Paper Review (arXiv 2025.07.09 - citation 139↑)

Cosmos World Foundation Model Platform for Physical AI

Susang Kim

https://arxiv.org/pdf/2501.03575

Contents

1.Introduction

2.Related Works

3.Cosmos (World Foundation Model)

1) Data Curation

2) Tokenizer

3) Architecture (Diffusion vs. Autoregressive)

4) World Foundation Model Pre-training

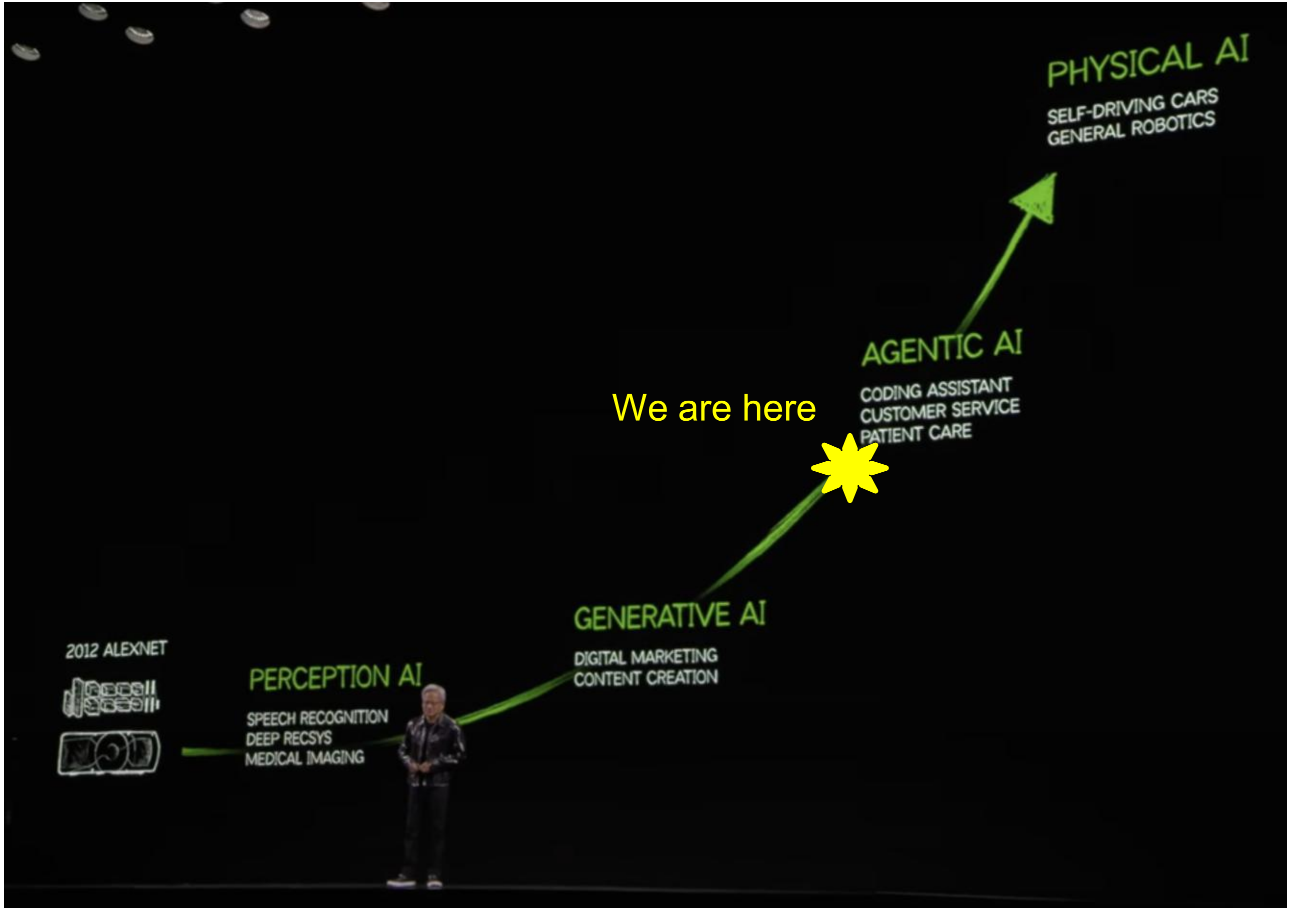
5) World Foundation Model Post-training

6) Guardrail

4. Conclusion

5. Cosmos World Foundation Models

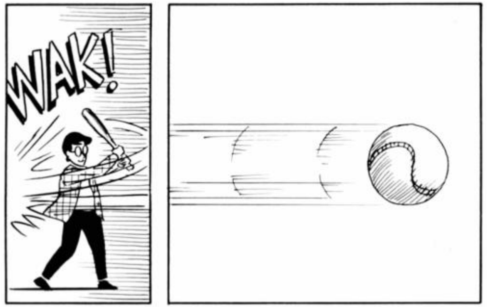
1.Introduction - NVIDIA CEO Jensen Huang Keynote at CES 2025



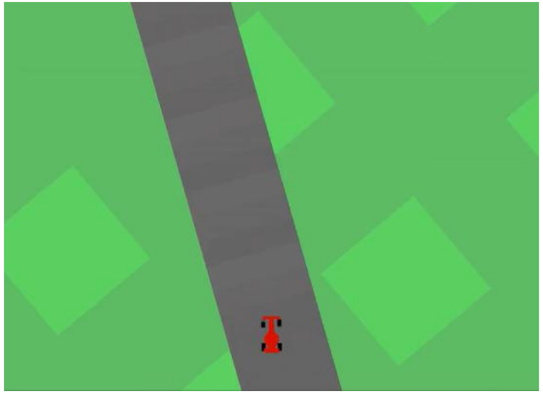
https://www.youtube.com/live/k82RwXqZHY8?si=19rRQC7xF0jg1nNS

