, Miika Aittala, Timo Aila, Samuli Laine, "Elucidating the Design Space of Diffusion-Based Generative Models," NeurIPS 2022.



Label-conditioned image generation을 위해 설계된 **DiT**(Peebles and Xie, 2023)를 기반으로 하는 **Denoiser network D_θ**설계

$$D_\theta(\boldsymbol{x};\sigma) = c_{\text{skip}}(\sigma)\,\boldsymbol{x} + c_{\text{out}}(\sigma)\,F_\theta\big(c_{\text{in}}(\sigma)\,\boldsymbol{x};\; c_{\text{noise}}(\sigma)\big)$$

- ✓ F_θ : the neural network to be trained
- ✓ $c_{\text{skip}}(\sigma)$: Modulates the skip connection
- ✓ $c_{\text{in}}(\sigma)$, $c_{\text{out}}(\sigma)$: Scale the input and output magnitudes
- ✓ $c_{\text{noise}}(\sigma)$: Map noise level σ into a conditioning input for F_θ.
- ✓ Taking a weighted expectation of Eq. 2 over the noise levels gives the overall training loss

$$\mathbb{E}_{\sigma,\boldsymbol{y},\boldsymbol{n}}\left[\lambda(\sigma)\,\|D(\boldsymbol{y}+\boldsymbol{n};\sigma)-\boldsymbol{y}\|_2^2\right]$$

where $\sigma \sim p_{\text{train}}$, $\boldsymbol{y} \sim p_{\text{data}}$, and $\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$

$$\mathbb{E}_{\sigma,\boldsymbol{y},\boldsymbol{n}}\Big[\underbrace{\lambda(\sigma)\,c_{\text{out}}(\sigma)^2}_{\text{effective weight}}\,\big\|\underbrace{F_\theta\big(c_{\text{in}}(\sigma)\cdot(\boldsymbol{y}+\boldsymbol{n});\,c_{\text{noise}}(\sigma)\big)}_{\text{network output}}-\underbrace{\tfrac{1}{c_{\text{out}}(\sigma)}\big(\boldsymbol{y}-c_{\text{skip}}(\sigma)\cdot(\boldsymbol{y}+\boldsymbol{n})\big)}_{\text{effective training target}}\big\|_2^2$$

Architecture of	F_θ	(any)
Skip scaling	$c_{\text{skip}}(\sigma)$	$\sigma_{\text{data}}^2/\left(\sigma^2+\sigma_{\text{data}}^2\right)$
Output scaling	$c_{\text{out}}(\sigma)$	$\sigma\cdot\sigma_{\text{data}}/\sqrt{\sigma_{\text{data}}^2+\sigma^2}$
Input scaling	$c_{\text{in}}(\sigma)$	$1/\sqrt{\sigma^2+\sigma_{\text{data}}^2}$
Noise cond.	$c_{\text{noise}}(\sigma)$	$\frac{1}{4}\ln(\sigma)$

DiT 기반 비디오 생성 구조

Label-conditioned image generation을 위해 설계된 DiT(Peebles and Xie, 2023)를 기반으로 하는 **Denoiser network D_θ**설계 –
Cosmos Diffusion WFM이 단순 text-to-video diffusion 모델이 아니라, **Physical AI를 위한 controllable video world simulator**가 되도록 DiT 구조를 재설계

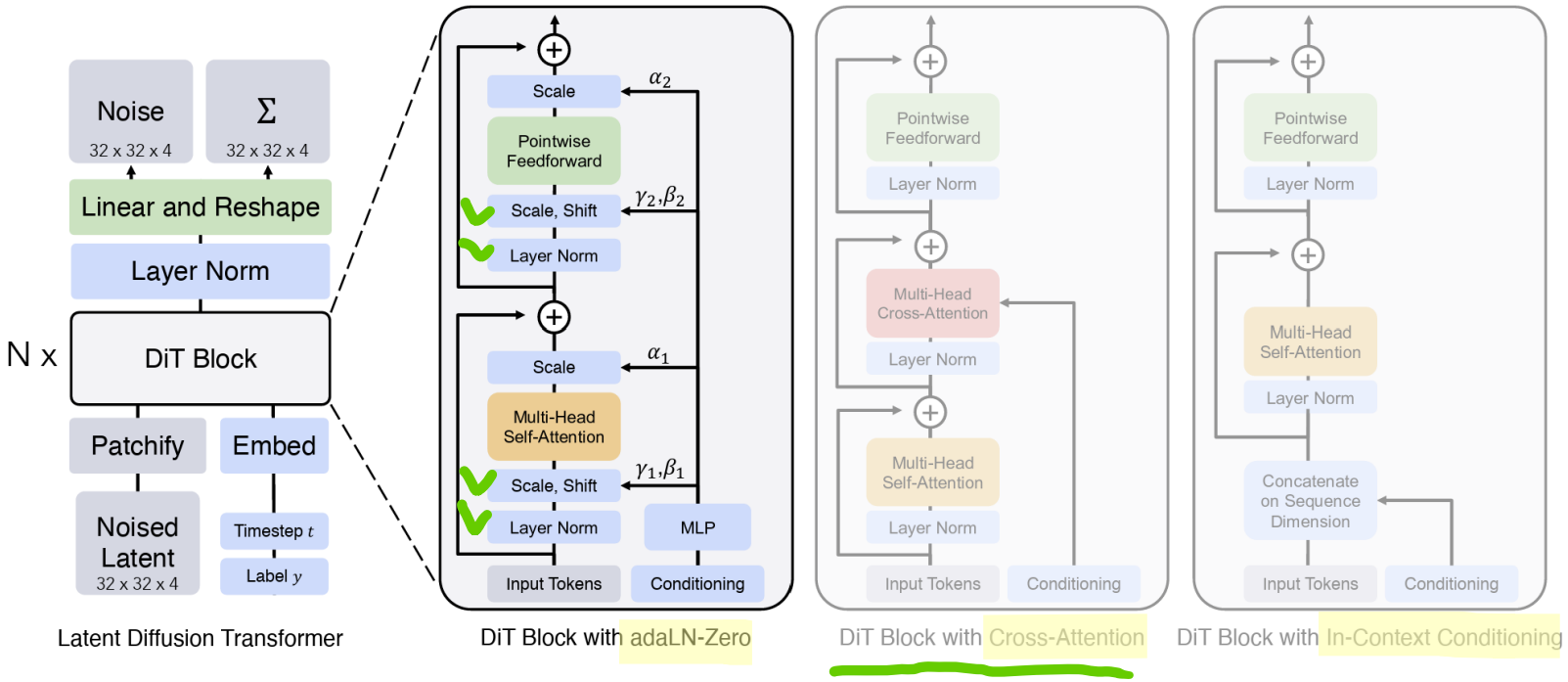


Figure 3. **The Diffusion Transformer (DiT) architecture.** *Left:* We train conditional latent DiT models. The input latent is decomposed into patches and processed by several DiT blocks. *Right:* Details of our DiT blocks. We experiment with variants of standard transformer blocks that incorporate conditioning via adaptive layer norm, cross-attention and extra input tokens. Adaptive layer norm works best.

Scalable Diffusion Models with Transformers
William Peebles, Saining Xie; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4195-4205

adaLN-Zero (Adaptive Layer Normalization-Zero)

- 0으로 초기화 (Zero-initialization): Timestep(t)과 Class(c) 등의 조건(Conditioning) 정보를 입력받아 레이어 정규화의 이동(shift, β)과 스케일(scale, γ) 파라미터를 출력하는 작은 신경망(MLP)의 최종 가중치와 편향을 모두 0으로 초기화
- 차원별 스케일 파라미터 (α): adaLN에 추가적으로 residual 연결 직전에 차원별 스케일 파라미터 α를 도입하며, 이 α를 출력하는 MLP 역시 0으로 초기화
- Identity Function로 시작: 학습을 시작하는 초기 시점에 α = 0, β = 0, γ = 0이 되므로, 첫 DiT 블록은 입력을 그대로 내보내는 항등 함수처럼 작동하여 학습 초기의 안정성을 극대화함

DiT 기반 비디오 생성 구조

Label-conditioned image generation을 위해 설계된 DiT(Peebles and Xie, 2023)를 기반으로 하는 **Denoiser network D_θ**설계 –
Cosmos Diffusion WFM이 단순 text-to-video diffusion 모델이 아니라, **Physical AI를 위한 controllable video world simulator**가 되도록 DiT 구조를 재설계

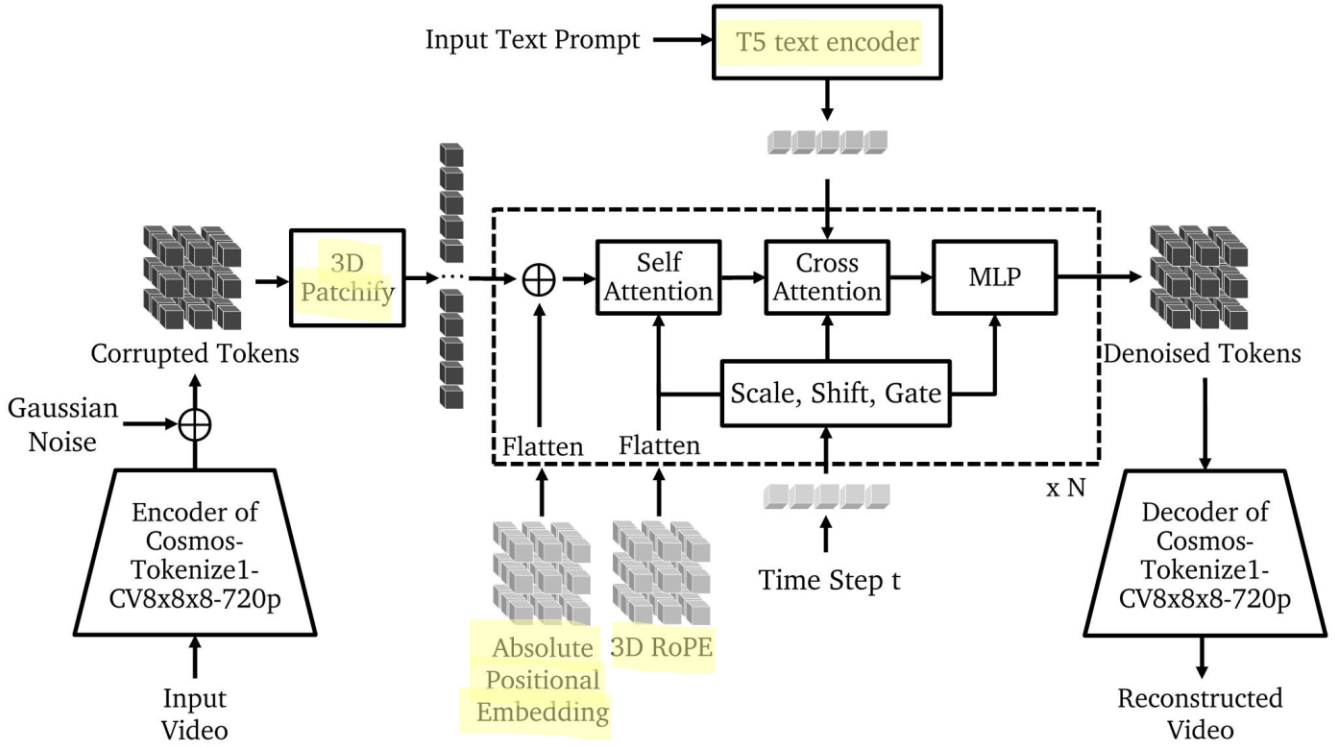


Figure 11 — Cosmos-Predict1 WFM 전체 구조: tokenizer encoder → Gaussian noise → 3D patchify → (Self-Attn · Cross-Attn · MLP) ×N → tokenizer decoder

설계 요소	역할
3D Patchification	latent → 시공간 token sequence (시간–공간 구조를 transformer token으로 변환)
FPS-aware 3D RoPE	시간T·H·W 위치 정보 반영 (다양한 FPS, 해상도, aspect ratio, video length 대응)
Learnable Absolute PE	morphing artifact 감소 및 학습 안정화
T5-XXL Cross-attention	text-conditioned controllable world generation
Query-Key Norm	attention logit 안정화 (attention collapse 방지)
AdaLN-LoRA	AdaLN parameter 효율화 (7B/14B급 모델의 parameter-efficient scaling)

adaLN-Zero : Large model일수록 adaLN 회귀를 위한 MLP 노드의 파라미터 크기가 급격히 커져 연산 병목(Bottleneck) 유발
→ adaLN-LoRA : LoRA(LowRank adapter) 구조로 학습 대상 파라미터를 극적으로 줄이고 연산 속도를 높임

핵심 설계 하이라이트

3D Patchification

- video latent (TCHW)를 시공간 token sequence (THW/$p_t p_h p_w$)로 변환 (image는 1-frame video로 통합 처리)
  - Denoiser 입력 위해, 상태를 patchify한 다음 flatten함. 이 과정에서 (p_t, p_h, p_w) 형태의 겹치지 않는 Cube를 네트워크의 개별 토큰 입력으로 투영하는 것을 포함함. THW/$p_t p_h p_w$ 길이의 1차원 시공간 시퀀스로 재구성됨. ($p_t = 1, p_h = p_w = 2$ 사용)

FPS-aware 3D RoPE

- 시간·높이·너비를 3분할해 RoPE 적용. FPS에 따라 temporal frequency 재조정 → 다양한 해상도/길이/종횡비 생성
  - 학습 비디오의 **FPS에 따라 시간 주파수를 재조정**하여 서로 다른 프레임 속도의 비디오 생성에서도 일관성을 유지함. 여기에 **학습 가능한 절대 위치 임베딩**과 **NTK-RoPE**를 결합해 점진적 학습 시 빠른 수렴과 생성 영상의 morphing artifact 감소를 유도함

Cross-attention for Text Conditioning (T5-XXL)

- 텍스트 prompt 조건을 각 block에서 반영
  - 각 트랜스포머 블록은 순차적인 Self-attention, Cross-attention 및 feedforward 계층으로 구성됨. Self-attention이 시공간 토큰에 대해 작동하는 반면, Cross-attention은 T5-XXL 임베딩을 키(key)와 값(value)으로 사용하여 의미론적 맥락을 통합하여 텍스트 조건화함

Query-Key Normalization

- 학습 초기 단계에서 attention logit 증가가 불안정하여 이를 완화
  - 어텐션 연산 이전에 쿼리 Q와 키 K를 정규화함. 네트워크 내의 모든 Self-attention 및 Cross-attention 계층에 대해 학습 가능한 스케일을 포함하는 Root Mean Square Normalization (RMSNorm)을 사용

AdaLN-LoRA

DiT의 adaptive layer norm(AdaLN) 계층들의 Dense linear projections을 Low-rank approximations로 분해하기 위해 LowRank Adaptation (LoRA)을 구현
Cosmos-Predict1-7B에서 parameter를 **11B → 7B (약 36% 감소)** 하면서도 성능을 유지.

핵심 설계 하이라이트

Cosmos-Predict1 diffusion 계열 모델의 주요 구성

Text2World는 text prompt와 FPS 정보를 조건으로 사용하고, Video2World는 conditional frames, 즉 과거 관측 frame을 추가 조건으로 사용

Predict2.5는 하나의 diffusion/flow backbone으로 텍스트, 이미지, 비디오 조건을 모두 처리하는 구조임. 공식 설명에서도 Predict2.5는 flow-based model이며, Text2World·Image2World·Video2World를 통합하고, 200M curated video clips와 RL 기반 post-training을 통해 품질과 prompt alignment를 개선

Table 11: Configuration details of Cosmos-Predict1 models.

Configuration	7B-Text2World	14B-Text2World	7B-Video2World	14B-Video2World
Number of Layers	28	36	28	36
Model Dimension	4,096	5,120	4,096	5,120
FFN Hidden Dimension	16,384	20,480	16,384	20,480
AdaLN-LoRA Dimension	256	256	256	256
Number of Attention Heads	32	40	32	40
Number of Key / Value Heads	32	40	32	40
MLP Activation	GELU			
Positional Embedding	Hybrid positional embedding			
Conditional Information	Text; FPS	Text; FPS	Text; FPS; Frames	Text; FPS; Frames
Base Learning Rate	2^{-15}	2^{-16}	2^{-15}	2^{-16}
Weight decay	0.1	0.2	0.1	0.2
Learning Rate Warmup	Linear scheduler with 2,500 iterations			
AdamW momentum and ϵ	$\beta_1, \beta_2 = 0.9, 0.99$; $\epsilon = 10^{-10}$			

항목	Cosmos-Predict1	Cosmos-Predict2.5
생성 패러다임	EDM 기반 latent diffusion	Flow Matching 기반 velocity prediction
모델 분리	Text2World → Video2World로 단계적 확장	Text2World, Image2World, Video2World를 단일 모델로 통합
Text Encoder	T5 계열 encoder	Cosmos-Reason1 VLM
위치 임베딩	Hybrid: 3D RoPE + learnable absolute PE	3D RoPE 중심, 절대 위치 임베딩 제거
조건 입력	Text, FPS, frames	Text + image/video condition
후처리/품질 향상	Prompt upsampler 등	model merging + RL 기반 post-training

Joint image-video + Progressive + Multi-aspect training

① Joint Image-Video Training

➢ 이미지의 고품질 시각 정보와 비디오의 시간적 물리 변화를 동시에 학습

- 이미지 batch와 비디오 batch를 번갈아 학습하여, 풍부한 **이미지** 데이터의 **appearance** 정보와 **비디오** 데이터의 **temporal dynamics**를 함께 학습함.
- 이미지와 비디오 latent 분포 차이를 줄이기 위해 **domain-specific normalization**을 적용하고, 비디오 latent는 frame-wise/channel-wise 표준화를 수행함

② Progressive Training

➢ 계산 안정성 확보, 긴 비디오·고해상도 생성 능력 점진적 확보

- (초기 대규모 학습) **낮은 해상도**인 **512p, 57 frames**로 학습하고, 이후 (목표 해상도 장시간 학습) **720p, 121 frames**로 확장함.
- 대규모 pre-training 이후에는 **고품질 subset**으로 약 O(10k) iterations 동안 learning rate를 선형 감소시키며 fine-tuning함.

③ Multi-aspect Training

➢ 다양한 화면비와 해상도의 비디오 생성을 안정적으로 지원

- 1:1, 3:4, 4:3, 9:16, 16:9 5개 aspect ratio bucket으로 다양한 종횡비 image/video 처리.

④ Text & Image/Video Conditioning

- Text2World는 T5-XXL text encoder. Video2World는 이전 frame을 conditional frame으로 입력 — conditional frame에 noise augmentation, binary mask로 조건/생성 frame 구분.

FSDP & Context Parallelism

Diffusion WFM의 Scaling UP ; 메모리 요구사항, 병렬화 전략

메모리 요구사항: 예) 14B Text2World — 모델 parameters, gradients, optimizer states **≈ 280GB** + activations **≈ 310GB**. 단일 H100 80GB로는 불가능
→ 병렬화 필수 (FSDP(Fully Sharded Data Parallelism)와 CP(Context Parallelism)으로 메모리 분산)

Fully Sharded Data Parallelism (FSDP)

❖ **모델의 크기(가중치)로 인한 메모리 부족**

- **parameter · gradient · optimizer state**를 장치 간에 샤딩(**sharding**)하여 메모리 효율성을 향상함
  - 연산 중에 필요할 때만 파라미터를 수집하고 이후에 해제함.
- FSDP는 parameter · gradient · optimizer state를 분산하며 각 장치는 자신의 Shard만 관리함
- 구현에서는 메모리와 통신 지연 시간의 균형을 맞추기 위해 7B 모델에는 32, 14B 모델에는 64의 Sharding 계수를 사용함

Context Parallelism (CP)

❖ **긴 비디오 시퀀스로 인한 어텐션(Activation) 메모리 폭발**

- **입력 데이터의 시퀀스 축 (프레임/공간 패치 수) 분산**
- CP는 Cosmos Diffusion WFM이 720p, 장시간 비디오처럼 매우 긴 spatiotemporal token sequence를 학습할 수 있게 하는 핵심 병렬화 전략임
- **FSDP**가 **model parameters·optimizer states를 분산**한다면, **CP**는 **activation과 attention computation을 sequence 방향으로 분산**하는 역할을 함

→ Cosmos는 모델 아키텍처뿐 아니라, 대규모 video diffusion WFM을 실제로 학습 가능하게 만드는 시스템 최적화를 함께 제시.