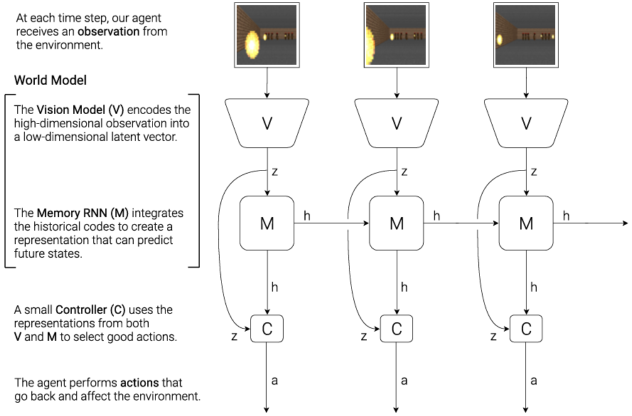
1.Introduction – World Model

2018–2019: **RNN-based latent world models**, such as PlaNet, laid the foundation for learning compact representations of the environment.
2020–2023: **The integration of reinforcement learning**, exemplified by the Dreamer series, and the spread of autoregressive prediction approaches significantly advanced world model capabilities.
2023–2024: **Diffusion-based methods, multi-modal world models**, and the practical deployment of **Sim2Real(ControlNet)** marked a new era of real-world applicability.



We learn to perceive time *spatially* when we read comics. According to cartoonist and comics theorist Scott McCloud, "*in the world of comics, time and space are one and the same.*" Art © Scott McCloud. [1]





At each time step, our agent receives an **observation** from the environment.

World Model

The **Vision Model (V)** encodes the high-dimensional observation into a low-dimensional latent vector.

The **Memory RNN (M)** integrates the historical codes to create a representation that can predict future states.

A small **Controller (C)** uses the representations from both **V** and **M** to select good actions.

The agent performs **actions** that go back and affect the environment.

Figure 4. Our agent consists of three components that work closely together: **Vision (V)**, **Memory (M)**, and **Controller (C)**

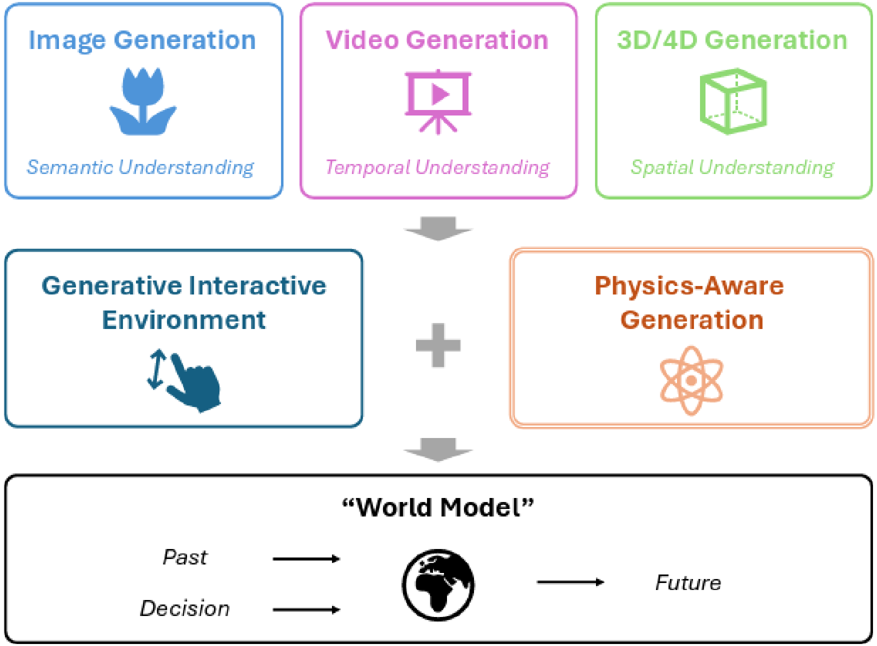
Functions of a World Model

- **Prediction**: Anticipates what is likely to happen next.
- **Planning and Exploration**: Enables the agent to think ahead and decide on future actions.
- **Inference**: Allows the agent to infer the overall situation even when some information is missing.
- **Risk Avoidance**: Helps avoid danger even in unfamiliar or novel situations.
- **Learning Enhancement**: Turns risky situations into opportunities to refine the internal model of the world.

Ha, David, and Jürgen Schmidhuber. "World models." NeurIPS 2018.

1.Introduction – World Model

Generative Interactive Environments: Allow models to respond to and adapt based on user interactions.
Physics-Aware Generation: Enables simulations that respect real-world physical laws and dynamics.
Together, these advancements contribute to building a comprehensive **"World Model"**, which **integrates past information** and **decisions to simulate and predict future outcomes** paving the way for general-purpose simulation models applicable across many domains.



Physical Materials – Define different types of entities with specific assumptions and constraints (e.g., rigid bodies, fluids).

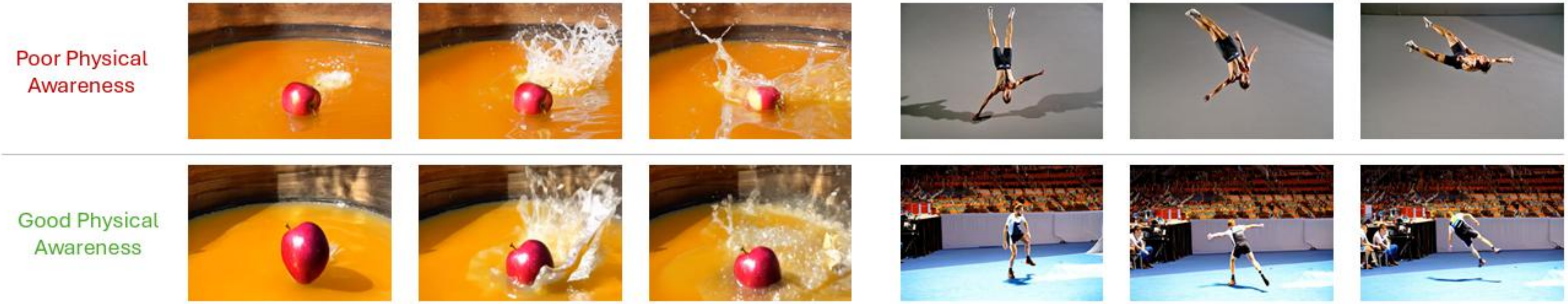
Physical Simulation – Provide computational tools to model the dynamics of these materials under physical laws. (Issac sim)

Physical Engines – Serve as practical, ready-to-use platforms that implement these simulations. (Blender, Issac Gym, Genesis, NVIDIA Physics)

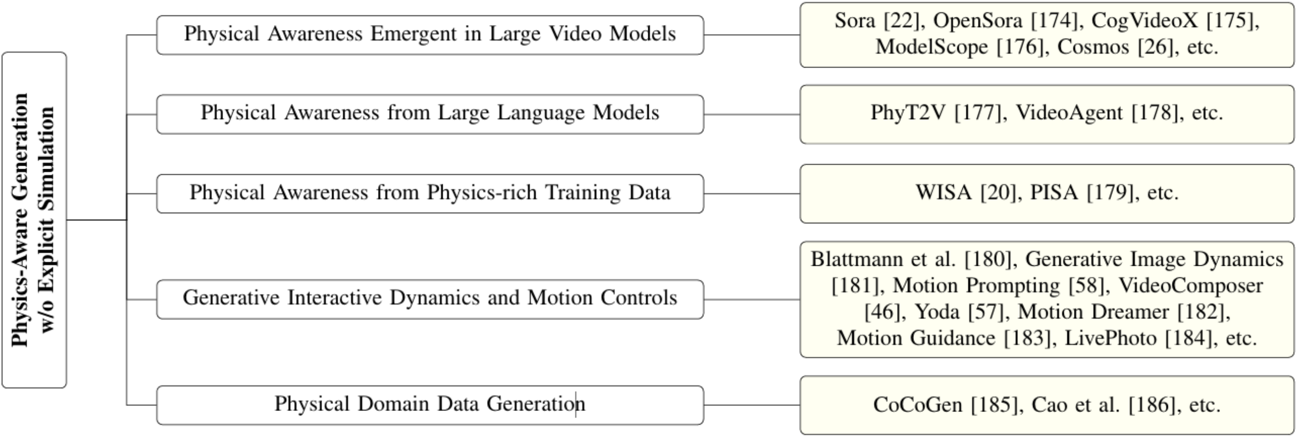
Liu, Daochang, et al. "Generative physical ai in vision: A survey." arXiv 2025.

1.Introduction – Generative Physical AI

Video plays a crucial role in generative AI for computer vision because it captures real-world information. As an implicit physical model, video helps AI understand and simulate complex real-world phenomena, enabling applications like autonomous driving, robotics, scientific simulation, and other embodied intelligence tasks.



The **Cosmos** is a recent open-source toolkit that provides a video data pipeline, tokenizers, and both pre-trained and post-trained models. It includes transformer-based diffusion and autoregressive models trained on large-scale video datasets, which can be fine-tuned **for physically grounded tasks like robotic manipulation, camera control, and autonomous driving**.



Physics-Aware Generation w/o Explicit Simulation

- Physical Awareness Emergent in Large Video Models — Sora [22], OpenSora [174], CogVideoX [175], ModelScope [176], Cosmos [26], etc.
- Physical Awareness from Large Language Models — PhyT2V [177], VideoAgent [178], etc.
- Physical Awareness from Physics-rich Training Data — WISA [20], PISA [179], etc.
- Generative Interactive Dynamics and Motion Controls — Blattmann et al. [180], Generative Image Dynamics [181], Motion Prompting [58], VideoComposer [46], Yoda [57], Motion Dreamer [182], Motion Guidance [183], LivePhoto [184], etc.
- Physical Domain Data Generation — CoCoGen [185], Cao et al. [186], etc.

Liu, Daochang, et al. "Generative physical ai in vision: A survey." arXiv preprint arXiv 2025.

2.Related Works - Video as the New Language for Real-World Decision Making (ICML 2024)

While both text and videos are abundant online, language models profoundly impact real-world applications, **whereas video generation primarily remains limited to entertainment**.

Videos inherently capture complex, physically grounded information about the real world, often beyond textual descriptions.

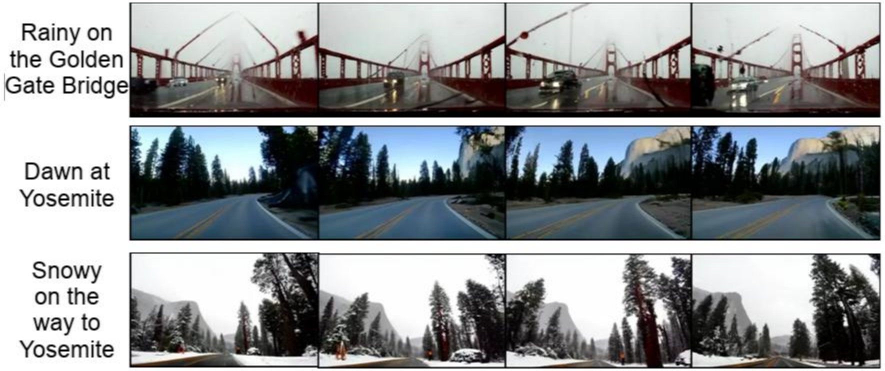
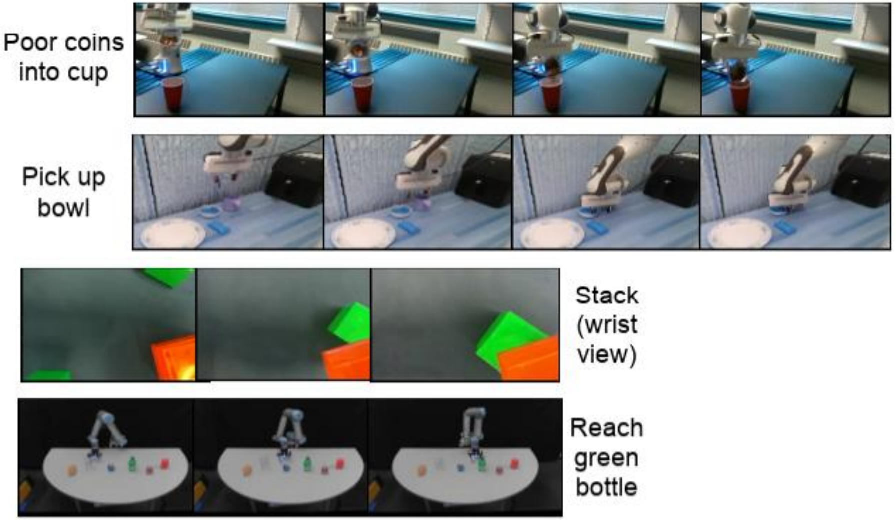