FSDP & Context Parallelism

Diffusion WFM의 Scaling UP ; 메모리 요구사항, 병렬화 전략

Context Parallelism (CP)

- 목적 : 긴 비디오 sequence를 처리할 때 발생하는 **self-attention의 메모리 사용량**과 **FLOPs 증가** 문제를 **여러 GPU로 분산**해 해결하는 병렬화 기법임
- 기본 원리
  - Query Q와 Key/Value (K, V)를 sequence dimension 기준으로 CP_SIZE 개의 chunk로 나눈다 (CP_SIZE: CP 그룹에 포함된 GPU 수)
  - 각 GPU는 자신에게 할당된 Q chunk를 처리하고, CP 그룹 내의 (K, V) block을 순차적으로 받아 부분 attention output을 누적함
- 통신 방식
  - CP 구현 방식에는 all-gather, P2P, all-to-all 등이 있음
  - Cosmos는 NVIDIA TransformerEngine의 **P2P** 방식을 사용함
  - P2P 방식은 GPU 간 (K, V) block 전송과 attention 계산을 동시에 수행하여 통신과 계산을 중첩함

- 성능 최적화
  - block size를 적절히 설정하면 GPU 간 데이터 전송 지연을 계산 시간으로 가릴 수 있어 communication latency를 효과적으로 숨길 수 있음
  - CP 그룹은 NVLink로 연결된 GPU 내부에 구성하여 고속 GPU 간 통신을 활용함
  - CP rank와 FSDP rank를 중첩하여 메모리 분산과 계산 효율을 동시에 높임
- 적용 범위
  - 긴 context를 갖는 **비디오 학습 iteration**에서는 **CP 사용**하며, sequence length가 짧은 이미지 iteration에서는 throughput 향상을 위해 CP를 비활성화함
  - Cross-attention layer는 text embedding의 (K, V) sequence 길이가 짧아 통신 지연을 숨길 만큼 계산량이 충분하지 않기 때문에 CP를 적용하지 않음

[Cosmos-Predict1-14B를 예로 들면]

- 64 Sharding 계수로 FSDP를 적용; parameter · gradient · optimizer state에 대한 메모리 요구 사항이 280 GB에서 GPU당 약 280 / 64 ≈ 4 GB로 감소함.
- CP_SIZE=8인 CP를 적용; 활성화 메모리가 310 GB에서 GPU당 약 310 / 8 ≈40 GB로 감소함.
- 실제로는 토크나이저와 샤딩되지 않은 파라미터에 의해 추가 메모리가 소비됨. CP에서 통신과 계산을 중첩하려면 각 GPU가 여러 개의 (K,V) 청크를 유지해야 함

학습/추론 prompt 분포 gap 완화

학습 prompt는 VLM이 만든 상세 description이지만, 사용자는 짧고 불완전한 prompt를 입력 → 분포 gap 발생. prompt upsampler가 이를 변환.

Prompt upsampler의 주요 요구사항

1 Fidelity to the input prompts

원래 prompt의 주요 객체·행동·의도 보존

2 Training distribution alignment

WFM 학습 prompt의 길이·구조·스타일과 유사하게 변환

3 Enhanced visual details

WFM이 더 정확한 이미지를 생성 유도하도록 설계

Text2World용

Mistral-NeMo-12B-Instruct를 fine-tuning하여 구축.

Video2World용

Pixtral-12B를 zero-shot prompt engineering으로 활용 (video condition + user prompt 동시 고려).

Pixtral-12B : Mistral AI가 공개한 첫 번째 Multimodal Language Model (2024)

Cosmos-1.0-Prompt-Upsampler-12B-Text2World

Diffusion WFM 결과

7B / 14B Text2World

높은 visual quality · motion dynamics · text alignment. 14B는 더 복잡한 visual detail과 intricate motion 생성에 강함.

7B / 14B Video2World

image/video conditioning 모두 지원, autoregressive 방식으로 긴 video generation 가능. 14B는 더 높은 scene richness·motion stability.

Diffusion WFM은 고품질·고충실도 세계 생성에 강하며, Text2World와 Video2World를 모두 지원하는 Cosmos의 핵심 생성 계열

PART 1 정리

5장 전반부 — Diffusion-based WFM

- 범용 세계모델(generalist) 학습: 10,000 H100 · 약 3개월
- Latent diffusion + EDM denoising score matching (uncertainty weighting)
- DiT 기반 denoiser: 3D patchify · FPS-aware 3D RoPE · T5 cross-attn · AdaLN-LoRA(11B→7B)
- Joint image-video + progressive(512p→720p) + multi-aspect 학습
- FSDP · Context Parallelism으로 대규모 학습 가능 · Prompt Upsampler로 분포 gap 완화

Part 2 예고: 5.2 Autoregressive-based WFM (next-token prediction · 실시간 추론 최적화 · Diffusion Decoder) + 5.3 Evaluation (3D Consistency · Physics Alignment)

NVIDIA COSMOS · WORLD FOUNDATION MODEL PLATFORM

Chapter 5 part2

WFM Pre-training: Autoregressive

Discrete video token 기반 next-token prediction WFM · 3D positional embedding · Medusa speculative decoding · Diffusion decoder · 3D consistency · Physics alignment 평가

Cosmos의 WFM pre-training은 대규모 비디오 데이터를 기반으로 물리 세계의 시각적 변화, 움직임, 3D 일관성, 기초 물리 동역학을 학습하는 범용 세계모델 구축 과정이다.

arXiv:2501.03575v3 · §5.2 Autoregressive WFM

Cosmos Autoregressive-based Model Foundation Model

5.1 Autoregressive-based World Foundation Model 핵심정리

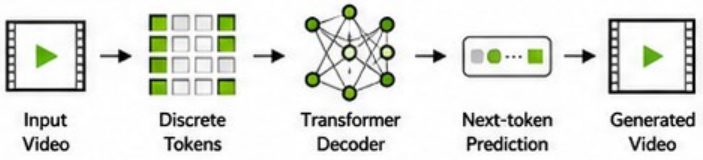
Physical AI를 위한 Cosmos WFM의 Autoregressive 기반 사전학습 구조

1 핵심 메시지

- 사전학습된 Autoregressive WFM은 discrete video token sequence를 다음 토큰 예측으로 생성하는 범용 세계 시뮬레이터
- 비디오 세계를 시공간 token 언어처럼 보고 next-token prediction으로 미래를 rollout
- 4B / 12B base WFM과 5B / 13B Video2World 계열로 구성
- diffusion decoder를 결합해 blur와 artifact를 완화

★ 핵심: Autoregressive WFM은 빠른 rollout과 LLM형 확장성이 강점인 Cosmos의 세계 생성 계열이다.

2 기본 개념과 모델 구성



$$L_{NLL} = \sum_i -\log P(v_i \mid v_1, \ldots, v_{i-1}; \Theta)$$

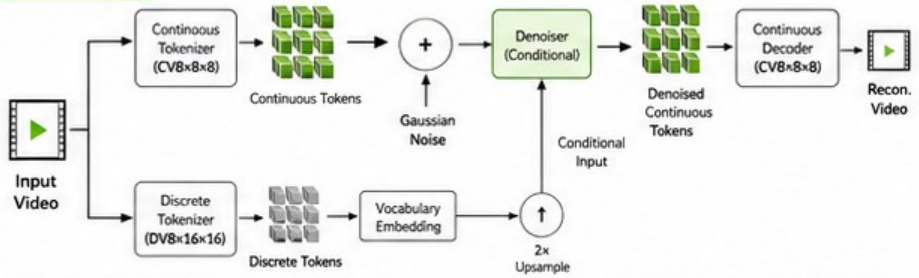
- discrete tokenizer: Cosmos-Tokenize1-DV8×16×16-720p
- text conditioning은 T5-XXL cross-attention으로 결합
- LLM-like decoder transformer 기반

모델 맵

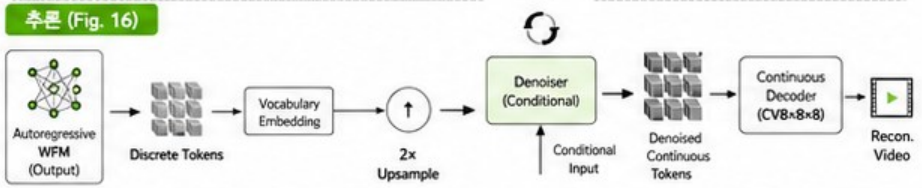
Base WFM	Video2World
• Cosmos-Predict1-4B	• Cosmos-Predict1-5B-Video2World
• Cosmos-Predict1-12B	• Cosmos-Predict1-13B-Video2World

3 Architecture + Diffusion Decoder (논문 Fig. 15–16)

학습 (Fig. 15)



추론 (Fig. 16)



Architecture 요약

- Llama3-style Transformer decoder
- 3D RoPE + 3D APE
- YaRN으로 temporal context length 확장
- QK normalization + z-loss로 안정화
- diffusion decoder가 discrete output을 continuous token으로 복원

Diffusion Decoder: blur 감소 + sharpness 향상

4 학습 전략과 스케일링

Stage 1	첫 frame 조건 → 16개 미래 frame 예측
Stage 1.1	34-frame context 확장 + temporal YaRN
Stage 2	text conditioning 추가, Video2World 확장
Cooling-down	high-quality image-video pairs로 마무리 학습
Scale-up	Tensor Parallelism + Sequence Parallelism
학습 해상도	640×1024 중심

LLM형 병렬화가 autoregressive WFM 학습의 핵심 기반

5 추론 최적화와 성능

- KV caching
- tensor parallelism
- torch.compile
- Medusa speculative decoding
- low-resolution adaptation (320×512)

8× H100 기준
약 10 FPS
실시간 video generation

9개의 Medusa head가 효율-성능 trade-off가 가장 우수

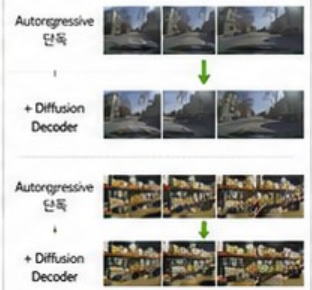
6 결과와 한계 (논문 Fig. 17–19)

A. 정성 결과 (Fig. 17)



큰 모델일수록 sharper detail과 better motion

B. Diffusion Decoder 효과 (Fig. 18)



autoregressive 단독 결과의 blur를 완화하고 sharpness 향상

C. Failure Case (Fig. 19)



일부 객체가 아래에서 갑자기 나타나는 failure case

한계 및 강건성

- image conditioning failure rate
  - 4B: 15%
  - 5B: 7%
  - 12B: 2%
  - 13B: 3%
- 9-frame video conditioning은 모든 모델에서 2% 미만 failure rate
- 작은 모델보다 큰 모델이 더 robust

5.2의 핵심 결론

1. Autoregressive WFM은 next-token 기반 세계 생성 모델이다.
2. LLM형 구조와 병렬화로 확장성과 속도가 뛰어나다.
3. Diffusion decoder가 화질과 선명도를 보완한다.
4. Medusa와 video conditioning이 실사용성을 높인다.

★ 한 줄 요약: Cosmos의 Autoregressive-based WFM은 discrete video token을 빠르게 rollout하며, diffusion decoder와 최적화 기법을 통해 Physical AI용 미래 세계를 효율적으로 생성하는 범용 world simulator이다.

핵심 원리 : World Simulation = Next-Token Prediction

Cosmos Discrete Tokenizer를 사용하여 비디오를 이산 비디오 토큰 시퀀스 $V = \{v_1, v_2, \cdots, v_n\}$으로 변환
그런 다음 LLM과 유사하게, 과거 비디오 토큰을 컨텍스트로 사용하여 다음 비디오 토큰을 예측하도록 Transformer 디코더를 학습함.
학습 목표는 다음의 음의 로그 우도(Negative Log-Likelihood, NLL) 손실을 최소화

Formulation · Negative Log-Likelihood over discrete video tokens

$$\mathcal{L}_{NLL} = \sum_i -\log P\left(v_i | v_1, v_2, \ldots, v_{i-1}; \Theta\right) \quad (9)$$

여기서 $V = \{v_1, v_2, \cdots, v_n\}$는 Cosmos-Tokenize1-DV8×16×16-720p로 얻은 discrete video token sequence, Θ는 Transformer decoder 파라미터.

Input Video
비디오 입력

Discrete Tokenizer (§4)
Discrete video token 시퀀스로 변환

Transformer Decoder
과거 token을 context로 next-token 예측 (LLM과 동일한 방식)

Future Video Tokens
미래 세계 상태를 token으로 rollout

GPT류 LLM(Brown et al., 2020; Dubey et al., 2024 등)과 동일한 학습 패러다임을 discrete video token 위에서 수행 — LLM의 학습·추론 인프라를 그대로 활용할 수 있는 구조적 기반

표준 Transformer에 대한 3가지 수정 (Figure 14)

video generation 과제에 맞춰 ① 3D-aware positional embedding, ② cross-attention(텍스트 입력), ③ QK-Normalization을 추가

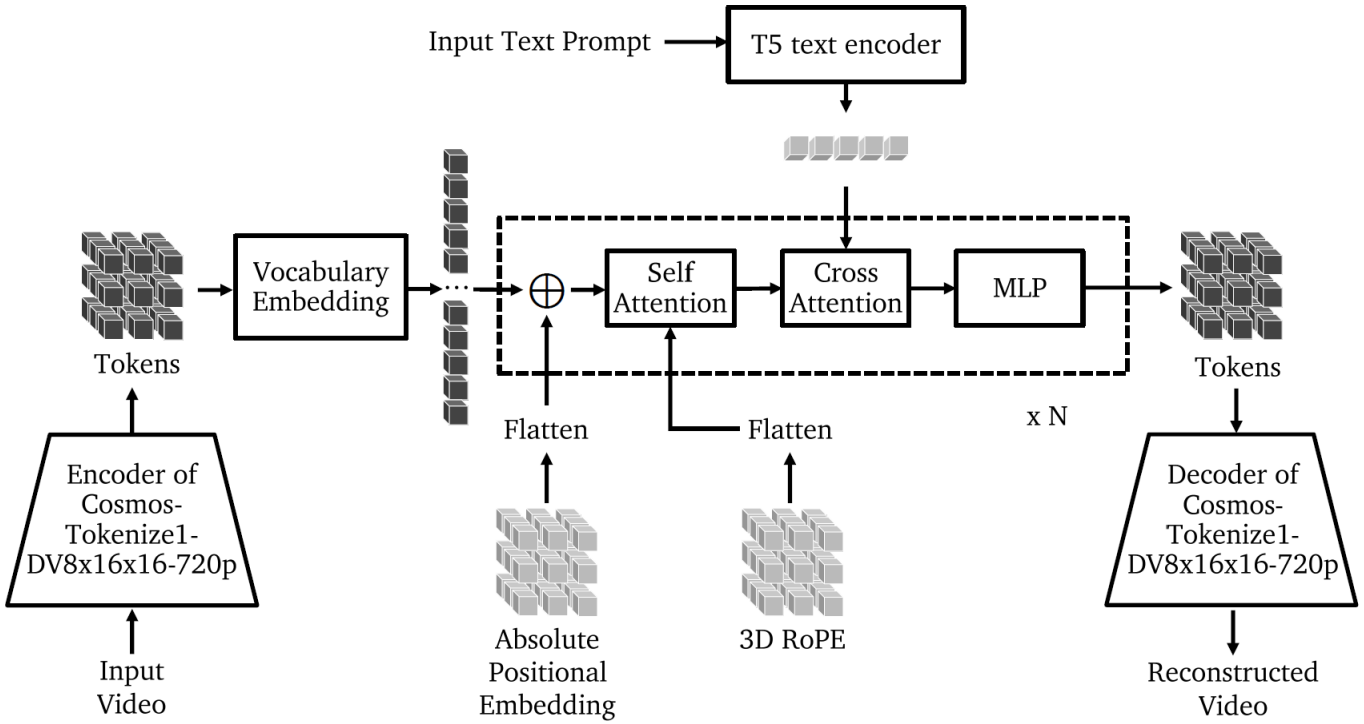


Figure 14. Architecture of Cosmos-Predict1-Video2World Model

설계 요소	역할
3D-aware Positional Embeddings	**3D factorized RoPE**(상대 위치)와 **3D factorized APE**(절대 위치)를 함께 사용. 두 메커니즘이 결합되어 네트워크 전반에 종합적인 시공간 정보를 제공
Cross-Attention (Text Conditioning)	텍스트 입력으로 더 나은 제어를 가능하게 하기 위해 cross-attention 계층을 추가. 모든 self-attention layer 뒤에 cross-attention block을 삽입, pre-trained T5-XXL text encoder의 embedding을 K/V로 사용
QK-Normalization	학습 안정성 향상을 위해 QKNorm(Wortsman et al., 2023) 도입. attention의 Q·K를 dot product 전에 정규화하여 softmax 포화를 방지

- 입력 비디오를 Cosmos-Tokenize1-DV8×16×16-720p의 인코더를 통해 인코딩하여 이산 토큰을 생성하는 것으로 시작하며, 이 토큰들은 학습된 임베딩(vocab embedding)으로 변환됨.
- 이러한 임베딩은 반복적인 트랜스포머 블록을 통해 처리되는데, 각 블록은 절대 위치 임베딩과 3D RoPE 구성 요소로 이루어져 있으며, 이는 Self Attention 모듈에 입력되기 전에 평탄화(flattened)됨.
- 각 블록에는 T5 텍스트 인코더를 통해 처리된 인코딩된 텍스트 프롬프트를 통합하는 크로스 어텐션 모듈이 포함되어 있으며, 그 뒤에는 2계층 MLP가 이어짐.
- 마지막으로, Cosmos-Tokenize1-DV8×16×16-720p의 디코더가 출력 토큰으로부터 비디오를 재구성함.

3D RoPE(+YaRN) & 3D APE — 상대·절대 위치의 결합

temporal · height · width 3축으로 factorize된 두 positional embedding이 결합되어 시공간 정보를 제공

3D Rotary Position Embedding (RoPE)

- **temporal · height · width** 차원에 걸친 **상대적 위치 정보**를 인코딩
- 학습은 multi-stage로 진행되며 stage가 진행될수록 비디오 sequence 길이가 증가
- 변화하는 temporal duration에 RoPE를 적응시키기 위해 **YaRN**(Peng et al., 2023) 사용 — (Yet another RoPE extensioN method: RoPE + LLM으로 10배 적은 토큰과 2.5배 적은 학습 단계로 Context Window를 128k까지 확장)
- video sequence 길이는 temporal 축으로만 증가하므로 **YaRN 확장은 temporal 축**에만 적용
- YaRN을 통해 초기 학습 단계에서 접한 것보다 긴 context 길이로 extrapolation 가능;

3D Absolute Positional Embedding (APE)

- 3D RoPE에 더해, 각 transformer block 내부에 3D APE를 추가하여 상대 위치 인코딩을 보완
- temporal · height · width 축으로 factorize된 **sinusoidal embedding**으로 절대 위치 정보를 인코딩
- 각 stage에서 input tensor에 직접 더해져 positional context를 보강
- 절대(APE) + 상대(RoPE) 결합 시 **모델 성능 향상 · training loss 감소 · 생성 비디오의 morphing artifact 최소화**
- *주의 : Diffusion WFM(§5.1.2)은 learnable embedding을 사용하지만, AR WFM의 APE는 sinusoidal 기반을 채택*

▶ **Table 14 설정 : 3D RoPE (θ = 500,000) + 3D APE**

Vocabulary · Cross-attention · QKNorm · Z-loss

LLM의 BPE tokenizer를 대신하는 FSQ 기반 vocabulary와, 대규모 학습을 지탱하는 안정화 기법

Vocabulary

- LLM의 vocab은 BPE 등으로 학습된 tokenizer(예: tiktoken)가 결정하지만, AR WFM은 **Cosmos-Tokenize1-DV8×16×16-720p**를 tokenizer로 사용
- **FSQ**(Finite-Scalar-Quantization)(Mentzer et al., 2023)로 6차원 latent space를 (8,8,8,5,5,5) level로 양자화
- Vocabulary size = 8×8×8×5×5×5 = **64,000**

Cross-attention for Text Conditioning

- Transformer의 self-attention block에 더해, 입력 텍스트를 조건으로 쓸 수 있도록 cross-attention 계층 추가
- Diffusion WFM(§5.1.2)과 유사하게, transformer feature와 pre-trained text encoder(**T5-XXL**)의 text embedding 사이에서 cross-attention 수행
- 실험에서는 **모든 self-attention layer 뒤에 cross-attention block 추가**

Query-Key Normalization (QKNorm)

- dot product 계산 전에 Q와 K를 정규화 → softmax 포화를 방지하고 더 효과적인 학습을 보장
- 정규화 후 dot product는 고정 상수 $1/\sqrt{d_k}$ 대신 **학습 가능한 파라미터 γ로 스케일링**
- γ가 attention score의 크기를 적응적으로 제어 → 유연성과 표현력 향상