Figure 9: **Generative Simulation for Self-Driving.** With internet knowledge, we can simulate different driving conditions at particular locations, such as "rain on Golden Gate Bridge" (top), "dawn in Yosemite" (middle), and "snow on the way to Yosemite" (bottom).



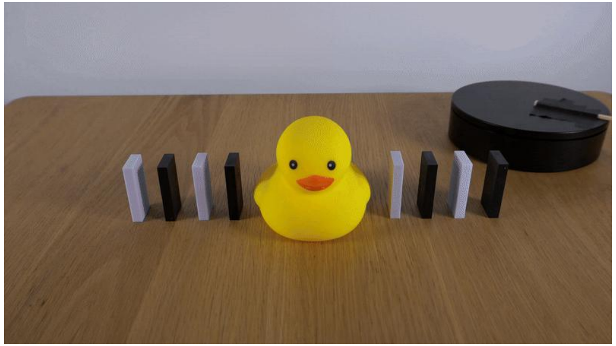
This paper advocates **treating video generation as a practical language of physical reality, enabling knowledge transfer, planning, and sophisticated decision-making**.

Yang, Sherry, et al. "Video as the new language for real-world decision making." ICML 2024.

2.Related Works - Physics IQ Benchmark arXiv (2025.02.27) - DeepMind

Do generative video models understand physical principles?

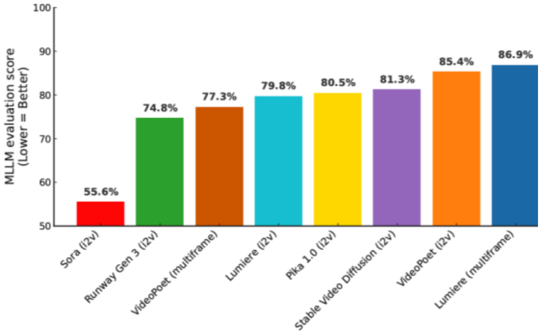
We develop the Physics-IQ benchmark and score, which reveals that current generative video models lack **physical understanding** despite sometimes achieving visual realism. Use our benchmark and dataset to assess your video model's physics understanding!

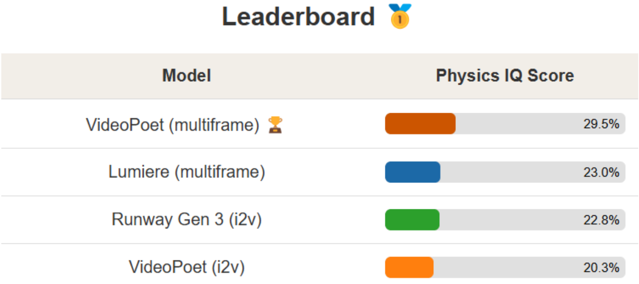


Solid mechanics

Fluid dynamics

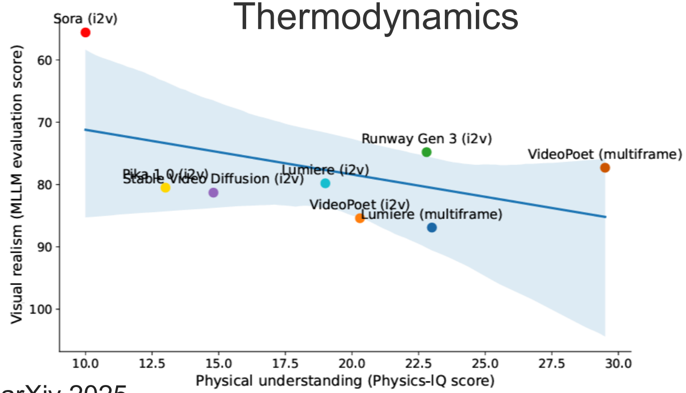
Thermodynamics





Leaderboard 🥇

Model	Physics IQ Score
VideoPoet (multiframe) 🏆	29.5%
Lumiere (multiframe)	23.0%
Runway Gen 3 (i2v)	22.8%
VideoPoet (i2v)	20.3%



Motamed, Saman, et al. "Do generative video models learn physical principles from watching videos?." arXiv 2025.