Z-loss

- logit이 0에서 벗어나는 것을 penalize하는 안정화 항 : $\mathcal{L}_{\text{z-loss}} = \lambda \cdot \sum_i z_i^2$
- 과도한 logit 값에 의한 numerical instability·gradient explosion 억제
- 다수의 GPU 노드로 학습을 확장할 때 gradient norm을 healthy range로 유지하는 데 결정적
- 경험적으로 계수 $\lambda = 3\times10^{-4}$가 성능 저하 없이 학습을 안정화하는 최적 균형점

메모리 분석과 병렬화 전략 (TP · SP)

학습 시 GPU 메모리 소비 요소를 분석하고, 12B 모델을 다중 GPU로 분산하는 방법

Memory Requirements (학습 시)

Model parameters	6 bytes/param	BF16 + FP32 모두 저장
Gradients	2 bytes/param	BF16 저장
Optimizer states	8 bytes/param	AdamW 1st/2nd moment, FP32
Activations	≈ 2 × L × 17 × seq × bs × d_model bytes	Korthikanti et al., 2023 참조

Tensor Parallelism (TP) — Shoeybi et al., 2019

- linear layer의 weight를 input 또는 output feature 차원으로 분할 — GPU 간 통신 최소화가 분할 기준
- 예 : 2-layer FFN에서 1층은 output 축, 2층은 input 축으로 분할 → 중간 activation을 GPU 간 통신 없이 로컬 처리
- 최종 출력은 all-reduce로 결합 · 각 GPU는 linear layer weight의 1/TP_SIZE만 저장
- 단, 기본 TP는 LayerNorm 등에서 sequence 축 activation이 중복됨

12B 모델 (Cosmos-Predict1-12B)

parameters + gradients + optimizer states ≈ 192 GB
→ H100 1장의 80GB HBM3 용량 초과 → TP·SP로 메모리·연산을 다중 GPU에 분산

Sequence Parallelism (SP) — Korthikanti et al., 2023

- TP를 확장하여 context를 sequence 차원으로도 분할
- LayerNorm · Dropout처럼 sequence 원소별로 독립 처리 가능한 연산에 적용
- SP 활성화 시 각 GPU는 activation의 1/TP_SIZE만 저장

Multi-stage Pre-training + Cooling-down

모든 모델은 고정 공간 해상도 640 × 1024로 학습

Stage 1	Stage 1.1	Stage 2	Cooling-down
Video Prediction	**Longer Context**	**Text Conditioning**	**High-quality Data**
context 17 frames	**context 34 frames**	**context 34 frames + text**	**30,000 iterations**
video prediction objective로 학습. 첫 frame을 입력 조건으로 주고 미래 video frame들을 예측 (1 frame 입력 → 16 frames 예측)	동일한 video prediction이지만 context length를 34 frames로 확대. temporal 축 YaRN extension으로 RoPE context length 증가	새로 초기화된 cross-attention으로 text embedding 결합. text-to-video 능력 향상 위해 §5.1.3의 joint image-video 학습 사용. image batch는 context가 짧아 더 큰 batch size 사용	pre-training 후 LLM 관행(Dubey et al., 2024)과 유사한 cooling-down 수행. 고품질 image-video pair로 학습하며 learning rate를 0까지 선형 감소

4개 모델 · 2계열

Cosmos-Predict1-4B / 12B = Stage 1 + 1.1로 학습된 pure next-video-token predictor (text prompt 미사용)
Cosmos-Predict1-5B / 13B-Video2World = 각 base 모델에 cross-attention 계층을 추가해 Stage 2까지 학습한 파생 모델

Cosmos-Predict1 모델 구성 상세 (Table 14)

4B·12B(base)와 5B·13B(Video2World)의 아키텍처·최적화 하이퍼파라미터

Table 14: Configuration details of Cosmos-Predict1 models.

Configuration	4B	5B-Video2World	12B	13B-Video2World
Number of Layers	16	16	40	40
Model Dimension	4,096	4,096	5,120	5,120
Cross Attention Layers	✗	✓	✗	✓
Base Learning Rate	1×10^{-3}	3×10^{-4}	1×10^{-3}	5×10^{-4}
Weight Decay	0.01			
Learning Rate Warmup	Linear scheduler with 5,000 iterations			
Activation Function	SwiGLU			
FFN Hidden Dimension	14,336			
Number of Attention Heads	32			
Number of Key / Value Heads	8			
Number of Tokens	12,800			
Vocabulary Size	64,000			
Positional Embedding	3D RoPE ($\theta = 500{,}000$) + 3D APE			

- **Cosmos-Predict1-4B**: 다음 비디오 토큰 예측을 위한 4B 트랜스포머 모델임. 이 모델은 다단계 학습 목표의 1단계와 1.1단계를 사용하여 학습됨
- **Cosmos-Predict1-5B-Video2World**: Cosmos-Predict1-4B에서 파생되어 다단계 학습 목표의 2단계로 추가 학습된 5B 트랜스포머 모델임
- **Cosmos-Predict1-12B**: 다음 비디오 토큰 예측을 위한 12B 트랜스포머 모델임. 이 모델은 다단계 학습 목표의 1단계와 1.1단계를 사용하여 학습됨
- **Cosmos-Predict1-13B-Video2World**: Cosmos-Predict1-12B에서 파생되어 다단계 학습 목표의 2단계로 추가 학습된 13B 트랜스포머 모델임

Towards Real-Time Generation — LLM 기법 + Medusa

LLM과의 아키텍처 유사성 덕분에, 확립된 LLM 추론 최적화 기법으로 sequential decoding 병목을 해소

기본 최적화 조합 : gpt-fast (PyTorch) 구현을 따름

- **Key-Value Caching** : 이전 step의 K·V 재사용
- **Tensor Parallelism** : weight를 다중 GPU에 분산
- **torch.compile** : graph compilation 가속

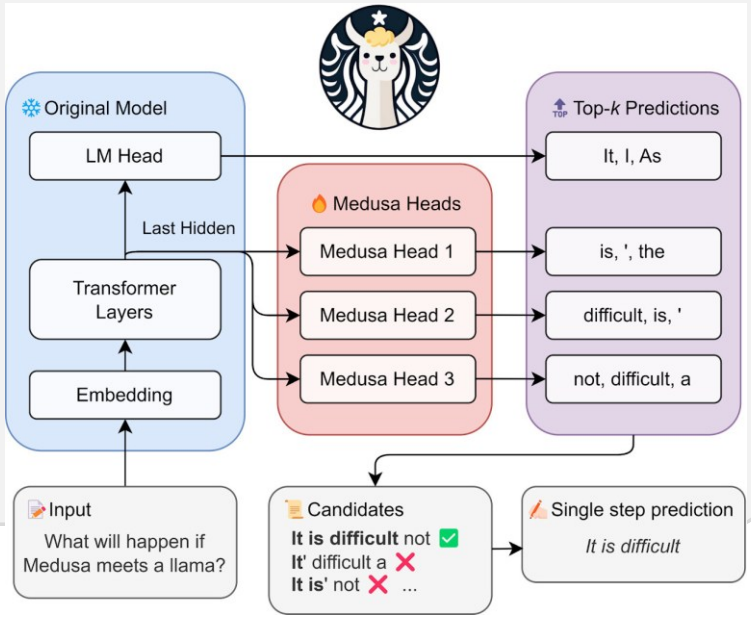
Medusa Speculative Decoding (Cai et al., 2024)
별도의 Draft Model 없이, LLM 생성 속도를 높이는 추론 가속화 프레임워크

동작원리

- 별도 draft model이 필요한 일반적 speculative decoding과 달리, **transformer backbone에 추가 decoding head를 확장하여 여러 후속 token을 병렬 예측**
- 예측된 token들을 rejection sampling으로 검증 → one-token-at-a-time 병목 완화
- 생성 품질 저하 없이 visual autoregressive 가속 가능성을 입증
- tree-based attention 메커니즘(Cai et al., 2024)은 사용하지 않음

구현 · fine-tuning 전략

- Medusa head는 마지막 transformer hidden state 뒤에 삽입 · backbone 파라미터와 final unembedding layer는 head 간 공유
- 각 head = SiLU activation + residual connection의 single-layer FFN · 여러 head의 weight matrix를 하나의 통합 FFN으로 병합해 병렬성 극대화
- head만 fine-tune → multi-token 예측 성능 저조 · full fine-tuning → 품질 저하 (catastrophic forgetting)
- **마지막 2개 transformer layer + final unembedding layer만 unfreeze** (backbone frozen)가 최적



Medusa head 수에 따른 throughput / forward pass (Table 15)

8 × H100 GPU · 640 × 1024 해상도의 unseen test video 50개로 평가

Table 15. Impact of Medusa head number on average token throughput and number of forward passes.
The experiments are conducted on 8 × H100 GPUs and 50 unseen test videos of 640 × 1024 resolution.

Model	Medusa Head Number	0	3	6	9	12
4B	Token Throughput (tokens/s)	444.95	663.51	829.59	**894.67**	890.64
	# of Forward Passes	7680	2860	2073	1812	**1682**
5B	Token Throughput (tokens/s)	303.61	659.94	758.58	**982.77**	978.80
	# of Forward Passes	10240	2857	2382	1799	**1673**

논문의 관찰 (paper findings)

- 4B 모델 : **최대 2.0× token throughput · 4.6× 적은 forward pass**
- 5B 모델 : **최대 3.2× token throughput · 6.1× 적은 forward pass**
- head 수를 늘리면 생성에 필요한 forward pass 수는 줄지만, 전체 token throughput은 오히려 느려질 수 있음
- **9개의 Medusa head**가 계산 효율과 모델 성능 간 최적의 trade-off

Medusa 결합 AR WFM 추론 성능 (640 × 1024) (Table 16)

32 frames 생성, 단일 conditioning frame 입력 기준 · BF16 · DD = diffusion decoder

Table 16: Performance analysis of Cosmos Autoregressive Models on test videos of 640 × 1024 resolution.

Model	GPU	No DD (s)	No DD+Medusa (s)	With DD (s)	With DD+Medusa (s)	VRAM (GB)
4B	1	31.04	23.52	61.49	53.08	29
	4	18.20	13.87	29.60	25.63	31
	8	17.62	9.91	30.30	22.83	34
5B	1	39.68	24.97	70.39	54.72	59
	4	25.59	20.96	37.29	33.35	51
	8	25.70	11.67	38.41	24.35	49
12B	1	84.78	—	116.66	—	45
	4	47.49	—	60.27	—	36
	8	45.69	—	58.81	—	37
13B	1	109.18	—	140.24	—	77
	4	67.80	—	80.76	—	55
	8	67.22	—	80.93	—	55

- BF16의 “max-autotune” 모드에서 이 표는 다양한 Cosmos Autoregressive WFM의 평균 추론 시간(초 단위)과 VRAM 사용량을 보고함
- 추론 시간은 단일 조건부 프레임을 입력으로 사용하여 32개의 프레임을 생성하는 기준으로 보고됨

- **No DD**: diffusion decode가 없는 경우의 시간
- **No DD+Medusa**: diffusion decode는 없지만 Medusa head가 있는 경우의 시간
- **With DD**: diffusion decode가 있는 경우의 시간
- **With DD+Medusa**: diffusion decoder와 Medusa head가 모두 있는 경우의 시간
- **VRAM**: 기가바이트(GB) 단위의 비디오 RAM 사용량

▶ **Medusa는 4B·5B 모델에서 GPU 수와 무관하게 일관된 추론 가속을 제공 · H100 GPU에서 평가 (12B·13B는 Medusa 미적용 수치만 보고)**

저해상도 적응을 통한 Real-Time 추론 달성

320 × 512 저해상도로 적응해 비디오당 token 수를 줄이고 실시간 생성 추구

STEP 1

Discrete tokenizer 저해상도 fine-tune

Cosmos-Tokenize1-DV8×16×16-720p를 target Physical AI domain의 320p 저해상도 비디오로 fine-tune

→

STEP 2

AR WFM 저해상도 fine-tune

640×1024로 pre-train된 Cosmos-Predict1-4B(§5.2.3)를 이 저해상도 tokenizer와 함께 target domain의 320×512 비디오로 fine-tune

→

STEP 3

Medusa head 추가

fine-tune된 저해상도 AR WFM에 Medusa head를 추가해 speculative decoding 적용

Table 17. 벤치마크 : 8 × H100 80GB · torch.compile “max-autotune” · BF16 · Physical AI domain의 10-FPS, 320 × 512 비디오

Model (320 × 512)	Token Throughput (tokens/s)	Video Throughput (frames/s)
Cosmos-Predict1-4B (with Medusa)	806.61	10.08

▶ **10개의 video frame을 1초 이내에 생성 — 10 FPS 실시간 비디오 생성 달성**

공격적 압축의 한계를 보완하는 Diffusion Decoder (Figure 15)

Cosmos-Predict1-7B-Text2World(§5.1)를 fine-tune하여 더 강력한 tokenizer decoder를 구축

Cosmos tokenizer는 경량 encoder-decoder로 공격적 압축을 수행해 WFM 학습 token 수를 줄인다. 그 결과, 특히 소수의 정수만으로 풍부한 비디오를 표현하는 discrete tokenization의 AR WFM 설정에서 생성 비디오에 흐림(blurriness)과 가시적 artifact가 발생할 수 있다.
→ OpenAI(2024) · Ramesh et al.(2022)의 diffusion decoder 설계를 채택해 이 한계를 해결

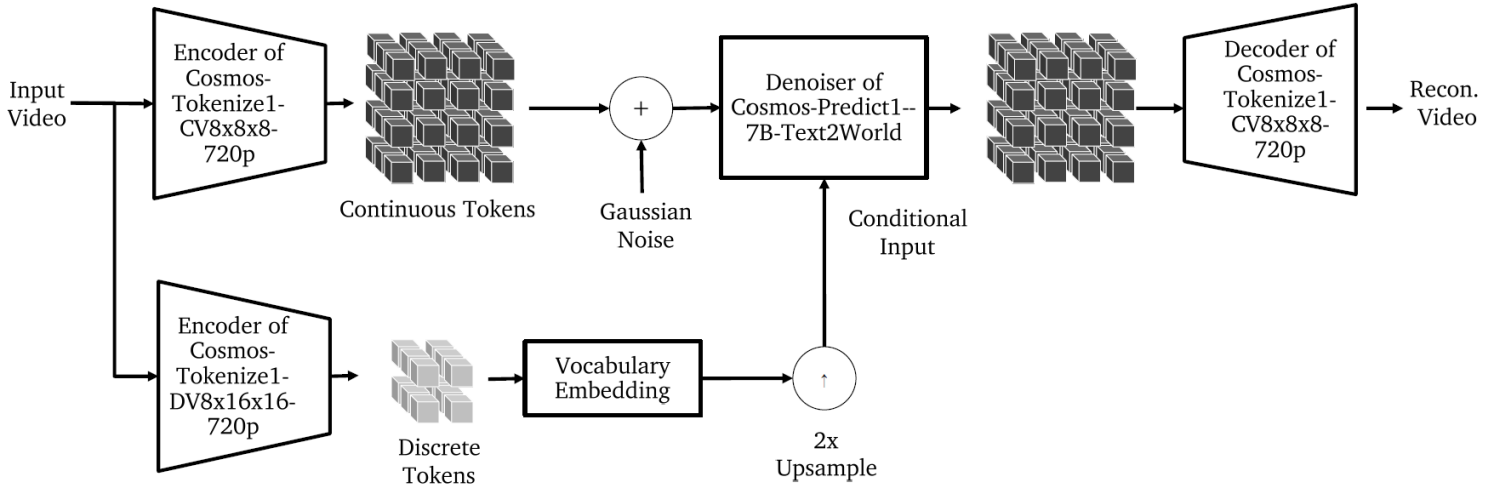
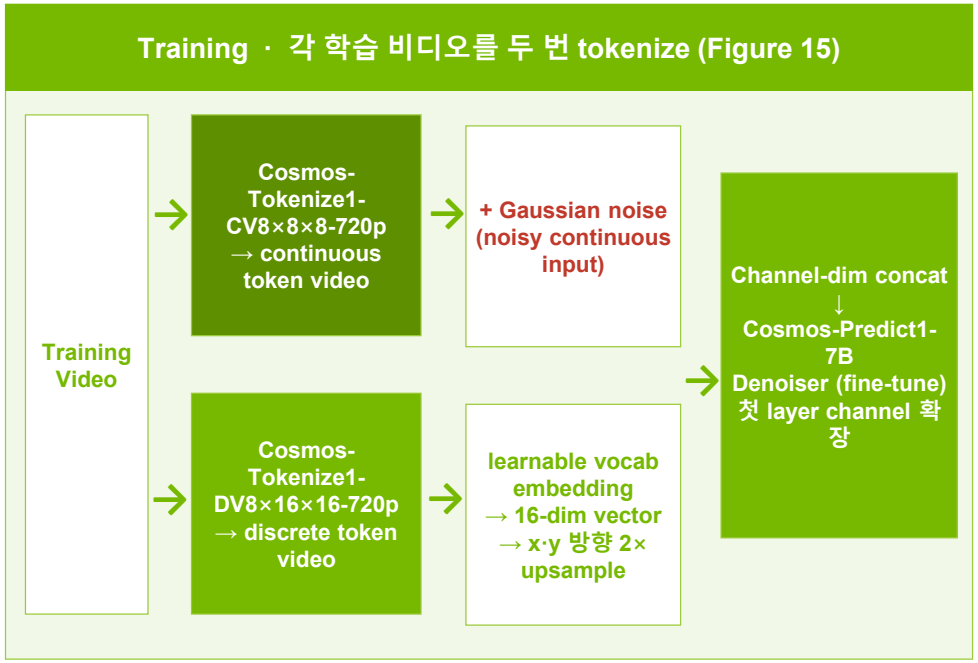


Figure 15. **Cosmos diffusion decoder training**
discrete token video는 noise 없이 conditional input으로 주어지므로, denoiser는 조건에 담긴 정보를 활용해 노이즈를 제거하도록 학습 → discrete token을 reverse diffusion 문제 풀이로 decode하는 고품질 decoder 획득



AR WFM 출력의 2단계 디코딩 (Figure 16)

AR WFM이 출력한 discrete token video(8×16×16 압축)를 두 단계로 RGB 비디오로 변환

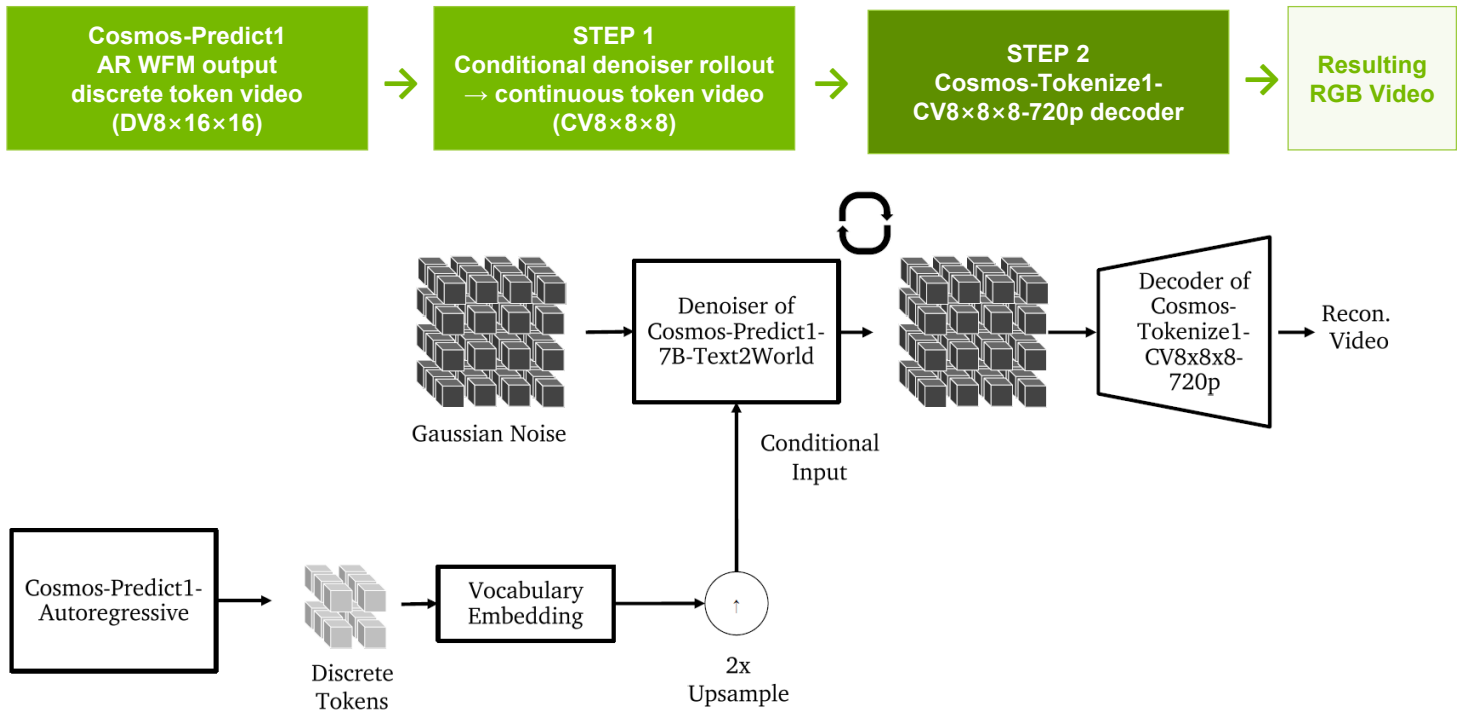


Figure 16. **Cosmos diffusion decoder inference**
During inference, the output video token from a Cosmos-Predict1 model is conditionally inputted to the denoiser.

설계 근거 (논문 기술 내용)

- **CV8×8×8-720p**는 DV8×16×16-720p보다 **더 부드러운(gentle) continuous tokenization과 덜 공격적인 압축 방법** (8×8×8 vs 8×16×16) 덕분에 더 높은 품질의 비디오 출력을 생성 가능
- conditional input 계산 : discrete token → learnable vocabulary embedding layer로 16차원 벡터 embed → x·y 방향 2× upsample → denoiser의 noisy input과 같은 크기로 정렬 → channel 차원 concat
- denoiser의 첫 layer는 새 입력 shape을 수용하도록 channel 차원이 확장됨 · 추가된 noise를 제거하도록 업데이트된 Cosmos-Predict1-7B를 fine-tune

모델 크기·Diffusion Decoder·Prompt Upsampler에 대한 정성적 결과

Figure 17 (모델 크기 비교)

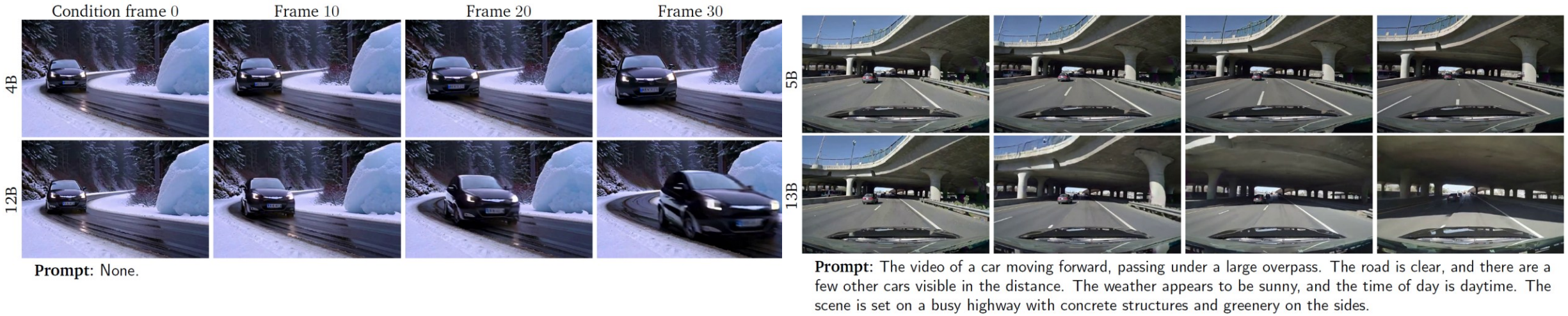


Figure 17: **Generated videos from Cosmos Autoregressive World Foundation Models**

상단 두 행은 4B 및 12B 모델을 사용한 비디오 생성 결과이며, 하단 두 행은 텍스트 프롬프트를 사용한 Video2World 결과임. 12B 및 13B 모델이 프롬프트가 있는 설정과 없는 설정 모두에서 4B 및 5B 모델보다 더 선명한 비디오와 더 나은 움직임을 보여줌

모델 크기 비교 (Figure 17)

- ➢ Unprompted setting : 4B vs 12B
  - 12B 모델이 더 나은 motion과 더 선명한(sharper) detail의 비디오를 생성
- ➢ Prompted setting : 5B-Video2World vs 13B-Video2World
  - 13B 모델이 5B 모델보다 더 나은 motion을 획득

모델 크기·Diffusion Decoder·Prompt Upsampler에 대한 정성적 결과

Figure 18 (diffusion decoder 효과)



Figure 18: **Diffusion decoder comparison**

- 상단 패널에서는 Cosmos-Predict1-13B-Video2World 모델을 사용한 비디오 생성 결과를 보여줌.
- 하단 패널에서는 autoregressive 모델의 출력을 diffusion decoder에 통과시킨 후 향상된 비디오를 보여줌
- autoregressive 모델만 사용할 경우 결과가 흐릿하게 나타나는 반면, diffusion decoder는 콘텐츠를 보존하면서 비디오의 선명도를 향상시킬 수 있음을 확인함

Diffusion Decoder 효과 (Figure 18)

- AR 모델 단독 출력은 discrete tokenizer의 lossy compression으로 인해 주로 흐릿(blurry)함
- Diffusion decoder를 통과하면 content를 보존하면서 detail·sharpness를 강화
- Cosmos-Predict1-13B-Video2World 출력에 대한 비교로 시연

Failure Case 분석 : 아래에서 갑자기 나타나는 객체 (Figure 19 · Table 18)

100개 Physical AI 입력으로 구성한 평가셋에서 모든 모델의 생성 비디오를 수동 검수하여 실패율 측정

관찰된 failure case & 평가 설계

- 대표적 실패 사례 : 생성 비디오에서 일부 객체가 **아래쪽에서 예기치 않게 나타남** (Figure 19, Cosmos-Predict1-4B 출력 예시)
- 평가셋 : AR WFM 입력용 **100개 Physical AI 입력**
- 두 가지 입력 모드로 모든 모델에서 비디오 생성 :
  ① **image (single-frame) conditioning**
  ② **video (9-frame) conditioning**
- 생성된 모든 비디오를 수동 검수(manual inspection)하여 failure rate 보고

Failure Rate Analysis

Table 18: **Failure rate analysis of Cosmos Autoregressive models**

Model	Image Conditioning	Video Conditioning (9 frames)
Cosmos-Predict1-4B	15%	1%
Cosmos-Predict1-5B-Video2World	7%	2%
Cosmos-Predict1-12B	2%	1%
Cosmos-Predict1-13B-Video2World	3%	0%

논문의 결론 :

- 작은 모델(4B·5B-Video2World)은 single-frame conditioning에서 더 높은 corruption rate를 보임
- 큰 모델(12B·13B-Video2World)은 더 robust
- 9-frame video conditioning에서는 모든 모델이 안정적 — failure rate **2% 미만**

PART 2 정리

5장 중반부 — Autoregressive WFM

5.2 Formulation	world simulation을 next-token prediction으로 정식화 · discrete video token 위 NLL loss로 Transformer decoder 학습 (수식 9)
5.2.1 Architecture	3D RoPE(+YaRN, temporal만) + 3D APE(sinusoidal) · T5-XXL cross-attention(모든 self-attn 뒤) · QKNorm(학습 가능 γ) · Z-loss($\lambda=3\times10^{-4}$) · FSQ vocab 64,000
5.2.2 Scaling Up	12B ≈ 192 GB > H100 80 GB → TP + SP로 분산 · MQA/GQA 미사용(LLM 정렬 우선) · CP/PP는 future work
5.2.3 Training	Stage 1(17f) → Stage 1.1(34f, YaRN) → Stage 2(text, joint image-video) → cooling-down 30K iter · 해상도 640×1024 고정 · 4B/12B base → 5B/13B-Video2World
5.2.4 Inference	KV cache + TP + torch.compile(gpt-fast) + Medusa 9-head · 4B 저해상도(320×512) 적응 → 806.61 tokens/s · 10.08 frames/s = 10 FPS 실시간
5.2.5 Diff. Decoder	비디오를 CV8×8×8·DV8×16×16로 이중 tokenize → discrete 조건 concat으로 7B 모델 fine-tune → AR 출력의 blur/artifact 보정
5.2.6-7 결과·한계	12B/13B가 motion·detail 우수 · prompt upsampler 효과 없음 (순수 video 사전학습 가설) · 'objects from below' 실패, 9-frame 조건 시 실패율 ≤2%

NVIDIA COSMOS · WORLD FOUNDATION MODEL PLATFORM

Chapter 5 part3

WFM Pre-training: Evaluation

5.3.1 3D Consistency — 생성 비디오의 기하학적 3D 일관성 평가
5.3.2 Physics Alignment — 렌더링된 동역학의 물리 법칙 준수 평가

Cosmos의 WFM pre-training은 대규모 비디오 데이터를 기반으로 물리 세계의 시각적 변화, 움직임, 3D 일관성, 기초 물리 동역학을 학습하는 범용 세계모델 구축 과정이다.

Pre-trained WFM 평가의 두 축

Pre-trained WFM은 visual world simulation의 generalist — 그 능력은 여러 측면에서 측정되어야 함

Axis 1 · 3D Consistency (§5.3.1)

이상적인 WFM은 기하학적으로 타당한(geometrically plausible) 3D 세계로부터 비디오 시뮬레이션을 생성해야 한다.

→ 생성된 비디오의 **3D 일관성**을 평가

- Geometric consistency (epipolar geometry 기반)
- View synthesis consistency (novel view 합성 기반)

Axis 2 · Physics Alignment (§5.3.2)

렌더링된 동역학(dynamics)이 물리 법칙을 얼마나 잘 준수하는지 계산한다.

→ 생성된 비디오의 **physics alignment**를 평가

- 물리 시뮬레이션 엔진 기반 controlled 벤치마크
- Pixel-level · Feature-level · Object-level 메트릭

저자 노트 · *"Evaluation of WFMs is a highly nontrivial task."*
평가에 필요한 다른 중요한 측면들이 여러 개 있음을 인정하며, 더 포괄적인 평가는 future work로 남긴다.

평가 설계 : Test Data & Baseline

WFM은 video 생성을 통해 3D 세계를 시뮬레이션하도록 설계됨 — 생성 비디오가 시각 세계의 3D 구조와 얼마나 일관되는지 평가

생성 비디오는 사실적으로 보이는 것에 더해, 시간에 걸쳐 장면의 물리적 원리와 일관성(coherence)을 유지해야 함 — downstream Physical AI 응용의 핵심 요구사항

Test Data

- **static scene 시나리오에 집중** — multi-view geometry 기반 기존 도구로 3D 일관성을 효과적으로 측정하기 위함
- **RealEstate10K**(Zhou et al., 2018) test set에서 무작위로 선택한 **500개 비디오**로 데이터셋 구성
- proprietary VLM으로 비디오를 추가 캡셔닝하여 **static scene으로 기술하는 text prompt** 획득 → metric 계산 시 scene motion을 고려할 필요가 없도록 함

Baseline & 평가 대상

- Baseline : **VideoLDM** (Blattmann et al., 2023)과 비교
- 평가 대상 Cosmos 모델 (Table 19) :
  Cosmos-Predict1-7B-Text2World
  Cosmos-Predict1-7B-Video2World
  Cosmos-Predict1-4B
  Cosmos-Predict1-5B-Video2World
- 참조(reference)로 실제 비디오(Real Videos)의 수치도 함께 보고

Geometric Consistency — Sampson Error & Pose Estimation

생성 비디오는 3D 시각 세계의 2D 투영 — epipolar geometry 제약의 충족 정도로 3D 일관성을 정량화

Sampson Error

$$\epsilon_{\text{samp}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left(\bar{\mathbf{x}}_i^\top \mathbf{F}\mathbf{x}_i\right)^2}{(\mathbf{S}\mathbf{F}\mathbf{x}_i)^2 + \left(\bar{\mathbf{x}}_i^\top \mathbf{S}\mathbf{F}\right)^2}, \text{where } \mathbf{S} = \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\end{bmatrix} \quad (10)$$

- 한 관심점(interest point)에서 다른 view의 대응 epipolar line까지 거리의 1차 근사(first-order approximation)
- frame pair의 N개 point correspondence $\{(\bar{\mathbf{x}}_i, \bar{\mathbf{y}}_i)\}_{i=1}^{N}$(homogeneous 좌표)로부터 fundamental matrix $\mathbf{F}$를 추정하여 계산
- metric이 pixel 단위로 직관적이도록 **square-root 버전의 error function** 사용

구현 상세

- keypoint 검출·매칭 : **SuperPoint + LightGlue**
- $\mathbf{F}$ 추정 : **OpenCV 8-point RANSAC** 알고리즘
- 평균 오차를 **960 × 540 canvas 기준 frame 대각선 길이**로 정규화

Camera Pose Estimation Success Rate

- 생성 비디오에 대해 camera pose estimation 알고리즘(Schönberger et al., 2016; Schönberger & Frahm, 2016)이 성공하는 비율
- epipolar geometry 제약이 충족되어야 pose 추정이 성공 → 전반적 품질과 3D 일관성을 함께 반영하는 지표

5.3.1 · METRICS ②

View Synthesis Consistency — 3D Gaussian Splatting 기반

interpolated novel viewpoint에서 underlying 3D 구조와 일관된 이미지를 합성하는 능력을 평가

STEP 1

Hold-out 분리

novel view synthesis 문헌의 관행(Mildenhall et al., 2020)에 따라 매 8 frames를 test frame으로 hold-out

→

STEP 2

3DGS 학습

나머지 training frame으로 3D Gaussian Splatting 모델(Kerbl et al., 2023)을 fit — Nerfstudio 라이브러리(Tancik et al., 2023) 기본 설정 사용

→

STEP 3

Test view 합성·비교

합성된 test view의 품질을 PSNR · SSIM · LPIPS(Zhang et al., 2018)로 정량화

평가 지표

PSNR ↑

Peak Signal-to-Noise Ratio — 픽셀 수준 재구성 품질

SSIM ↑

Structural Similarity — 구조적 유사성

LPIPS ↓

Learned Perceptual Image Patch Similarity — perceptual 유사성

Base Cosmos 모델의 3D Consistency 평가 (Table 19)

Geometric Consistency(좌) · View Synthesis Consistency(우) — baseline 대비 큰 폭의 우위

Table 19: Evaluation of 3D consistency on base Cosmos models.

	Geometric Consistency		View Synthesis Consistency		
Method	Sampson error ↓	Pose estimation success rate (%) ↑	PSNR ↑	SSIM ↑	LPIPS ↓
VideoLDM (Blattmann et al., 2023)	0.841	4.4%	26.23	0.783	0.135
Cosmos-Predict1-7B-Text2World	**0.355**	62.6%	**33.02**	**0.939**	**0.070**
Cosmos-Predict1-7B-Video2World	0.473	**68.4%**	30.66	0.929	0.085
Cosmos-Predict1-4B	0.433	35.6%	32.56	0.933	0.090
Cosmos-Predict1-5B-Video2World	0.392	27.0%	32.18	0.931	0.090
Real Videos (Reference)	0.431	56.4%	35.38	0.962	0.054

Results

- Cosmos WFM은 geometric·view synthesis consistency 모두에서 baseline보다 유의미하게 우수한 3D 일관성 달성
- interest point가 더 3D-consistent할 뿐 아니라, camera pose estimation success rate도 눈에 띄게 높음 — **실제 비디오(real-world videos) 수준에 도달**
- pose 추정에 성공한 케이스에서, 합성된 held-out view는 모든 이미지 합성 지표에서 더 높은 품질 → **Cosmos WFM이 3D 일관된 비디오를 생성하는 효과적인 world simulator임을 입증**

동기 : 물리 법칙에 대한 이해를 측정하다

이상적인 WFM은 물리 법칙을 강하게 이해하고, 이를 준수하는 미래 관측을 생성해야 한다

현재 상태에 대한 저자들의 진단

- Pre-trained WFM은 일정 수준의 물리 이해(physics understanding)를 보이며 state-of-the-art를 전진시켰지만, **물리 법칙을 위반하는 예시를 여전히 쉽게 생성할 수 있음**
- 물리적으로 implausible한 비디오를 제거하는 데이터 큐레이션 추가 단계와 개선된 모델 설계가 필요하다고 판단
- 강력한 physics-aligned WFM 구축은 future work로 남기되, **대규모 데이터 기반 pre-training에서 intuitive physics가 얼마나 자연스럽게 발현되는지(emerge) 측정**하는 데 관심

접근 방법 : 물리 시뮬레이션 기반 controlled 벤치마크

- Kang et al.(2024)에서 영감을 받아, 물리 시뮬레이션 엔진으로 controlled 벤치마크 데이터셋을 설계
- 관심 물리 법칙별 테스트 시나리오에 대해 물리적으로 정확한 photorealistic 비디오를 시뮬레이션으로 생성 — pre-trained WFM의 **Newtonian physics** · **rigid body dynamics** 준수를 테스트
- 이 reference “ground truth” 비디오를, 공유된 context(과거 관측 + perturbation)를 받은 WFM이 생성한 “predicted” 비디오와 비교

PhysX + Isaac Sim 기반 8개 3D 시나리오

서로 다른 물리 효과(physical effects)를 평가하도록 설계된 8개 시나리오

1 Free-falling object(s)	2 Tilted planar slope	3 U-shaped slope	4 Stable stack
평면에 낙하하는 물체 *gravity, collision 등*	**경사면을 굴러 내려가는 물체** *gravity, moment of inertia 등*	**U자형 슬로프를 굴러가는 물체** *potential·kinetic energy 등*	**평형 상태의 물체 스택** *balanced forces*

5 Unstable stack	6 Dominoes	7 Seesaw	8 Gyroscope
불균형 상태의 물체 스택 *gravity, collision 등*	**연쇄적으로 넘어지는 직육면체 블록** *transfer of momentum, collision 등*	**시소 양쪽의 물체** *torque, rotational inertia 등*	**평면 위에서 회전하는 팽이** *angular momentum, precession 등*

데이터 생성 상세

- 시나리오마다 동적 객체의 수·종류(크기·텍스처·형태 변화)를 **Omniverse assets**에서 선택해 randomize · 배경 외관(background appearance)도 randomize
- 시간에 따른 객체의 kinematic state를 시뮬레이션하고, **4개의 static camera view**에서 출력 비디오를 렌더링
- 총 **800개의 1080p 비디오, 각 100 frames** 렌더링
- 존재 모호성(existence ambiguity)을 피하기 위해, 각 시뮬레이션 roll-out의 객체는 첫 frame부터 모두 보이도록 배치

Conditioning 설계와 3단계 평가 지표

simulated GT 비디오와 WFM 예측 비디오를 직접 비교 (Figure 20 : tilted slope · U-shaped slope · unstable stack 예시)

Conditioning

- 미래 관측 생성을 위해 GT 비디오의 **첫 1 또는 9 frames**를 condition으로 WFM에 입력
- 해당되는 경우, 시뮬레이션된 객체의 kinematic state에 초점을 둔 text prompt(conditioning frame들을 proprietary VLM으로 캡셔닝)도 추가 condition으로 사용

① Pixel-level · PSNR / SSIM

WFM rollout의 predicted frame과 GT 비디오의 reference frame을 pixel 수준에서 비교

② Feature-level · DreamSim

약간 더 높은 의미(semantic) 수준의 비교를 위해, predicted-reference frame 간 DreamSim 유사도 점수(Fu et al., 2023) 계산 — feature 유사도 metric

③ Object-level · Avg. IoU (SAMURAI tracking)

관심 객체가 물리 현상에 어떻게 영향을 받는지가 핵심이므로, tracking으로 confounder(배경 변화·시각 품질 등)를 제거한 object-level metric 계산. 합성 테스트 조건이므로 GT instance segmentation mask 접근 가능 — SAMURAI(Yang et al., 2024)로 첫 frame의 GT instance mask를 예측 비디오 나머지 frame에 전파해 track 추출, frame·객체별 GT-predicted mask 간 IoU 계산

평균화 : 비디오 내 frame 전체 · 평가셋 내 비디오 전체 · rollout의 4개 random seed에 걸쳐 평균 — AR 변형이 지원하는 최대 길이인 33 frames에 대해 계산, PSNR·SSIM은 conditioning frame 제외 전체 frame에서 계산

Conditioning 설계와 3단계 평가 지표

simulated GT 비디오와 WFM 예측 비디오를 직접 비교 (Figure 20 : tilted slope · U-shaped slope · unstable stack 예시)