2.Related Works - Physics IQ Benchmark arXiv (2025.02.27) - DeepMind

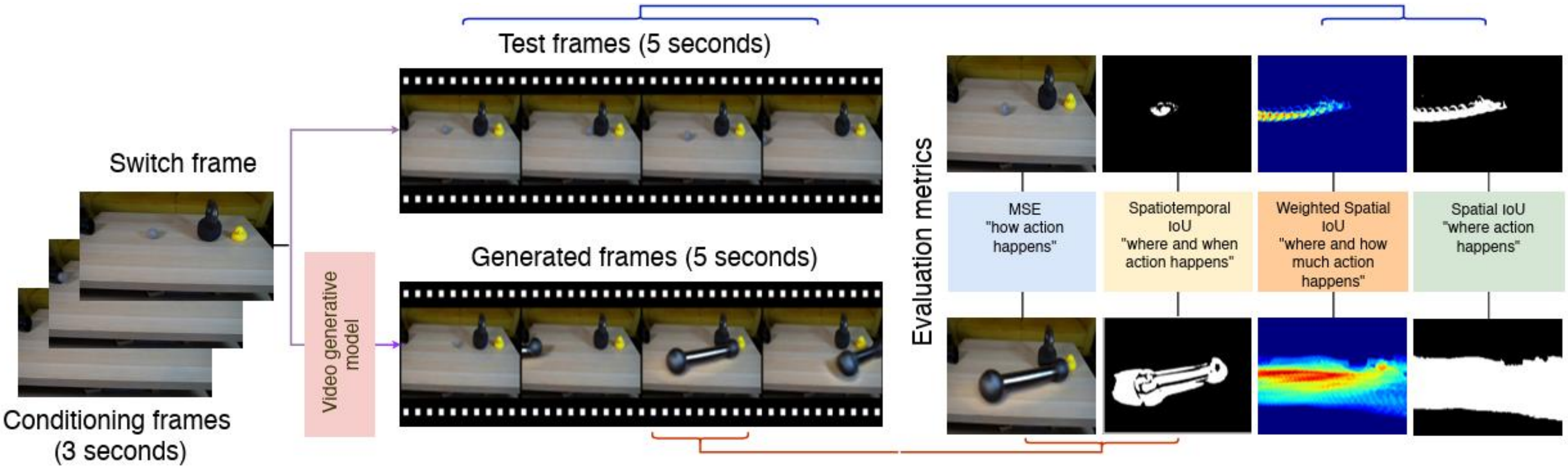

Test frames (5 seconds)
Switch frame
Generated frames (5 seconds)
Video generative model
Conditioning frames (3 seconds)
Evaluation metrics
MSE "how action happens"
Spatiotemporal IoU "where and when action happens"
Weighted Spatial IoU "where and how much action happens"
Spatial IoU "where action happens"

Fig. 2. Overview of the Physics-IQ evaluation protocol. A video generative model produces a 5 second continuation of the conditioning frame(s), optionally including a textual description of the conditioning frames for models that accept text input. They are compared against the ground truth test frames using four metrics that quantify different properties of physical understanding. The metrics are defined and explained in the methods section. Code to run the evaluation is available at Physics-IQ-benchmark.

- Where does action happen? **Spatial IoU**
- Where & when does action happen? **Spatiotemporal IoU**
- Where & how much action happens? **Weighted spatial IoU**
- How does action happen? **MSE**

$$\text{Weighted-spatial-IoU} = \frac{\sum_{i=1}^{n} \min\left(M_{\text{real,i}}^{\text{weighted,spatial}}, M_{\text{gen,i}}^{\text{weighted,spatial}}\right)}{\sum_{i=1}^{n} \max\left(M_{\text{real,i}}^{\text{weighted,spatial}}, M_{\text{gen,i}}^{\text{weighted,spatial}}\right)}$$

$$\text{Spatiotemporal-IoU}(M_{\text{real}}, M_{\text{gen}}) = \frac{|M_{\text{real}} \cap M_{\text{gen}}|}{|M_{\text{real}} \cup M_{\text{gen}}|}$$

$$\text{MSE}(f_{\text{real}}, f_{\text{gen}}) = \frac{1}{n}\sum_{i=1}^{n}(f_{\text{real,i}} - f_{\text{gen,i}})^2$$

2.Related Works - GEN3C (CVPR 2025) - Nvidia

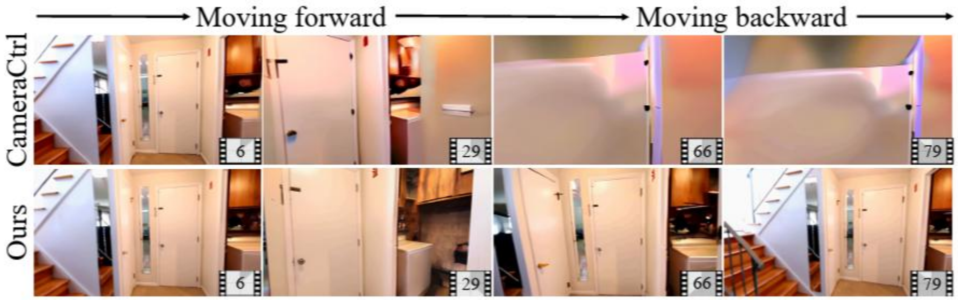


Figure 2. **Motivation:** Our model can generate consistent videos when the camera covers the same region multiple times, while previous work produces severe artifacts due to the lack of explicit modeling of the history.

GEN3C is a generative video model that ensures **precise camera control and temporal 3D consistency** by leveraging a 3D cache of point clouds derived from depth predictions. Unlike prior models that infer structure from camera inputs, GEN3C uses 2D renderings of this 3D cache for each new frame, allowing it to focus on generating new content rather than reconstructing past frames.

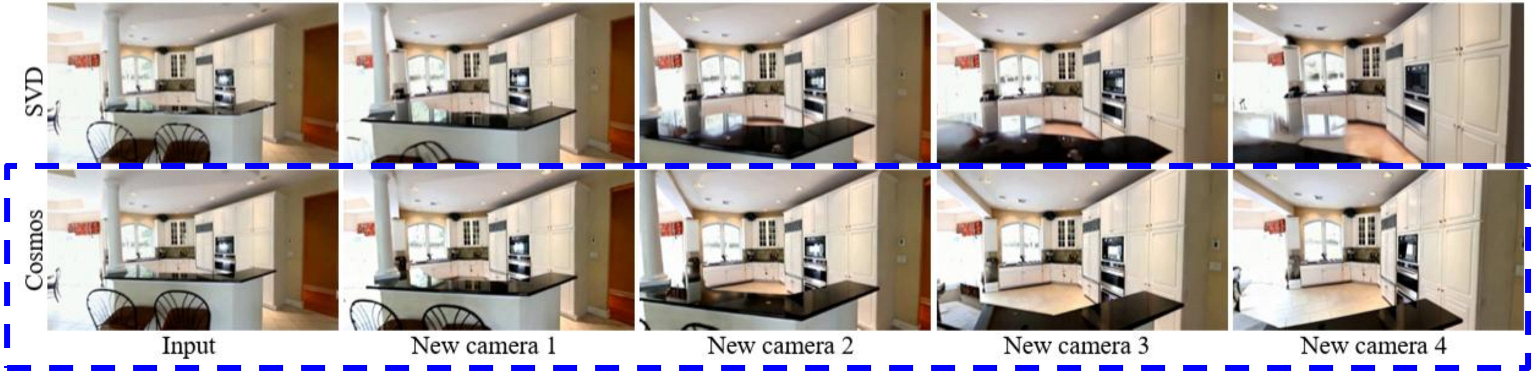


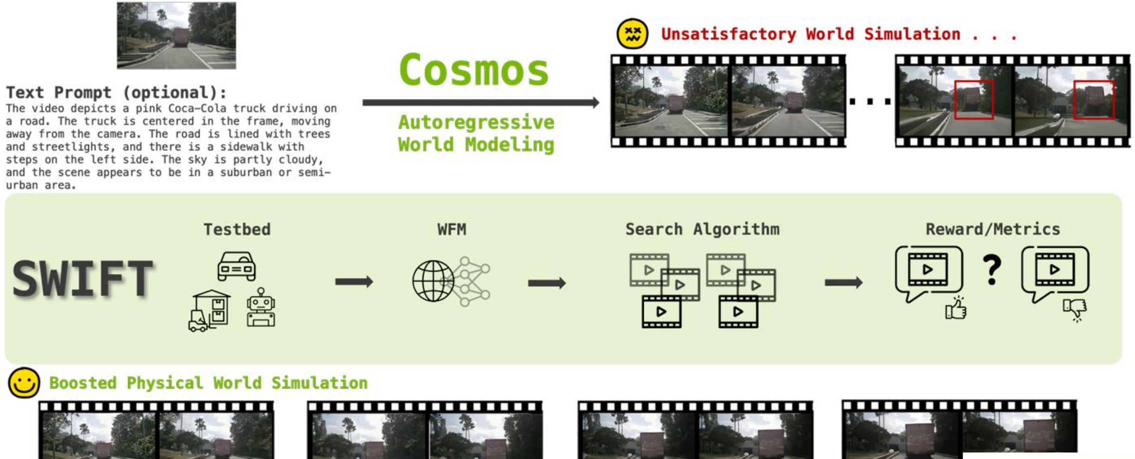
Figure 11. Qualitative comparison on using different base models: Stable Video Diffusion (SVD) [4] v.s. Cosmos [1]. When having a more powerful video generation model, GEN3C is able to generate more realistic output with less artifacts. Note that the slight misalignment between the two results is due to the models using different video resolutions.

Ren, Xuanchi, et al. "Gen3c: 3d-informed world-consistent video generation with precise camera control." CVPR 2025.

2.Related Works - SWIFT (arXiv 2025.03.31) - Nvidia

SWIFT: Can Test-Time Scaling Improve World Foundation Model?

The first test-time scaling framework specifically designed for world foundation models (WFMs). Addressing the high computational cost and data limitations of training and scaling WFMs, SWIFT offers an efficient alternative by reallocating computation during inference.



Text Prompt (optional):
The video depicts a pink Coca-Cola truck driving on a road. The truck is centered in the frame, moving away from the camera. The road is lined with trees and streetlights, and there is a sidewalk with steps on the left side. The sky is partly cloudy, and the scene appears to be in a suburban or semi-urban area.

Cosmos
Autoregressive World Modeling
Unsatisfactory World Simulation . . .
SWIFT
Testbed
WFM
Search Algorithm
Reward/Metrics
?
Boosted Physical World Simulation

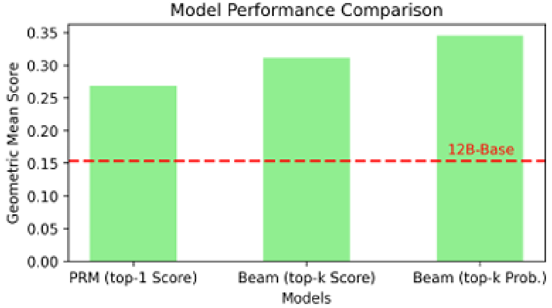
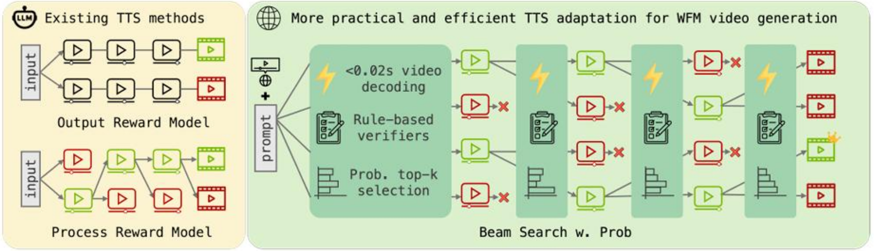


Figure 6: Improvement based on COSMOS-4B using different search algorithms. Beam search with probability boosts the performance most.