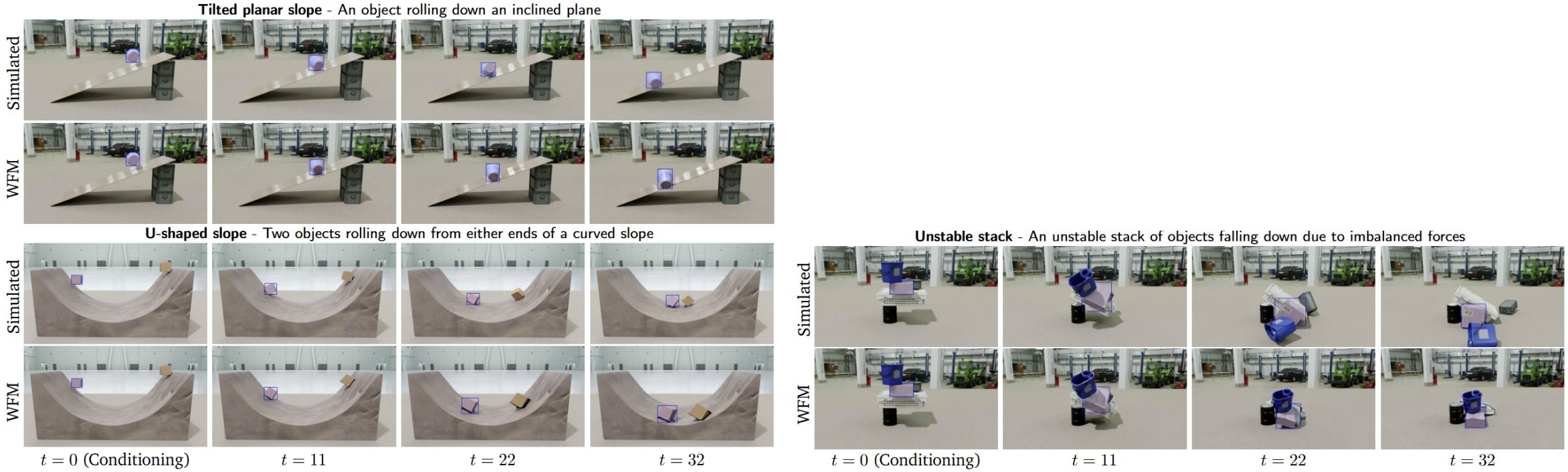


Figure 20: **Physics-scenario rollouts in simulation vs. pre-trained WFM**. We demonstrate three exemplar scenarios of increasing complexity as obtained from the reference (physically correct) simulation (first row in each group) and Cosmos-Predict1-7B-Video2World rollouts (second row in each group). We condition the WFM on 9 frames and a prompt focusing on the kinematic state of the simulated objects. We show one tracked object (blue bounding box and mask) per example used to compute our object-level metrics (average IOU)

Physics Alignment 결과 (Table 20)

pixel-level · **feature-level** · **object-level** 메트릭으로 Cosmos WFM 변형들을 비교 · 33 frames 기준

Table 20: **Physics alignment results**. We compare different variants of Cosmos WFMs in terms of accurate future prediction of a physical scenario using the pixel-level, feature-level, and object-level metrics. Metrics are calculated over 33 frames, the maximum length supported by the autoregressive variants of the Cosmos WFMs

		Pixel-level		Feature-level	Object-level
Model	Condition(s)	PSNR ↑	SSIM ↑	DreamSim ↑	Avg. IoU ↑
Cosmos-Predict1-7B-Video2World	prompt + 1 frame	17.34	0.538	0.836	0.332
Cosmos-Predict1-7B-Video2World	prompt + 9 frames	**21.06**	**0.691**	0.859	0.592
Cosmos-Predict1-14B-Video2World	prompt + 1 frame	16.81	0.521	0.836	0.338
Cosmos-Predict1-14B-Video2World	prompt + 9 frames	20.21	0.635	0.860	**0.598**
Cosmos-Predict1-4B	1 frame	17.91	0.486	0.827	0.394
Cosmos-Predict1-4B	9 frames	18.13	0.482	0.859	0.481
Cosmos-Predict1-5B-Video2World	prompt + 1 frame	17.67	0.478	0.818	0.376
Cosmos-Predict1-5B-Video2World	prompt + 9 frames	18.29	0.481	0.864	0.481
Cosmos-Predict1-12B	1 frame	17.94	0.486	0.829	0.395
Cosmos-Predict1-12B	9 frames	18.22	0.487	**0.869**	0.487
Cosmos-Predict1-13B-Video2World	prompt + 1 frame	18.00	0.486	0.830	0.397
Cosmos-Predict1-13B-Video2World	prompt + 9 frames	18.26	0.482	0.865	0.482

초록 음영 = 9-frame conditioning — 같은 모델의 1-frame 조건 대비 전 지표에서 일관된 향상

결과 해석과 관찰된 물리 위반 유형

정량·정성 결과에 기반한 논문의 관찰과, rigid-body 시뮬레이션이 드러낸 구체적 실패 사례들

관찰 (quantitative & qualitative)

① 더 많은 conditioning frame → 더 나은 전반적 object kinematics 예측
- 속도·가속도 같은 1차·2차 물리량(quantities)을 더 잘 추론할 수 있기 때문 (당연한 결과로 서술)

② 9-frame 조건에서 diffusion WFM이 AR WFM보다 pixel-level 예측 우수
- diffusion 기반 WFM이 더 높은 시각 품질의 비디오를 렌더링한다는 시각적 관찰과 상관

③ 더 큰 모델이 physics alignment에서 더 낫다는 결과는 나타나지 않음
- 더 큰 모델이 더 높은 시각 품질을 렌더링하긴 하지만, **모든 WFM이 physics 준수에 똑같이 어려움을 겪음** — 더 나은 데이터 큐레이션과 모델 설계 필요

Failure Rate Analysis

rigid-body 시뮬레이션만으로도 이미 WFM의 한계를 시험하며, 구체적 실패 사례 식별에 유용한 도구로 기능 :

- Low-level 문제 : **object impermanence**(객체의 자발적 출현·소멸), **deformation**(형태 변화)
- 복합적 문제 : **implausible kinematics · violation of gravity** 등

이러한 structured simulation은 physics alignment 테스트에 유용한 methodology — 향후 다음과 같이 개선 예정 :

- **더 복잡한 시나리오 도입**
- **photorealism 강화로 sim-to-real gap 축소** (WFM pre-training 데이터는 실제 비디오이므로)
- **더 포괄적인 물리 이해 평가를 위한 evaluation metric 정제**

PART 3 정리

5.3 Evaluation

5.3 평가 축	3D Consistency와 Physics Alignment 두 축으로 pre-trained WFM 평가 · 포괄적 평가는 future work
5.3.1 설계	RealEstate10K 500개 static scene · VLM 캡션 · VideoLDM baseline · Sampson error(수식 10, SuperPoint+LightGlue, 8-point RANSAC) + pose 성공률 + 3DGS novel view(PSNR/SSIM/LPIPS)
5.3.1 결과	Sampson 0.841→0.355 · pose 성공률 4.4%→62.6/68.4% (실제 비디오 56.4% 상회) · held-out view 합성 품질 전 지표 우위 → 효과적인 world simulator 입증
5.3.2 설계	PhysX+Isaac Sim 8개 시나리오(낙하·경사면·U-슬로프·스택×2·도미노·시소·자이로스코프) · 800개 1080p·100f 비디오 · 4 static view · 1/9-frame 조건 + kinematic caption
5.3.2 지표	Pixel(PSNR/SSIM) · Feature(DreamSim) · Object(SAMURAI 전파 mask IoU) 3-tier · 33 frames · 4 seeds 평균
5.3.2 결과	9-frame 조건 일관 우세(속도·가속도 추론) · pixel 수준은 Diffusion > AR · 모델 스케일은 physics alignment을 개선하지 않음
실패·전망	object impermanence · deformation · implausible kinematics · gravity 위반 관찰 → 시나리오 확장·photorealism 강화·metric 정제 예정

NVIDIA COSMOS · WORLD FOUNDATION MODEL PLATFORM

Chapter 6

Post-trained World Foundation Model

6.1 Post-training WFM for Camera Control
Dataset · Fine-tuning · Evaluation — 3D navigable world generation

6.2 Post-training WFM for Robotic Manipulation
Datasets · Fine-tuning · Evaluation — instruction/action 기반 예측

6.3 Post-training WFM for Autonomous Driving
Datasets · Fine-tuning · Evaluation — multi-view 주행 세계 생성

Cosmos Post-trained World Foundation Model

6 6. Post-trained World Foundation Model 핵심정리

Physical AI를 위한 Cosmos WFM의 도메인 특화 Post-training 구조

1 핵심 메시지

- Post-trained WFM = 범용 pre-trained WFM을 특정 Physical AI 환경에 맞게 fine-tuning한 도메인 특화 world simulator
- 논문은 production model이 아니라 sample application / reference implementation 성격
- 핵심 응용: Camera Control, Robotic Manipulation, Autonomous Driving

Pre-trained WFM → Domain Data + Condition Signal → Post-trained WFM

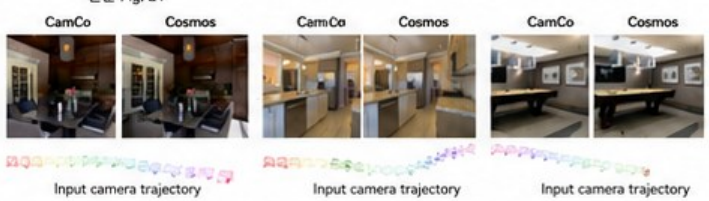
★ 핵심: WFM은 비디오 생성모델을 넘어 action-conditioned simulator로 확장된다.

2 Post-training 모델 맵 (Table 21 요약)

응용 분야	모델	조건 입력	역할
Camera Control	Cosmos-Predict1-7B-Video2World-Sample-CameraCond	Text + Image + Cameras	3D navigable world generation
Robot Manipulation	Sample-Instruction	Text + Video	instruction-conditioned future prediction
Robot Manipulation	Sample-ActionCond	Action + Video	action-conditioned next-frame prediction
Autonomous Driving	Sample-MultiView	Text	multi-view driving video generation
Autonomous Driving	Sample-MultiView-TrajectoryCond	Text + Trajectory	trajectory-following multi-view simulation
Autonomous Driving	Sample-MultiView Video2World	Text + Video	long-horizon multiview extension

3 6.1 Camera Control

논문 Fig. 21



- 입력: single reference image + camera trajectory
- 방법: camera pose를 Plücker embedding으로 변환하여 latent와 결합
- 데이터셋: DL3DV-10K 학습, RealEstate10K 평가
- 활용: moving forward / backward / left / right 형태의 3D 탐색 세계 생성

정량 결과 (Table 22)

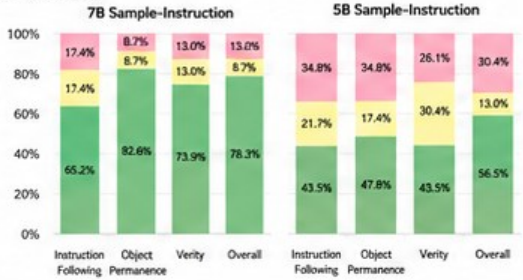
모델	Pose success ↑	Rot err (°) ↓	Trans err (m) ↓	FID ↓	FVD ↓
CamCo	43.0%	8.277	0.185	57.49	433.24
Cosmos CameraCond (Ours)	82.0%	1.646	0.038	14.30	120.49

✓ CameraCond는 고품질 영상과 3D 일관성을 유지하면서 trajectory를 정확히 따른다.

4 6.2 Robotic Manipulation

A) Instruction-based Video Prediction

- 데이터셋: Cosmos-1X, humanoid robot EVE, 약 200시간, 12,000 episodes
- 입력: 현재 video + text instruction
- 출력: instruction 수행 예상 미래 video

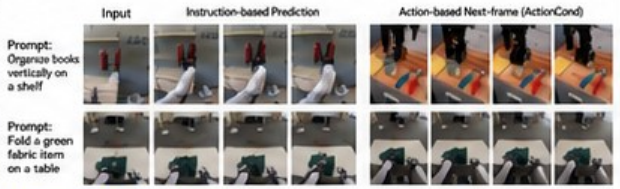


7B Sample-Instruction 전체 선호도(Overall) **78.3%**

5B Sample-Instruction 전체 선호도 56.5%

B) Action-based Next-frame Prediction 논문 Fig. 25–26

- 데이터셋: Bridge, 약 20,000 episodes, kitchen robot arm, 7D action vector
- 입력: 현재 frame + action vector
- 출력: next frame, autoregressive rollout 가능



ActionCond 정량 결과 (Table 23)

모델	PSNR (↑)	SSIM (↑)	Latent L2 (↓)	FVD (↓)
IRASim-Action	19.13	0.64	0.38	593
Cosmos 5B (ActionCond)	19.95	0.80	0.36	434
Cosmos 7B (ActionCond)	21.14	0.82	0.32	190

✓ Post-trained WFM은 instruction-following과 action rollout 기반 로봇 planning/simulation에 강점

5 6.3 Autonomous Driving

논문 Fig. 27 (Multi-view Generation)



논문 Fig. 27 (Trajectory-conditioned Multi-view)



- 특징: 6-view multi-camera 자율주행 world model 구축
- 데이터셋: RDS (Real Driving Scene), 약 360만 개 20초 surround-view clips, 약 20,000시간, 6 cameras (front/left/right/rear/rear-left/rear-right)
- 모델: MultiView / MultiView-TrajectoryCond / Video2World-Sample-MultiView
- 특징: view embedding, view-dependent cross-attention, trajectory control condition

A) Multi-view 일관성 (Table 24)

모델	FID (↓)	FVD (↓)	TSE (↓)	CSE (↓)
VideoLDM-MultiView	60.14	884.46	1.24	6.48
Cosmos MultiView	32.95	210.23	0.68	2.11
Cosmos TrajectoryCond	-	-	0.59	2.02
Real Videos	-	-	0.69	1.71

B) Trajectory 일관성 (Table 25)

모델	TAE-ATE (m) (↓)	RPE-R (deg/m) (↓)	RPE-t (m/m) (↓)	TFE (cm) (↓)
VideoLDM-MultiView	0.88	22.94	0.77	-
Cosmos MultiView	0.77	4.25	0.29	-
Cosmos TrajectoryCond	0.54	4.31	0.18	20.20
Real Videos	0.49	4.60	0.14	13.49

✓ TrajectoryCond는 입력 경로를 정밀하게 따라가며 multi-view geometry consistency를 향상시킨다.

6 구조적 의미와 시사점

1) Generalist → Specialist
대규모 pre-training 후 소규모 도메인 데이터로 특화

2) Conditioned Simulation
camera pose, instruction, action, trajectory가 세계 변화의 조건 입력

3) Physical AI 활용
policy evaluation, planning, MPC, synthetic data generation의 기반

갚의 핵심 걸론

1. Camera control → 3D navigable world
2. Manipulation → instruction/action-conditioned simulator
3. Driving → multi-view trajectory-aware simulator
4. Post-training → Physical AI용 도메인 특화 world simulator 완성

한 줄 요약: 6장 Post-trained WFM은 범용 Cosmos WFM용 Camera, Robot, Driving 조건으로 Fine-tuning하여 Physical AI용 도메인 특화 시뮬레이터로 전환하는 방법을 제시함

Pre-trained WFM → 다양한 Physical AI 응용으로 fine-tune

Camera control(3D navigable world) · 두 Robotic setup의 action control · AV 학습을 위한 multi-view 지원

Table 21: A map of Post-trained WFMs discussed in Sec. 6.

Section	Model	Condition(s)
Sec. 6.1	Cosmos-Predict1-7B-Video2World-Sample-CameraCond	Text + Image + Cameras
Sec. 6.2	Cosmos-Predict1-7B-Video2World-Sample-Instruction	Text + Video
Sec. 6.2	Cosmos-Predict1-7B-Video2World-Sample-Instruction	Text + Video
Sec. 6.2	Cosmos-Predict1-7B-Video2World-Sample-ActionCond	Action + Video
Sec. 6.2	Cosmos-Predict1-7B-Video2World-Sample-ActionCond	Action + Video
Sec. 6.3	Cosmos-Predict1-7B-Text2World-Sample-MultiView	Text
Sec. 6.3	Cosmos-Predict1-7B-Text2World-Sample-MultiView-TrajectoryCond	Text + Trajectory
Sec. 6.3	Cosmos-Predict1-7B-Video2World-Sample-MultiView	Text + Video

"-Sample" 접미사의 의미 : pre-trained WFM의 sample application 제공이 목적임을 강조 — 이 모델들은 어떤 실제 응용을 위한 완전한 시스템이나 production 모델이 아님
개발자는 자신의 Physical AI setup과 target 응용을 위해 custom dataset으로 WFM을 fine-tune 해야 함

Camera Control Post-training & Dataset (6.1.1)

Camera pose conditioning으로 Cosmos-Predict1-7B-Video2World를 효과적인 3D world simulator로 전환

목표 · Cosmos-Predict1-7B-Video2World-Sample-CameraCond

- **단일 reference 입력 이미지로부터 3D 세계를 생성**하는 데 집중 — camera control을 활용해, 지정된 camera trajectory로부터 시점 변화가 장면의 underlying 3D 구조와 정렬되는 **temporally coherent · 3D-consistent 비디오 시뮬레이션**을 생성

6.1.1 Dataset · DL3DV-10K (Ling et al., 2024)

데이터	**static scene**의 대규모 비디오 데이터셋 **DL3DV-10K** 사용
전처리	모든 비디오를 256 frames 단위 clip으로 chunk
Pose 주석	**GLOMAP**(Pan et al., 2025)으로 chunk된 clip에 **structure-from-motion** 실행 → clip 내 모든 frame에 **dense camera pose 주석** 획득
좌표 기준	첫 frame의 camera pose를 identity transform으로 설정하고, 이후 모든 frame의 relative camera pose 계산
캡션	proprietary VLM으로 비디오를 캡셔닝하여 static scene으로 기술하는 text prompt 획득

Plücker Embedding 기반 Camera Conditioning

Sampled latent embedding에 latent와 동일한 공간 차원의 Plücker embedding을 concat하여 camera control 조건 추가

Plücker Coordinates (수식 11 · Sitzmann et al., 2021)

$$r = (d, m) \in \mathbb{R}^6 \text{ where } m = c \times d \qquad (11)$$

- c : camera center 위치, d : 각 latent pixel의 unit ray direction (latent embedding을 downsample된 이미지로 취급)
- 모든 camera pose는 initial frame에 대한 상대(relative) pose

Temporal 압축과의 정렬

- Cosmos-Predict1-7B-Video2World가 사용하는 **Cosmos-Tokenize1-CV8×8×8-720p**는 temporal 압축률 8×
- 따라서 매 8 frames마다 **4번째 frame의 Plücker embedding**을 해당 latent representation과 concat

학습 설정

- 학습 비디오 입력 frame을 **704 × 1252로 resize** 후 **reflection으로 704 × 1280 padding**
- 학습 중 **57 frames 샘플링**
- 학습 objective와 기타 hyper-parameter는 base Diffusion WFM 학습(§5.1.3)과 동일

CamCo 대비 평가 : 설정과 지표

단일 reference image가 주어지고, 입력 이미지로부터 future rollout 비디오를 생성하는 설정

Baseline & Test Data

Baseline : CamCo (Xu et al., 2024)

- 이 설정에서 camera-controllable video generation의 SOTA 모델
- 공정한 비교 위해 DL3DV-10K 학습셋으로 fine-tune된 CamCo 사용
- 본 모델은 57 frames, CamCo는 14 frames만 생성 → 같은 57-frame trajectory를 CamCo용으로 4× temporal downsample하여 비교
- CamCo 비디오 해상도는 256×256으로 제한 → 입력·테스트 frame을 최대 center-crop하여 평가

Test data

- §5.3.1과 동일한 RealEstate10K test set 500개 샘플
- 초기 frame을 reference image로, 데이터셋 제공 camera trajectory를 control 입력으로 사용
- trajectory 양 끝 간 거리가 1로 정규화되도록 rescale

Metrics (Xu et al., 2024를 따름)

① Video generation quality

- FID (frame 수준) · FVD (video 수준)
- 같은 test data를 reference video로 사용해 계산 (pixel 수준 비교에는 사용하지 않음)

② 3D consistency — camera trajectory alignment

- structure-from-motion 라이브러리로 camera pose를 재추정하여 입력 camera control trajectory와 비교
- N frames에 대해 :

$$\epsilon_{rot} = \frac{1}{N}\sum_{i=1}^{N} \cos^{-1}\left(\frac{\mathrm{trace}\left(R_i^{\top}\hat{R}_i\right)-1}{2}\right), \quad \epsilon_{trans} = \frac{1}{N}\sum_{i=1}^{N}\|t_i - \hat{t}_i\| \qquad (12)$$

- pose 추정의 similarity transform 모호성 보정 : Lin et al.(2021)을 따라 Procrustes analysis로 예측 trajectory를 GT에 정렬

정량 비교 (Table 22) & 정성 비교 (Figure 21)

생성 품질(FID/FVD)과 camera trajectory alignment 모두에서 CamCo를 큰 폭으로 상회

Table 22: Quantitative comparison of post-trained WFM with camera control.