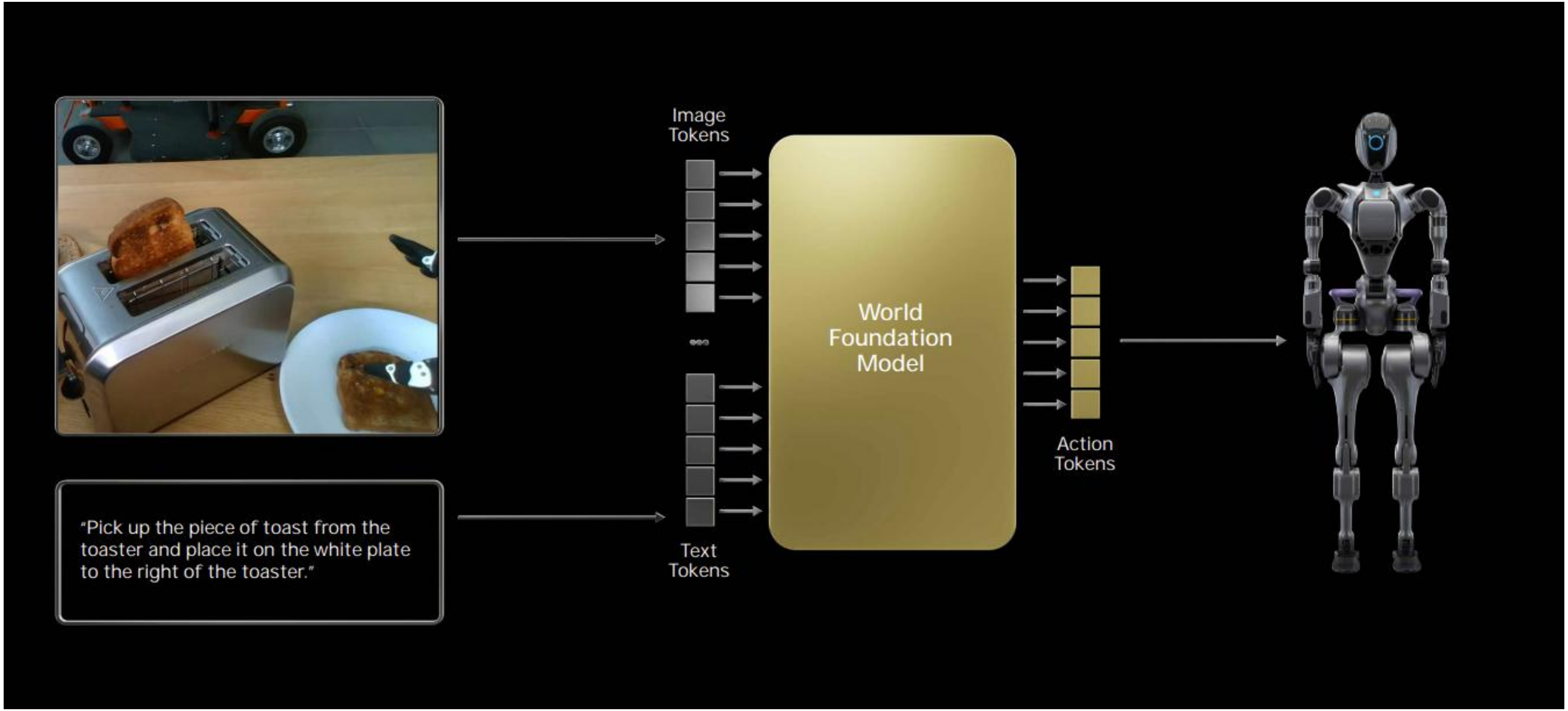
The First Test-Time Scaling for World Foundation Models.



Cong, Wenyan, et al. "Can Test-Time Scaling Improve World Foundation Model?." arXiv 2025.

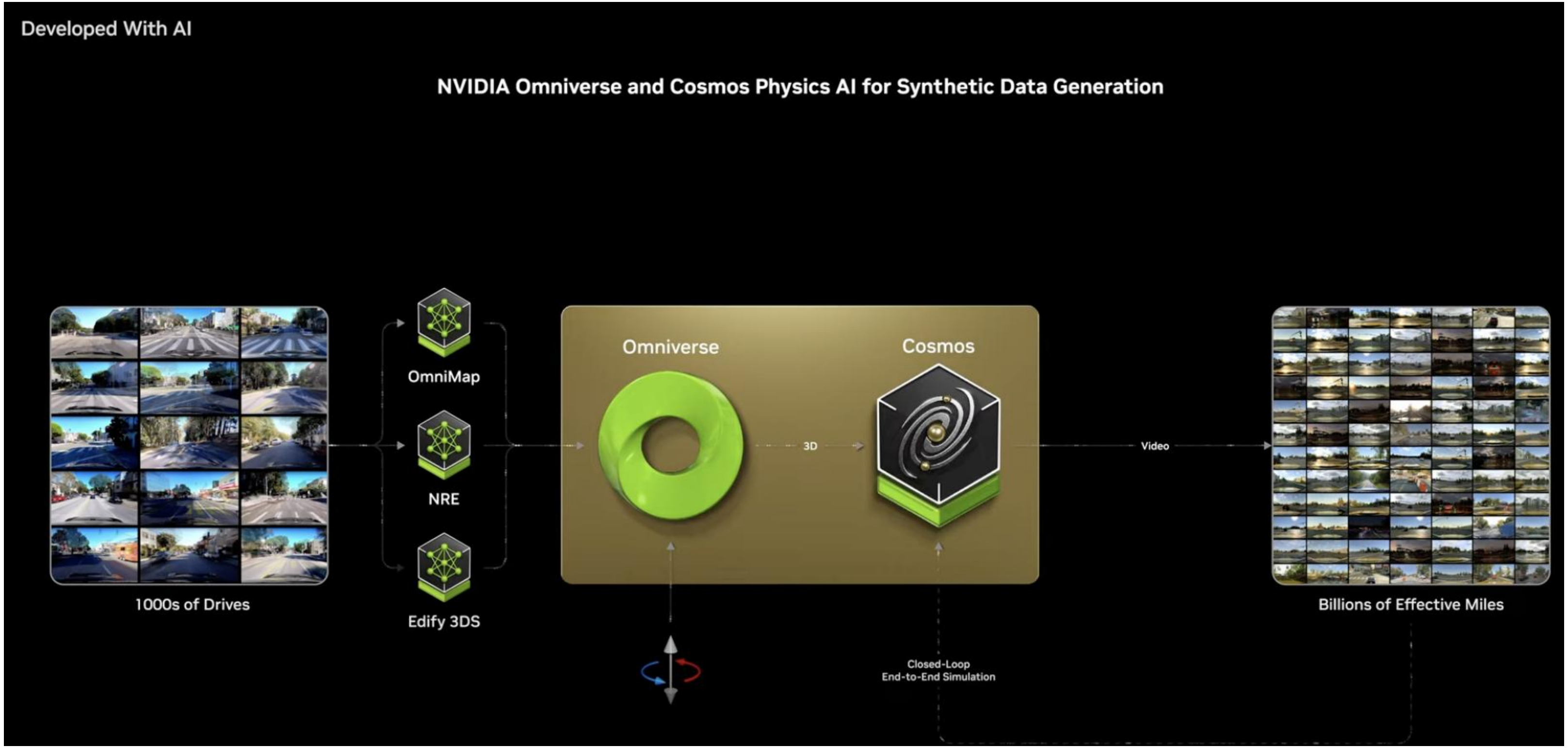
3.Cosmos(World Foundation Model)