	Camera Trajectory Alignment			Video Generation Quality	
Method	Pose estimation success rate (%) ↑	Rotation error (°) ↓	Translation error ↓	FID ↓	FVD ↓
CamCo (Xu et al., 2024)	43.0%	8.277	0.185	57.49	433.24
Cosmos-Predict1-7B-Video2World-Sample-CameraCond	**82.0%**	**1.646**	**0.038**	**14.30**	**120.49**

논문의 비교 결과 (Comparisons)

- 낮은 FID/FVD(높은 시각 품질)와 높은 pose estimation 성공률 → 현실적이고 일관된(coherent) 3D 세계 생성
- camera trajectory 재추정 결과가 원래 control 입력에 훨씬 근접 → **더 우수한 camera control**
- CamCo는 입력 이미지 너머의 content 생성에 어려움을 겪고, data distribution shift로 부정확한 trajectory 또는 pose 추정 불가능한 out-of-distribution 합성 생성
- **두 모델 모두 DL3DV-10K로 post-train, RealEstate10K로 평가 — 큰 분포 이동에도 Cosmos 모델은 이를 극복하고 unseen 입력 camera trajectory에 일반화**



Figure 21: **Qualitative comparison of camera control models**.
Given the input frame and camera trajectory (color-coded temporally from red to violet), we compare Cosmos-Predict1-7B-Video2World-Sample-CameraCond against CamCo (Xu et al., 2024) on the generated future frames as well as the re-estimated camera poses. CamCo suffers from the data distribution shift and often generates inaccurate trajectories or even out-of-distribution image syntheses that lead to un-estimatable camera poses. In contrast, the Cosmos camera control model can successfully generate future frames aligned with the camera control input while also maintaining high video quality and 3D consistency.

Joystick 제어 · 시드 다양성 (Figure 22 · 23)

카메라 제어로 시뮬레이션된 세계를 탐색하고, 같은 조건에서 서로 다른 미래를 시뮬레이션

Figure 22 · Joystick-like Control

- 4가지 joystick식 카메라 제어 : **moving forward · moving backward · rotating left · rotating right**
- 사용자가 joystick으로 모델을 제어하며 미래 video frame을 생성 — 시뮬레이션된 세계를 내비게이션하는 use case 시연
- Physical AI 에이전트도 이러한 제어로 서로 다른 시나리오 하의 세계의 미래를 예측 가능
- *생성 비디오에서 frames 14 · 28 · 42 · 57 시각화*

Figure 23 · 같은 입력, 다른 random seed

- 같은 입력 이미지·camera control에 대해 서로 다른 random seed로 생성한 결과 제시 (moving backward · rotating right 그룹, seed 3개씩)
- 비디오의 3D 공간·시간적 일관성(coherence)을 유지하면서도 **서로 다른 세계를 생성**
- 현재 상태가 주어졌을 때 **서로 다른 가능한 미래(possible futures)를 시뮬레이션**하는 데 활용 가능

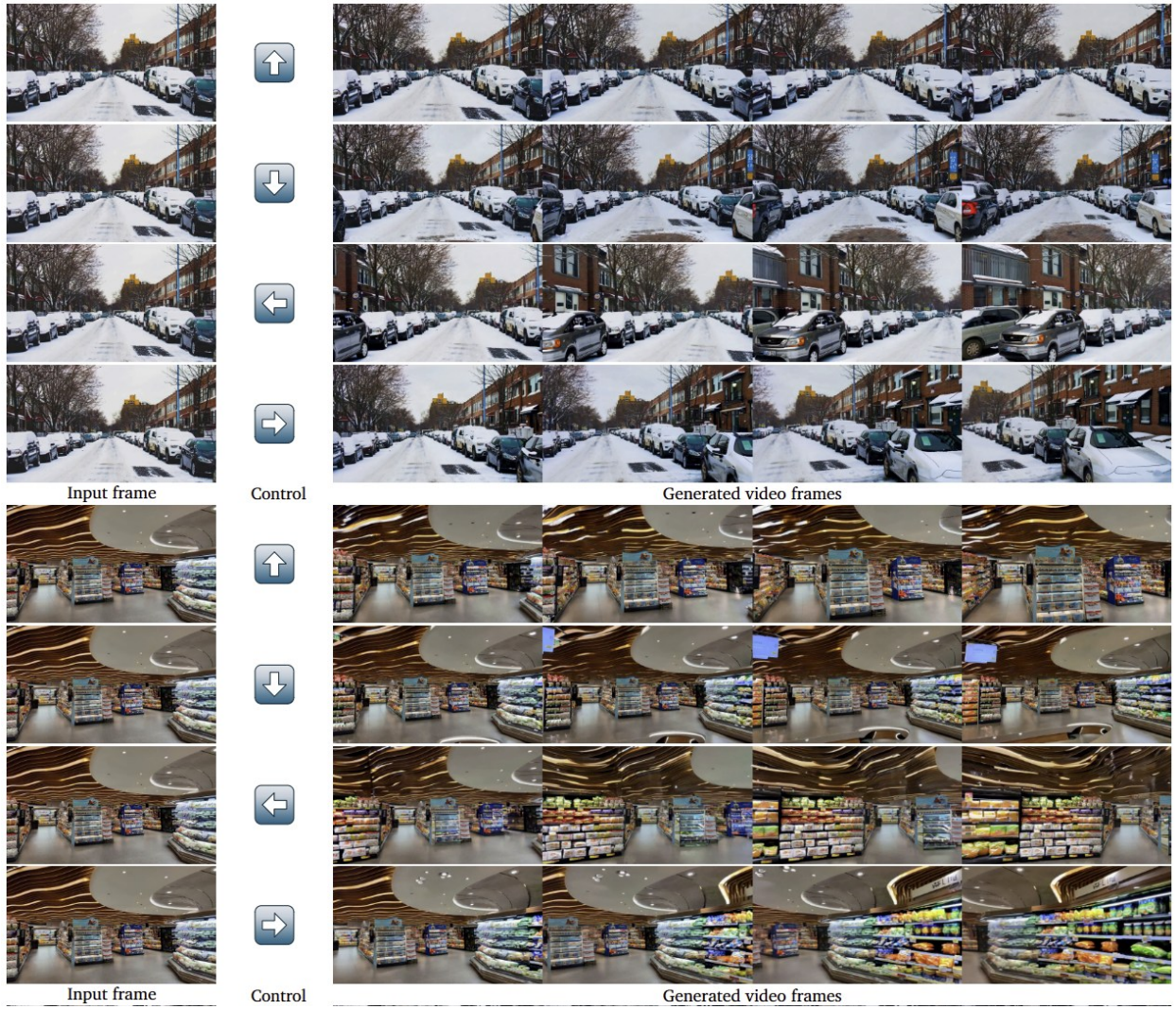


Figure 22: **Cosmos-Predict1-7B-Video2World-Sample-CameraCond results with joystick control.**
For each input frame (left-most column), we apply 4 different camera trajectories created with joystick-like control moving forward, moving backward, rotating left, and rotating right. We visualize frames 14, 28, 42, and 57 from the generated videos.



Figure 23: **Cosmos-Predict1-7B-Video2World-Sample-CameraCond results with different seeds**.
We show the capability of simulating diverse futures with our camera control model given the same input image and camera condition. For each group, we apply the same input frame and camera condition created with joystick. The first group shows moving backward and second group shows rotating right. Within each group, we show the generated videos with 3 different random seeds in each column. We visualize frames 19, 38, and 57 from the generated videos.

World Model = 로봇 조작을 위한 Planner & Simulator

Pre-trained WFM을 두 가지 robotic manipulation 과제로 fine-tune

Task 1 · Instruction-based Video Prediction

입력

- 로봇의 현재 video frame
- text instruction

출력

- instruction을 따르는 로봇의 예측 비디오

Task 2 · Action-based Next-Frame Generation

입력

- 로봇의 현재 video frame
- 현재-다음 frame 사이의 action vector

출력

- 지정된 action 수행 결과를 보여주는 예측 next frame

➢ **Action sequence가 주어지면, 모델을 autoregressive하게 실행하여 로봇이 해당 action들을 수행하는 비디오를 예측할 수 있음 — world model이 로봇 조작의 강력한 planner이자 simulator가 될 잠재력**

Cosmos-1X (instruction) · Bridge (action)

두 과제 각각을 위해 큐레이션한 두 개의 데이터셋

Cosmos-1X · Instruction-based Video Prediction용

- 내부(internal) 데이터셋
- **1x.Tech의 휴머노이드 로봇 EVE**가 다양한 과제를 수행하며 촬영한 **egocentric 비디오 약 200시간**
- 과제 : navigation, folding clothes, cleaning tables, picking up objects 등

- raw 비디오에서 **1~9초 길이의 episode 약 12,000개** 선별
- 각 episode에 한 문장 instruction 라벨 부여 → 이후 proprietary VLM으로 upsample

- 촬영 조건 : **30 FPS · 512 × 512 해상도**

Bridge (Ebert et al., 2022) · Action-based Generation용

- 공개(public) 데이터셋 — 비교를 위해 선행 연구(IRASim, Zhu et al., 2024)와 동일한 구성 사용
- 주방 환경에서 다양한 과제를 수행하는 robot arm의 **third-person view episode 약 20,000개**
- 촬영 조건 : **320 × 256 해상도 · 5 FPS**

- 각 video frame의 action은 OpenVLA(Kim et al., 2024)와 같이 gripper 좌표 공간의 **7차원 벡터**로 정의 :

$$\Delta x, \Delta y, \Delta z, \Delta\theta_r, \Delta\theta_p, \Delta\theta_y, \Delta\text{Gripper}$$

Diffusion(7B)·AR(5B) 각각을 두 과제로 fine-tune — 총 4개 모델

Cosmos-Predict1-7B-Video2World(§5.1) · Cosmos-Predict1-5B-Video2World(§5.2)를 instruction/action 과제에 각각 적용

Instruction-based Video Prediction

- 모델 : **Cosmos-Predict1-7B-Video2World-Sample-Instruction** · **Cosmos-Predict1-5B-Video2World-Sample-Instruction**
- instruction의 **T5 embedding**을 계산하여 **cross-attention**을 통해 base 모델 fine-tuning에 결합

Action-based Next-Frame Prediction — action은 pre-training에 없던 새 modality

Cosmos-Predict1-5B-Video2World-Sample-ActionCond

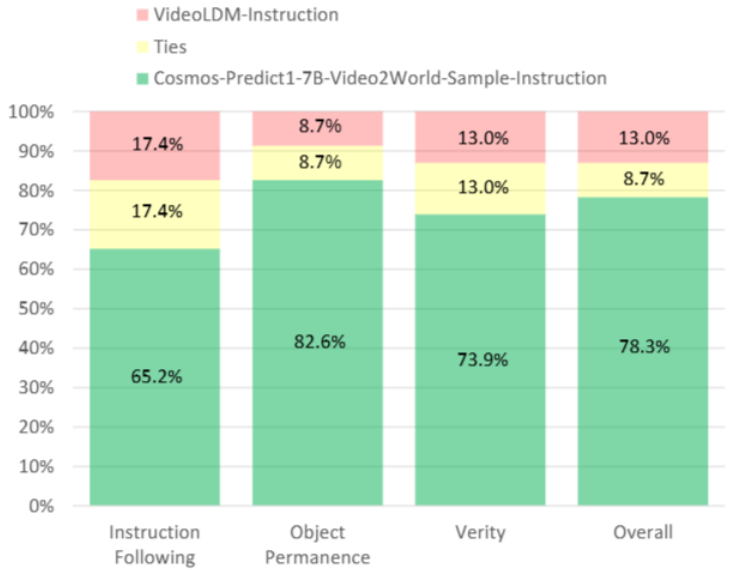
- action embedder MLP 추가 : action vector를 tensor로 project
- 이를 **cross-attention**을 통해 모델에 결합

Cosmos-Predict1-7B-Video2World-Sample-ActionCond

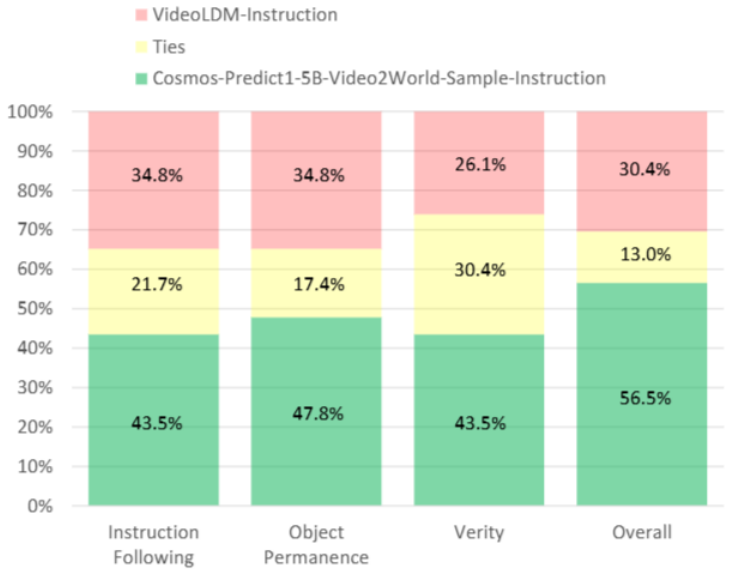
- 마찬가지로 action embedder MLP로 action을 tensor로 project
- 단, **DiT 모듈의 timestamp embedding에 더하는 방식**으로 결합

Human Evaluation : VideoLDM-Instruction 대비 (Figure 24)

VideoLDM을 Cosmos-1X로 fine-tune한 VideoLDM-Instruction을 baseline으로, 익명 비디오 쌍 비교



(a) Cosmos-Predict1-7B-Video2World-Sample-Instruction vs.VideoLDM-Instruction.



(b) Cosmos-Predict1-5B-Video2World-Sample-Instruction vs.VideoLDM-Instruction.

Figure 24: **Human evaluation results for instruction-based video prediction on Cosmos-1X dataset**. The results show that compared with the baseline model (VideoLDM-instruction), our fine-tuned models (Cosmos-Predict1-7B-Video2World-Sample-Instruction and Cosmos-Predict1-5B-Video2World-Sample-Instruction) have higher preferences in the four evaluation dimensions.

평가 설계 & 결과

4개 평가 차원

- Instruction following — 입력 언어 지시와의 정렬
- Object permanence — 장면 내 객체가 비디오 내내 유지되는가
- Verity — 예상 밖의 상상 객체 없이 실제 세계를 충실히 표현하는가
- Overall — 로봇이 계획을 세우기에 합리적인 비디오인가

절차

- 같은 instruction으로 생성된 익명 비디오 쌍을 4개 차원에서 비교
- 인간 평가자 10명 · 테스트 episode 23개

결과

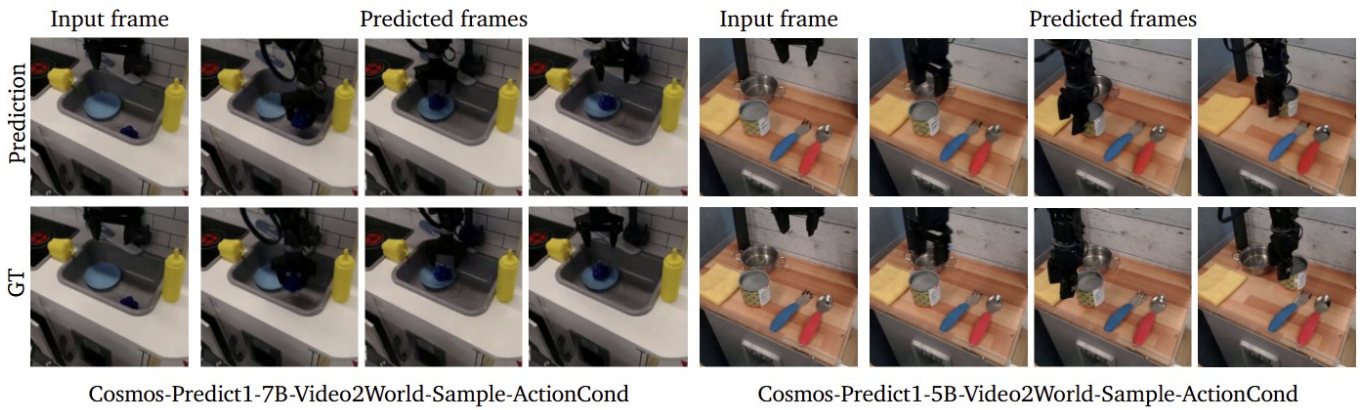
- 7B·5B 모두 4개 차원 전부에서 VideoLDM-Instruction보다 높은 선호도
- 7B-Sample-Instruction의 Overall 선호도 **78.3% vs 13.0%** (VideoLDM-Instruction)

Bridge test set 평가 : Table 23 & Figure 26

IRASim을 fine-tune한 IRASim-Action baseline과 비교
next-frame 예측을 autoregressive하게 수행해 비디오 생성 · 공식 Bridge test set에서 무작위 100 episodes

Table 23: Evaluation of action-based next-frame prediction on Bridge dataset.

Method	PSNR ↑	SSIM ↑	Latent L2 ↓	FVD ↓
IRASim-Action	19.13	0.64	0.38	593
Cosmos-Predict1-5B-Video2World-Sample-ActionCond	19.95	0.80	0.36	434
Cosmos-Predict1-7B-Video2World-Sample-ActionCond	**21.14**	**0.82**	**0.32**	**190**



결과 요약

- 5B·7B ActionCond 모두 baseline(IRASim-Action)을 모든 지표에서 상회
- Figure 26 : 예측 video frame이 GT video frame과 매우 유사 — 두 모델 모두 고품질 예측
- *평가지표 : PSNR · SSIM · Latent L2(Zhu et al., 2024) · FVD*

Figure 26: **Action-based next-frame prediction samples on the Bridge dataset**. The left is the results of the Cosmos-Predict1-7B-Video2World-Sample-ActionCond model, and the right is the results of the Cosmos-Predict1-5B-Video2World-Sample-ActionCond model. As shown, the predicted video frames closely match the GT video frames for both models.

Multi-view 주행 World Model & RDS Dataset (6.3.1)

대부분의 자율주행차는 여러 방향의 카메라 장착 — 이상적인 AV world model은 target 차량의 센서 setup에 맞는 multi-view 모델이어야 함

Real Driving Scene (RDS) Dataset · 내부 데이터셋

- NVIDIA 내부 driving platform으로 수집한 **약 360만 개의 20초 surround-view video clip (≈ 20,000시간)**
- 각 clip은 **6개 camera view** (front · left · right · rear · rear-left · rear-right)로 기록 · ego-motion 정보 포함(trajectory 데이터 구성에 사용) · front camera timestamp로 다른 view 들의 frame 동기화

속성(attribute) 기반 분포 매칭

대규모 라벨링 데이터에서 target 분포에 맞게 선별 — 속성 태그 :

- Contender vehicle density (none/low/medium/high)
- Weather (clear/raining/snowing/fog)
- Illumination (day/night)
- Ego vehicle speed (standing/low/local/highway)
- Ego vehicle behavior (곡률·가속 high/medium/low)
- Road type/population density (OpenStreetMap 기준 rural/residential/urban)

증강 & 캡션

- 희귀 도로 구조물(tollbooths · bridges · tunnels · speed bumps 등)을 포함한 clip 의 최소 개수 확보를 위해 2차 data-mining으로 증강
- 각 camera view의 비디오를 개별적으로 캡셔닝, 다음 템플릿 문자열로 시작 :

“The video is captured from a camera mounted on a car. The camera is facing forward|left|right|backward|rear-left|rear-right.”

Cosmos-Predict1-7B-Text2World → 3개의 Multi-View World Model

일관된 multi-view 생성을 위해 §5.1 아키텍처를 약간 수정, 6개 카메라 view를 동시에 생성하도록 fine-tune

1 **Cosmos-Predict1-7B-Text2World-Sample-MultiView**
text prompt 입력을 기반으로 6개 camera view를 생성하는 multi-view world model

2 **Cosmos-Predict1-7B-Text2World-Sample-MultiView-TrajectoryCond**
위 모델 위에 구축, trajectory 입력을 추가 conditional 입력 신호로 받음

3 **Cosmos-Predict1-7B-Video2World-Sample-MultiView**
video 기반 conditioning을 지원하도록 fine-tune — 이전 frame들을 생성 과정에 결합. Text2World-Sample-MultiView의 video 출력을 받아 그 연장(extension)을 생성 가능

➢ **3개 모델 모두 848 × 480 해상도로 6개 view × 57 frames의 비디오를 출력**

참고 : 선행 연구들처럼 per-interval action vector 기반의 더 세밀한 제어 신호도 가능하지만(§6.2의 robotic manipulation처럼), 이러한 확장은 future work로 남김

Multi-view 일관성을 위한 3가지 설계

View embedding · View-dependent cross-attention · Trajectory control condition

View-independent PE + View Embedding

- FPS-aware 3D RoPE를 view 차원까지 확장하는 대신, §5.1과 동일한 positional embedding을 각 view에 독립적으로 적용
- view 간 차이 표현을 위해 denoising function $D\theta$가 추가 view embedding을 입력으로 받도록 수정
- **즉, camera view 정보는 positional embedding이 아니라 global view embedding으로 공급**

STEP 2

- 같은 장면의 6개 view는 각각 다른 video description을 가짐
- 6개 view 전체를 diffusion process의 state로 취급하여 모든 요소 간 self-attention을 수행하되,
- **텍스트 입력에 대해서는 view-dependent cross-attention이 유리 : 각 view의 cross-attention은 해당 view의 텍스트 설명에만 attend**

STEP 3

- text 조건에 더해, 주어진 미래 trajectory 경로를 따르는 비디오를 생성하도록 fine-tune → 에이전트의 더 정밀한 제어
- **trajectory = 3D 공간의 64개 점 시퀀스 — 초기 위치 (0,0,0)에서 목적지까지의 translation, 각 점은 0.1초 간격**
- trajectory 입력의 embedding을 계산해 fine-tune된 모델 denoiser의 conditional 입력으로 사용 · 실제 주행 기록 trajectory와 텍스트로 지정한 주행 환경을 모두 준수하는 고유한 시나리오 생성 가능

Text-conditioned 6-view 생성과 일반화 (Figure 27 · 28)

Text2World-Sample-MultiView로 57-frame 6-view 생성 → Video2World-Sample-MultiView로 201 frames까지 연장



Prompt: The video captures a highway scene with a white truck in the foreground, moving towards the camera. The truck has a large cargo area and is followed by a motorcyclist wearing a full-face helmet. The road is marked with white lines and has a metal guardrail on the right side. The sky is partly cloudy, and there are green trees and bushes visible on the roadside. The video is taken from a moving vehicle, as indicated by the motion blur and the changing perspective of the truck and motorcyclist.

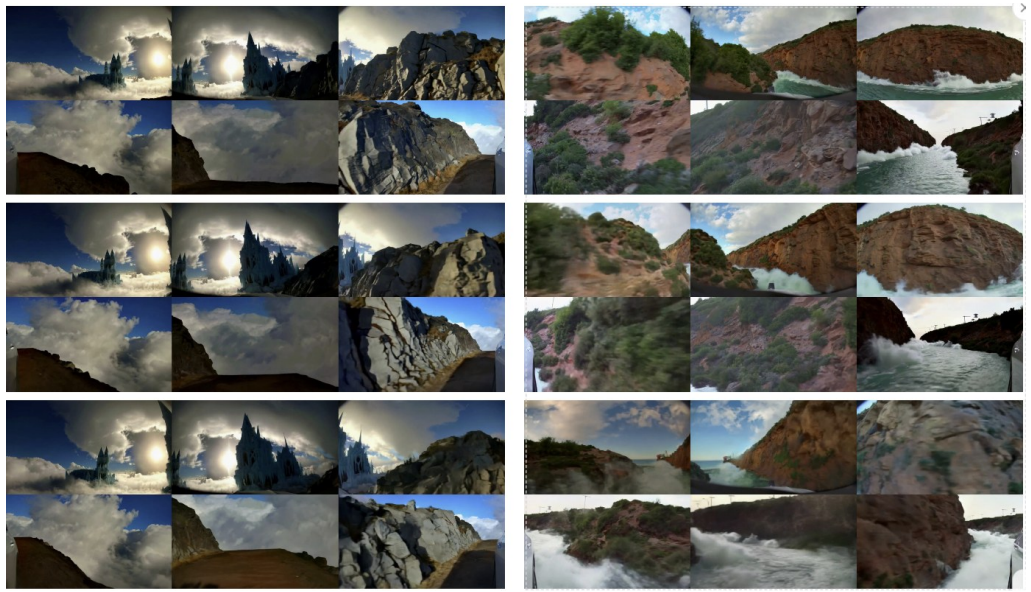
Prompt: The footage shows a multi-car pile-up on a foggy highway. Visibility is severely reduced due to thick fog, with only the taillights of vehicles ahead visible. Suddenly, brake lights flash, and cars begin to swerve and stop abruptly. The highway is cluttered with stopped and crashed vehicles. The surroundings are obscured by fog, adding to the chaos and confusion of the scene.

Figure 27: **Text-conditioned samples generated by Cosmos-Predict1-7B-Text2World-Sample-MultiView,extended to 8 seconds by Cosmos-Predict1-7B-Video2World-Sample-MultiView**. This figure visualizes all six camera views in a group, with each row corresponding to a specific timestamp. The left example depicts a highway scene where a motorcycle is riding alongside a large truck. The right example shows the ego car following a sedan as it takes a right turn in a heavy snowy day.

➢ **post-trained 모델이 pre-trained WFM의 일반화 능력을 효과적으로 보존 — RDS 데이터셋의 rare/out-of-domain 장면 생성 가능**

Text-conditioned 6-view 생성과 일반화 (Figure 27 · 28)

Text2World-Sample-MultiView로 57-frame 6-view 생성 → Video2World-Sample-MultiView로 201 frames까지 연장



Prompt: There are towering, intricately designed ice castle illuminated from within. Ahead, the sky showcases a vibrant sunset and a luminous moon positioned to the left of the castle, casting a blue hue across the scene. Fast-moving, dark, and dramatic clouds add to the otherworldly atmosphere. The 3D realistic art style focuses on lighting and texture, creating a striking visual effect.

Prompt: The footage captures a coastal area where the ocean meets a rugged, rocky cliff. Ahead, vibrant blue waves crash against the rocks, creating white foam. The cliff is a mix of brown and green hues, indicating vegetation and possibly moss or algae.

Figure 28: **Text-conditioned samples generated by Cosmos-Predict1-7B-Text2World-Sample-MultiVieware extended to 8 seconds by Cosmos-Predict1-7B-Video2World-Sample-MultiView.** The post-trained world model effectively preserves its generalization ability. In the left example, the ego car is driving towards an ice castle, while in the right example, the ego car is shown driving on a river.

➢ **post-trained 모델이 pre-trained WFM의 일반화 능력을 효과적으로 보존 — RDS 데이터셋의 rare/out-of-domain 장면 생성 가능**

FID/FVD · TSE/CSE (Table 24)

baseline : 동일 레시피로 fine-tune한 **VideoLDM-MultiView** · 품질 1000 samples, 일관성은 4개 trajectory 유형(직진/좌회전/우회전/기타)별 200개 = 800 samples

Generation Quality

- FID · FVD로 실제 비디오 대비 생성 품질 측정
- 각 비디오에서 16 frames를 추출해 view별 score 계산 → 전체 view 평균을 보고

Multi-View Consistency (확장 Sampson error)

- RDS fisheye intrinsics의 median calibration으로 keypoint를 960×540 · 120° HFoV pinhole로 undistort
- **TSE** : view별 인접 frame 간 Sampson error의 median — 시간적 일관성
- **CSE** : 생성된 view 간 Sampson error의 시간 평균 — F는 전체 temporal frame에 걸쳐 누적된 keypoint로 추정

Table 24: Evaluation on post-trained multi-view world models for multi-view driving video generation.

	Generation Quality		Multi-View Consist.	
Method	FID ↓	FVD ↓	TSE ↓	CSE ↓
VideoLDM-MultiView	60.84	884.46	1.24	6.48
Cosmos-Predict1-7B-Text2World-Sample-MultiView	**32.16**	**210.23**	0.68	2.11
Cosmos-Predict1-7B-Text2World-Sample-MultiView-TrajectoryCond	-	-	**0.59**	**2.02**
Real Videos (Reference)	-	-	0.69	1.71

➢ **두 Cosmos 모델 모두 FID/FVD·TSE/CSE에서 VideoLDM-MultiView를 크게 상회 · trajectory 조건의 명시적 3D guidance로 일관성이 추가 개선 (TrajectoryCond가 최상위)**

TAE / TFE + Object Tracking Consistency (Table 25)

Liang et al.(2024)류의 robust multi-view pose 추정 파이프라인(online dynamic mask + dense bundle adjustment) 사용

TAE (Trajectory Agreement Error) · "front + left-front"와 "front + right-front" 두 camera 구성으로 front camera pose를 각각 추정하고, 두 trajectory의 오차(ATE · RPE-t · RPE-R)로 일치도를 계산 — multi-view 생성의 일관성 반영. trajectory 길이는 1.0으로 정규화, 미세한 카메라 움직임 사례(신호 대기 등)는 제외

Table 25: Trajectory consistency evaluation on post-trained multi-view world models for multi-view driving video generation. The numbers of TAE are scaled by 10^2 for convenience, and the unit for TFE is cm.

Method	TAE-ATE ↓	TAE-RPE-R ↓	TAE-RPE-t ↓	TFE ↓
VideoLDM-MultiView	0.88	22.94	0.77	-
Cosmos-Predict1-7B-Text2World-Sample-MultiView	0.77	**4.25**	0.29	-
Cosmos-Predict1-7B-Text2World-Sample-MultiView-TrajectoryCond	**0.54**	4.31	**0.18**	**20.20**
Real Videos (Reference)	0.49	4.60	0.14	13.49

TFE (Trajectory Following Error)

- 같은 pose 추정 파이프라인으로 front camera trajector를 추정, GT trajectory 조건과 비교 — 모델이 주어진 경로를 얼마나 잘 따르는지 측정
- GT oracle 대비 오차가 **7cm 미만의 근소한 차이** → AV 에이전트 학습에 중요한 정밀한 trajectory 추종 입증

Object Tracking Consistency

- 생성된 8초 비디오에 YOLOv11x로 객체 검출·추적 · 인간 주석자가 물리적으로 불가능한 시나리오(예 : 사람과 차가 하나의 tracked entity로 병합) 식별
- 무작위 20개 비디오 · 157개 객체 중 **물리적으로 불가능한 사례 0건** → 물리적 일관성과 object permanence 입증

Trajectory 조건 생성 (Figure 29)

좌측 trajectory 입력이 주어지면, 이를 따르는 multi-view 비디오 생성 — front camera view 시각화



Figure 29: **Trajectory-conditioned generated samples from Cosmos-Predict1-7B-Text2World-Sample-MultiView-TrajectoryCond**. Given the trajectory inputs on the left-most column, we generate multiview videos that follow the given trajectory. We visualize the front camera view in this figure

CHAPTER6 정리

6 Post-Trained WFM

6.1 Camera Control

DL3DV-10K + GLOMAP pose · Plücker embedding concat → 단일 이미지에서 3D navigable world · CamCo 대비 pose 성공률 43→82%, FVD 433→120

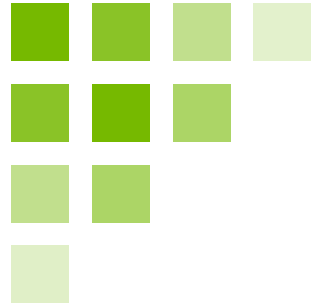
6.2 Robotic Manipulation

RealEstate10K 500개 static scene · VLM 캡션 · VideoLDM baseline · Sampson error(수식 10, SuperPoint+LightGlue, 8-point RANSAC) + pose 성공률 + 3DGS novel view(PSNR/SSIM/LPIPS)

6.3 Autonomous Driving

RDS 20,000시간 6-view · view embedding + view-dependent cross-attn + trajectory 조건 · FID/FVD·TSE/CSE·TAE 전반 우위, TFE < 7cm 차이, 157개 객체 물리 모순 0건

→ 하나의 pre-trained WFM에서 세 도메인의 특화 world model이 파생 — pre-train → post-train 패러다임의 실효성 시연



NVIDIA COSMOS · WORLD FOUNDATION MODEL PLATFORM

Chapter 7

Guardrails

7.1 Pre-Guard
7.1.1 Keyword Blocking · 7.1.2 Aegis Guardrail — 유해 입력(prompt) 차단

7.2 Post-Guard
7.2.1 Video Content Safety Filter · 7.2.2 Face Blur Filter — 유해 출력(video) 차단

7.3 Red Team Effort
10,000+ adversarial prompt-video 쌍 기반 지속 검증

Cosmos Guardrails

7. Guardrails 핵심정리

Physical AI용 WFM의 입력·출력 안전성 계층

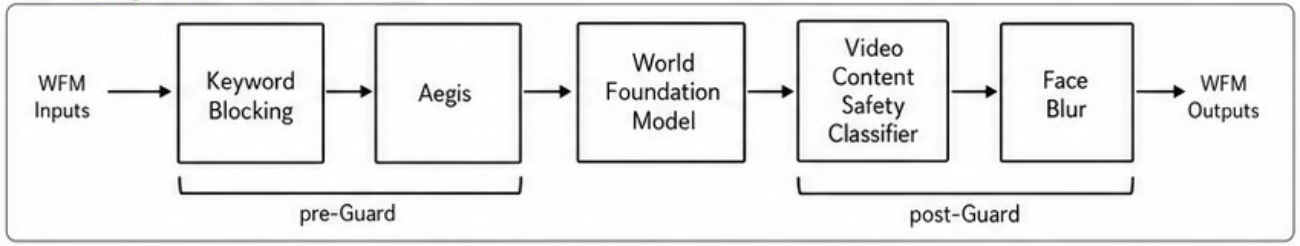
1 핵심 메시지

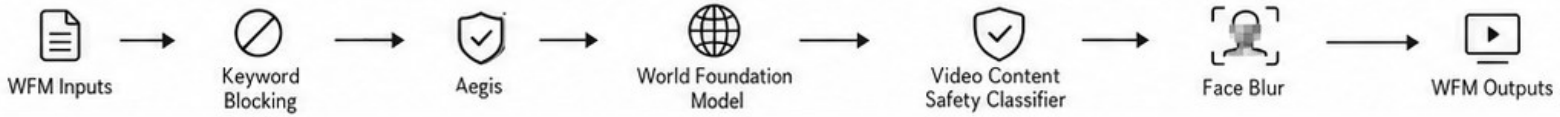
- Guardrails = WFM의 안전한 사용을 위한 입력·출력 양방향 safety envelope
- pre-Guard는 유해 prompt를 사전 차단
- post-Guard는 생성 video output을 사후 검사
- Face Blur와 Red Team까지 포함한 운영형 안전 체계

핵심: 생성 전 + 생성 후를 모두 통제하는 다층 안전 구조

2 전체 Guardrail 구조

논문 Fig. 30 Cosmos Guardrail overview





- 입력 전: Keyword Blocking + Aegis
- 생성 후: Video Content Safety Classifier + Face Blur
- 목적: harmful prompt 차단, unsafe video 차단, 얼굴 비식별화

3 7.1 Pre-Guard

A Keyword Blocking

- hard-coded blockiist 기만 1차 방어선
- WordNetLemmatizer로 단어 원형화 후 비교
- 명시적 유해 키워드 발견 시 prompt 전체 거부

B Aegis Guardrail

- Aegis-AI-Content-Safety-LlamaGuard-LLM-Defensive-1.0 사용
- 의미적으로 복잡한 prompt까지 판별하는 2차 방어선
- unsafe로 판정되면 video 생성 중단

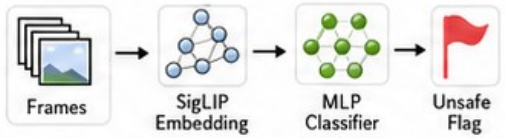
Aegis 위험 범주

- Violence
- Sexual
- Criminal Planning
- Weapons
- Substance Abuse
- Suicide
- Child Sexual Abuse Material
- Hatred / Harassment
- Threat / Profanity

4 7.2 Post-Guard

A Video Content Safety Filter

- frame-level multi-class classifier
- 학습 테이터: 실제 dataset frame + synthetic video + human gold labels
- SigLIP embedding + MLP classifier
- 한 frame이라도 unsafe이면 전체 video를 unsafe로 flag



B Face Blur Filter

- RetinaFace로 high-confidence 얼굴 탐지
- 20×20 pixel보다 큰 얼굴 영역 pixelation
- scene composition은 유지, 개인 식별성은 감소



5 7.3 Red Team Effort

- standard + adversarial prompt로 적극적 probe
- expert annotator가 harm severity 1–5 scale 평가
- unsafe content의 start/end frame까지 라벨링
- 10,000개 이상 prompt-video pair 테스트·주석화

10,000+
prompt-video pair
테스트·주석화 완료

6 구조적 의미

1) Input Safety
유해 instruction과 우회 prompt 차단

2) Output Safety
unsafe visual content와 얼굴 노출 통제

3) Physical AI 보호
downstream policy learning, simulation, synthetic data의 안전성 확보

➢ **한 줄 요약: 7장 Guardrails은 Cosmos WFM의 입력 promp와 출력 video를 모두 관리하여, Physical AI용 world simulator를 안전하게 운영하기 위한 필수 안전 계층을 제시한다**

2단계 Guardrail 시스템 : pre-Guard + post-Guard (Figure 30)

WFM의 안전한 사용을 위해 포괄적인(comprehensive) guardrail 시스템을 개발

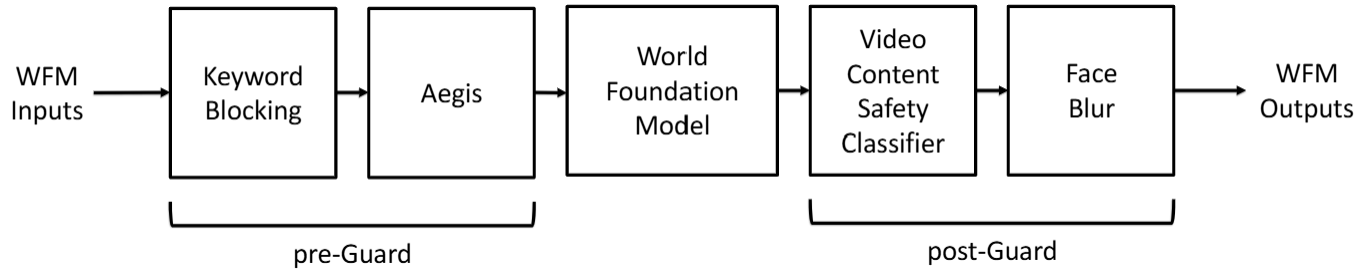


Figure 30: **Cosmos Guardrail overview**. Cosmos Guardrail contains pre-Guard and post-Guard, where pre-Guard blocks inputs based on Aegis (Ghosh et al., 2024) and keywords, while post-Guard blocks outputs based on a video content safety classifier and blurs output face

Stage 1 · pre-Guard (§7.1)

- **Aegis**(Ghosh et al., 2024)와 **keyword list**를 활용해 **유해 prompt를 차단**
- 텍스트 도메인 guardrail — 의미적으로 복잡한 prompt를 위한 LLM 기반 guardrail + 명시적 유해 keyword를 위한 blocklist 기반 checker로 구성

Stage 2 · post-Guard (§7.2)

- **video content safety classifier**와 **face blur filter**로 유해한 시각적 출력을 차단
- 비전 도메인 guardrail — 생성된 출력에 대해 동작

Blocklist 기반 1차 방어선

명시적으로 유해한 생성을 차단하기 위한 blocklist heuristic — unsafe content 생성 위험 완화의 first line of defense

STEP 1

입력 단어 Lemmatization

입력 단어를 **WordNetLemmatizer**로 lemmatize — 영어 lexical database(Miller, 1995)를 사용해 변형(variants)으로부터 root word 추출
예 : “abacii” → root word “abacus”

→

STEP 2

Blocklist 대조

lemmatize된 단어들을 hard-coded blocklist와 비교 — blocklist는 explicit·objectionable 단어들의 대규모 corpus로 구성

→

STEP 3

Prompt 거부

profanity가 하나라도 발견되면 전체 prompt를 거부(reject)

➢ **사용자를 최대한 보호(maximally protect)하기 위해 포괄적인(comprehensive) keyword 집합을 사용**

LLM 기반 2차 방어선 : Aegis Guardrail (LlamaGuard-LLM-Defensive-1.0)

Llama-Guard(Inan et al., 2023)를 NVIDIA Aegis Content Safety Dataset으로 fine-tune한 모델 (Ghosh et al., 2024)

모델 구성과 동작

- NVIDIA의 광범위한 taxonomy — **13개 critical safety risk 카테고리**를 커버하는 Aegis Content Safety Dataset으로 학습
- AEGIS 1.0은 두 가지 버전 : **defensive** 버전과 **permissive** 버전 — defensive 버전이 더 엄격한(tighter) permission boundary 채택
- Cosmos는 **defensive 버전**을 사용해 유해 콘텐츠 생성을 시도하는 잠재적 유해 user prompt를 차단
- 입력 prompt가 unsafe로 분류되면 : **비디오를 생성하지 않고 error message 표시**

Prompt Filter의 Unsafe 분류 카테고리

Violence (폭력)	Sexual (성적 콘텐츠)
Criminal Planning (범죄 계획)	Weapons (무기)
Substance Abuse (약물 남용)	Suicide (자살)
Child Sexual Abuse Material	Hatred (혐오)
Harassment (괴롭힘)	Threat (위협)
Profanity (욕설)	

위 카테고리에 해당하지 않는 prompt는 prompt-filtering 관점에서 safe로 간주

Frame-level Multi-class Safety Classifier

자체 비디오 데이터셋과 생성 결과로 학습 · 일부 class는 safe, 일부는 unsafe · 핵심 과제는 false positive와 false negative의 균형

3종의 Ground-truth 주석 데이터

① VLM 라벨링

자체 데이터셋에서 대규모 비디오를 샘플링, frame을 추출하고 VLM으로 class 결정

② WFM 합성 비디오

corner case와 최소 대표(least-represented) 콘텐츠 카테고리 커버리지 확보를 위해, prompt 집합으로 자체 WFM이 합성 비디오 생성

③ 인간 “gold standard”

인간 주석자가 데이터 일부에 gold standard 라벨 제공 — 필수적인 검증 계층을 더하고 분류기 정확도를 지속 개선

분류기 구조

- 각 video frame에서 **SigLIP**(Zhai et al., 2023) **embedding** 추출
- embedding 위에 **간단한 MLP classifier**를 학습 · 분류 오류 최소화를 위해 학습 데이터를 신중히 balancing

추론 시 동작

- 모든 frame에 대해 SigLIP embedding 생성 → classifier 적용
- **어느 한 frame이라도 unsafe로 분류되면 전체 비디오를 unsafe로 flag**

RetinaFace 기반 얼굴 영역 pixelation

생성 출력의 얼굴 영역을 감지·비식별화하되, Physical AI 응용을 위한 전체 장면 구성은 보존

Detection	Threshold	Obscuring
RetinaFace (Deng et al., 2020) state-of-the-art 얼굴 검출 모델로 높은 confidence score의 facial region 식별	**20 × 20 pixels 초과** 검출된 얼굴 영역 중 20×20 pixels보다 큰 모든 영역이 처리 대상	**Pixelation 적용** 해당 영역에 pixelation을 적용해 비식별화하면서, Physical AI 응용을 위해 전체 장면 구성(scene composition)은 보존

➢ **비식별화(익명성 보호)와 Physical AI가 요구하는 장면 정보 보존의 균형 — post-Guard의 두 번째 vision-domain 필터**

전담 Red Team의 적대적 검증

standard + adversarial example로 시스템을 능동적으로 probing — 내부 attack prompt dataset으로 수집

전문 주석 프로세스

- red team이 생성한 video 출력을, 본 과제를 위해 특별히 훈련된 **expert annotator 팀**이 주석
- §7.1.2의 taxonomy와 관련된 여러 harm 카테고리에 대해 생성 비디오를 **1–5 척도**로 분류
- 주석에는 unsafe content가 감지된 **start / end frame**도 명시 → 고품질 annotation 생성

컴포넌트별 probing & 규모

- red team은 각 guardrail 컴포넌트를 targeted example로 **독립적으로 probing** → 약점 식별 및 edge case 성능 개선

10,000+

distinct prompt-video pairs

논문 출판 시점 기준, 광범위한 unsafe content를 커버하도록 신중히 설계되어 테스트·주석 완료

NVIDIA COSMOS · WORLD FOUNDATION MODEL PLATFORM

Chapter 8

Related Work

① **World Models** — Ha & Schmidhuber(2018)에서 출발한 세계모델 연구의 계보
② **Video Generative Models** — Diffusion·Autoregressive 비디오 생성의 발전
③ **Video Generation with Camera Control** — 3D 일관성과 카메라 제어
④ **Generative Models for Robotic Control** — 로봇 제어를 위한 생성모델
⑤ **Generative Models for Autonomous Driving** — 자율주행 시뮬레이션
⑥ **Tokenizer** — continuous·discrete visual tokenizer 연구

Cosmos Related Works

8. Related Work 핵심정리

NVIDIA

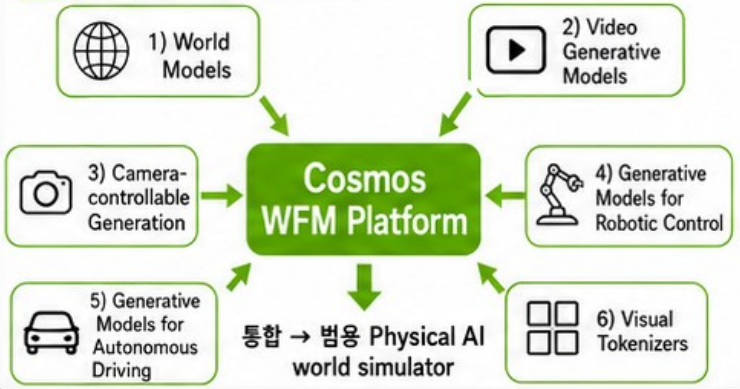
Physical AI용 WFM의 학술적 위치와 연구 계보

1 핵심 메시지

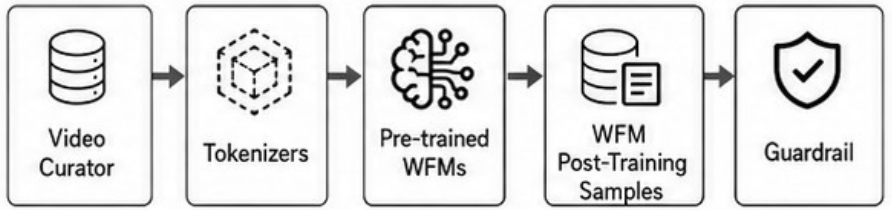
- Cosmos WFM은 기존 world model, video generation, robotic control, autonomous driving, tokenizer 연구를 통합·확장한 플랫폼
- 단순 비디오 생성모델이 아니라 Physical AI용 범용 world simulator 지향
- 핵심 연결축: World Models, Video Generative Models, Camera Control, Robotic Control, Autonomous Driving, Tokenizers
- 관련연구의 통합이 Cosmos의 차별성

핵심: Cosmos는 '개별 생성기술'이 아니라 'Physical AI 개발 플랫폼'으로 위치된다.

2 전체 Related Work 맵

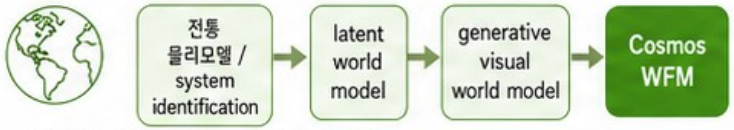


7 논문 Fig. 4 기반 플랫폼 개요



Related Work의 각 축은 결국 Cosmos 플랫폼 구성요소로 수렴

3 World Models



- 현재 상태와 입력으로 미래 상태 예측
- 예측뿐 아니라 planning·decision-making의 기반
- Cosmos는 특정 시스템용 모델을 넘어 대규모 비디오 기반 generalist world model로 확장

4 Video Generative Models

주요 과제	
Text-to-Video	text → video
Image-to-Video	image → future video
Video-to-Video	reference video → transformed/extended video
Action-to-Video	action → outcome video

주요 계열	
Diffusion-based	높은 visual quality
Autoregressive-based	LLM식 scaling/inference 활용

Cosmos는 diffusion WFM + autoregressive WFM을 모두 구축

5 Camera / Robot / Driving

Camera Control

- 3D-consistent view synthesis와 navigable world 생성
- camera trajectory 조건 기반 시점 변화 예측
- Robotics-navigation 응용과 직접 연결

Robotic Control

- diffusion policy / generative planning / future frame prediction 흐름
- instruction-conditioned / action-conditioned simulation으로 확장
- Cosmos는 로봇 policy 자체보다 '미래 세계 변화 예측기'에 초점

Autonomous Driving

- text / image / trajectory / multi-view driving video generation
- 기존 한계: data scale, resolution, view 수 부족
- Cosmos는 multi-view, trajectory-aware driving simulator로 확장

6 Visual Tokenizers

유형	특징	연결 모델
Continuous tokenizer	연속 latent 표현	Diffusion WFM
Discrete tokenizer	양자화 token index	Autoregressive WFM

- image/video를 compact token으로 압축해 대규모 WFM 학습 효율화
- Cosmos는 causal image-video tokenizer suite로 확장

8 구조적 의미

학문적 위치
world model + video model + control 연구의 융합

기술적 차별성
tokenizer, pre-training, post-training, guardrail까지 포함한 end-to-end 구조

Physical AI 관점
예측, 계획, 시뮬레이션, synthetic data 생성 기반

K-Port 시사점
자율항만도 video/sensor 기반 domain-specific WFM으로 확장 가능

➢ **한 줄 요약: 8장 Related Work는 Cosmos WFM이 world model, video generation, camera control, robot control, autonomous driving, tokenizer 연구를 통합해 Physical AI용 범용 world simulator 플랫폼으로 확장된 연구임을 보여줌**

6개 연구 흐름과 Cosmos의 위치

논문 §8이 다루는 연구 영역별 핵심 논지와 Cosmos의 대응을 한 장으로 정리

연구 흐름	§8의 핵심 논지	Cosmos의 위치 (논문 서술)
World Models	RNN 기반 latent 모델에서 시각 관측 공간의 조건부 비디오 생성모델(AR·diffusion)로 발전	AR·diffusion 양 계열의 WFM을 모두 구축
Video Generative Models	diffusion이 시각 품질에서 우위, AR은 통합 modality 처리에 이점	두 계열의 장단점을 모두 취해 Physical AI 개발자에게 제공
Camera Control	neural rendering → data-driven transformer → diffusion 스케일링으로 발전, camera controllability가 활발한 연구 주제	robotics·자율 내비게이션 응용을 위한 camera 제어 post-training (§6.1)
Robotic Control	visuomotor policy · generative motion planning · generative pre-training의 세 흐름	video 예측 pre-training 후 instruction/action 조건화 (§6.2)
Autonomous Driving	기존 연구는 데이터 규모·해상도·카메라 view 수에 제약	pre-trained WFM 활용 — 고해상도·높은 frame rate·multi-view 일관성 달성 (§6.3)
Tokenizer	continuous(AE/VAE)는 diffusion용, discrete(quantizer)는 LLM/VLM 통합용	FSQ·causal 구조 위에 효율·고품질 tokenizer suite 구축 (§4)

기원과 전통적 접근의 한계

"world models" 개념은 Ha & Schmidhuber(2018)의 선구적 연구에서 출발

기원 · Ha & Schmidhuber (2018)

- 현재 상태와 입력이 주어졌을 때 미래 상태를 예측하도록, neural network로 실제 세계의 표현(representation)을 학습하는 개념 제안
- 물리 세계의 정확한 표현은 신뢰할 수 있는 미래 상태 예측뿐 아니라 **정보에 기반한 의사결정(informed decision-making)**도 가능하게 함

전통적 접근 : 물리 법칙 기반 수학적 모델

- 물리 세계 모델링은 새로운 개념이 아님 — 전통적 자동화·로봇 산업은 물리 법칙과 system identification에 기반한 수학적 모델을 planning·control 알고리즘에 오랫동안 활용 (Murray et al., 2017)

- 그러나 이러한 **system-specific 모델은 저차원 state space에 국한** → 시스템 간 일반화·지식 전이가 제한되고, 새 과제·환경 적용 시 모델 재사용이 어려움

딥러닝·생성 AI의 전환

- 최근 딥러닝, 특히 **생성 AI(generative AI)**의 발전으로 **시각 관측(visual observations)으로부터 직접 world model을 학습**하는 것이 가능해짐

→ 저차원 물리 모델의 일반화 한계를 대규모 시각 데이터 학습으로 극복하는 방향 전환

Backbone 분류와 활용·검증 도메인

현대 world model 파이프라인은 backbone 아키텍처에 따라 분류 가능

RNN 기반 latent 모델

대부분의 기존 연구(Hafner et al., 2019·2021·2023 · Hansen et al., 2024 · Kim et al., 2020·2021)와 원 논문(Ha & Schmidhuber)은 autoencoder로 학습한 latent space에서 recurrent neural network로 시스템 상태 변화를 모델링

시각 관측 공간의 생성모델

최근 추세는 world model을 시각 관측 공간의 생성모델로 보는 것 — 흔히 conditional video 생성모델 형태 (action-to-video · text-to-video). Autoregressive(Bruce et al., 2024 · Liu et al., 2024 · Micheli et al., 2023 · Robine et al., 2023 · Yang et al., 2023) 또는 본 연구처럼 diffusion 기반(Alonso et al., 2024 · Ding et al., 2024 · Valevski et al., 2024)

Generative Simulation

또 다른 유망한 접근 — 생성 AI와 물리 시뮬레이터를 결합해 실제 세계를 모델링 (Hua et al., 2024 · Nasiriany et al., 2024)

활용 방식과 검증 도메인

- 잘 학습된 world model의 활용 : **verification**(Hu et al., 2023) · **planning 기반 MPC**(Bar et al., 2024 · Hansen et al., 2024) · **model-based RL**(Alonso · Ding · Robine · Yang · Zhang et al.)
- 효과가 입증된 도메인 : **computer games · real-world robots · autonomous driving** — 저자들은 foundational world model이 이 산업들에 transformative한 영향을 줄 것으로 전망

비디오 생성모델의 발전과 Cosmos의 양 계열 전략

초기의 짧은 저해상도 비디오에서, 생성 AI 연구의 최전선으로 급속히 발전 (Ho et al., 2022 · Huang et al., 2024)

- 최근 수년간 **Sora · Dream Machine · Gen 3 · Kling** 등 realistic·고해상도 비디오를 생성하는 인상적인 모델 등장 (OpenAI · Luma · Runway · KuaiShou, 2024) — 최초의 비디오 생성모델 공개 후 불과 몇 년 만의 성과

과제(task) 유형

Text-to-Video
대부분의 연구가 집중 — text prompt로 비디오 생성

Image-to-Video
주어진 image frame에서 시작하는 비디오 생성

Video-to-Video
reference 비디오로부터 새 비디오 생성

Action-to-Video
action 기반 생성 — world model·embodied AI 발전이 견인

Diffusion 프레임워크

- 대다수 비디오 생성모델이 채택 — noise를 점진적으로 비디오 시퀀스로 변환
- AR 모델이 유망함을 보였음에도, **시각 품질에서는 여전히 diffusion 기반이 우위**

Autoregressive 프레임워크

- 비디오와 다른 modality를 **통합된 방식(unified manner)으로 처리**할 수 있는 장점 (Deng · Kondratyuk · Liu et al.)
- LLM 커뮤니티가 개발한 온갖 기법을 활용 가능

3D 일관성 비디오 생성 연구의 계보

view synthesis·3D reconstruction의 초기 연구에서 camera controllable 비디오 생성까지

① Neural Rendering 시대

- 3D representation에 neural rendering을 적용해 3D 일관 비디오 생성 (Kerbl et al., 2023 · Mildenhall et al., 2020 · Zhou et al., 2018 등)
- single-image 3D view synthesis는 특히 어려움 — multi-view 데이터셋에서 강한 3D prior 모델 학습이 필요하나 3D prior는 확장성(scale)이 낮음

→

② 순수 Data-driven 접근

- scalable Transformer 아키텍처(Dosovitskiy et al., 2021 · Vaswani et al., 2017)로 explicit 3D prior 없이 학습 기반 view synthesis 탐구 (Rombach et al., 2021 · Sajjadi et al., 2022)
- camera 입력을 조건으로 neural network가 novel view를 직접 합성 (Tatarchenko et al., 2016)

→

③ Diffusion 스케일업

- 이 패러다임은 diffusion 모델로 성공적으로 확장 (Liu et al., 2023) → 3D asset generation에 광범위하게 응용 (Poole et al., 2023 · Lin et al., 2023 · NVIDIA, 2024 등)
- 최근 비디오 생성 품질의 발전은 학습 비디오 데이터 스케일링으로 완전한 3D 일관성 달성 가능성을 시사 (Brooks et al., 2024)

➢ **이러한 모델의 camera controllability는 robotics·autonomous navigation 응용 잠재력으로 인해 활발한 연구 영역 (He et al., 2024 · Wang et al., 2024 · Xu et al., 2024)**

로봇 제어를 위한 생성모델의 세 흐름

딥 생성모델의 발전이 robotic control 응용에 대한 관심을 크게 촉발

Diffusion 기반 Visuomotor Policy

- diffusion 모델을 visuomotor policy로 직접 활용하는 연구 흐름
- 다양한 robotic task의 imitation learning에서 상당한 성능 향상을 입증 (Chi et al., 2023 · Ke et al., 2024 · Prasad et al., 2024 · Wang et al., 2024)

Generative Motion Planning

- 본 연구와 더 관련된 두 흐름 중 하나
- explicit action sequence 대신 중간 시각 sub-goal을 생성해 unseen 환경에 대한 일반화를 강화 (Black et al., 2023 · Du et al., 2024 · Finn & Levine, 2017 · Ko et al., 2024 · Zhou et al., 2024)
- **image·video sub-goal은 다양한 환경 setup에 일반화되는 반면, action sequence는 환경·과제 특화적 — 시각 표현 전략이 더 robust**

Generative Pre-training

- 대규모 image·video 데이터셋을 pre-training에 활용하는 흐름 (Cheang et al., 2024 · Gupta et al., 2024 · He et al., 2024)
- Gupta et al.(2024) : pre-trained text-to-image diffusion 모델의 feature를 추출해 이후 policy learning을 가이드
- **Cheang et al.(2024)·He et al.(2024) : 2단계 — 미래 frame 예측 pre-training 후, action과 미래 frame을 공동 예측하도록 fine-tune**

자율주행 시뮬레이션을 위한 비디오 생성모델

다양한 입력 modality를 조건으로 realistic 주행 비디오를 생성 — AV 시뮬레이션을 혁신할 잠재력

Diffusion 기반 Visuomotor Policy

Text | Images | Trajectories | 3D Data | Maps

기존 접근의 한계 (논문 지적)

잠재력에도 불구하고, 기존 접근들은 다음 제약으로 포괄적인 driving world simulator로서의 효과가 제한됨 :

- **데이터 규모(data scale)의 제약** (Gao et al. · Jia et al. · Lu et al. · Wang et al.)
- **해상도(resolution)의 제약** (Hu et al., 2023 · Yang et al., 2024)
- **카메라 view 수의 제약** (Blattmann et al., 2023 · Gao et al., 2024)

Cosmos의 대응 (논문 서술)

- 이러한 한계를 극복하기 위해, **강력한 pre-trained WFM의 능력을 활용**해 유연하고 확장 가능한 driving simulator를 개발
- 본 모델이 달성한 것 :
  고해상도 (high resolution)
  높은 frame rate (elevated frame rates)
  multi-view 일관성 (multi-view consistency)

Visual Tokenizer 연구의 흐름과 Cosmos Tokenizer의 기반

입력 시각 데이터를 재현하는 latent feature 학습은 오랜 역사 (Hinton et al., 1995 · Kingma, 2013 · van den Oord et al., 2017 · He et al., 2022)

- 최근에는 대규모 생성모델의 학습 효율을 높이는 필수 구성요소로 널리 채택 (Esser et al., 2021 · Rombach et al., 2022)

Continuous Visual Tokenizer (AE · VAE)

- 시각 데이터를 continuous latent space로 압축 → 그 위에서 diffusion 기반 모델을 효율적으로 학습 (Ho et al., 2020 · Lipman et al., 2022 · Song et al., 2020)
- 추론 시 생성된 latent를 tokenizer decoder가 RGB 공간으로 복원
- 다양한 diffusion 모델이 이 방식으로 image·video 생성에 학습됨 (FLUX · Rombach · Blattmann · Brooks et al. 등)

Discrete Visual Tokenizer (Quantizer 포함)

- continuous latent를 discrete space로 추가 양자화하는 quantizer 포함 (Lee et al., 2022 · Mentzer et al., 2023 · van den Oord et al., 2017 · Yu et al., 2024 등)
- text·audio 등 다른 modality와 함께 **LLM·VLM에 쉽게 통합 가능**
- visual understanding과 image·video 생성 과제에 폭넓게 배치됨 (Esser · Chang · Kondratyuk · Villegas et al. 등)

➢ **Cosmos tokenizer는 FSQ(Mentzer et al., 2023) · causal 구조(Yu et al., 2023) 등 선행 연구 위에 광범위하게 구축 — 목표는 효율적이고 고품질인 tokenizer suite의 구축**