A state-of-the-art generative World Foundation Model Platform that produces videos based on physical laws, dynamics, and properties. By integrating with NVIDIA Omniverse, a platform for creating 3D virtual environments, it enables the generation of high-quality synthetic data closely resembling the real world.

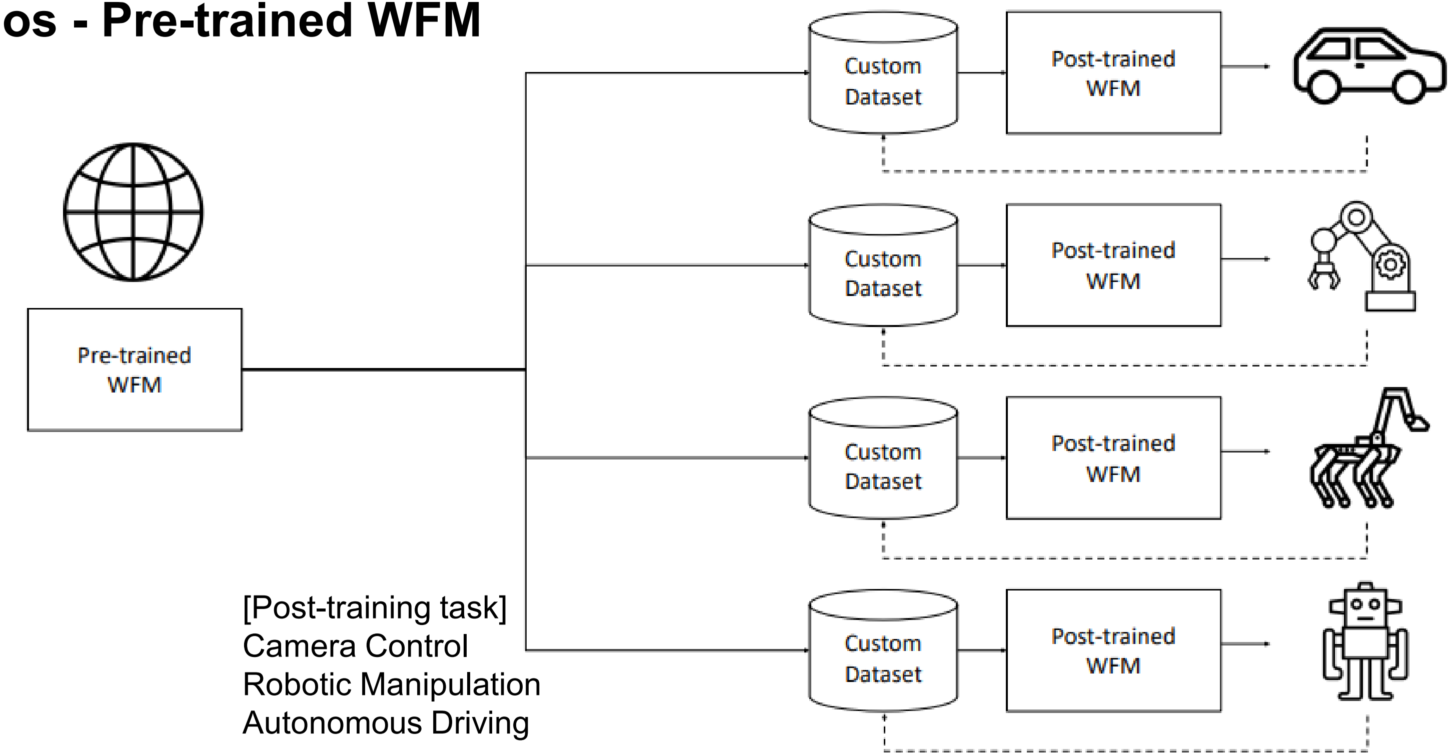

"Pick up the piece of toast from the toaster and place it on the white plate to the right of the toaster."
Image Tokens
Text Tokens
World Foundation Model
Action Tokens

3.Cosmos(World Foundation Model)

Physical AI refers to AI systems interacting with the physical world through sensors and actuators.
Real-world training is risky and inefficient, hence the need for training in a digital environment first.
To enable this, a World Foundation Model (WFM) is proposed, acting as a digital twin simulating the real world, built using the Cosmos platform.



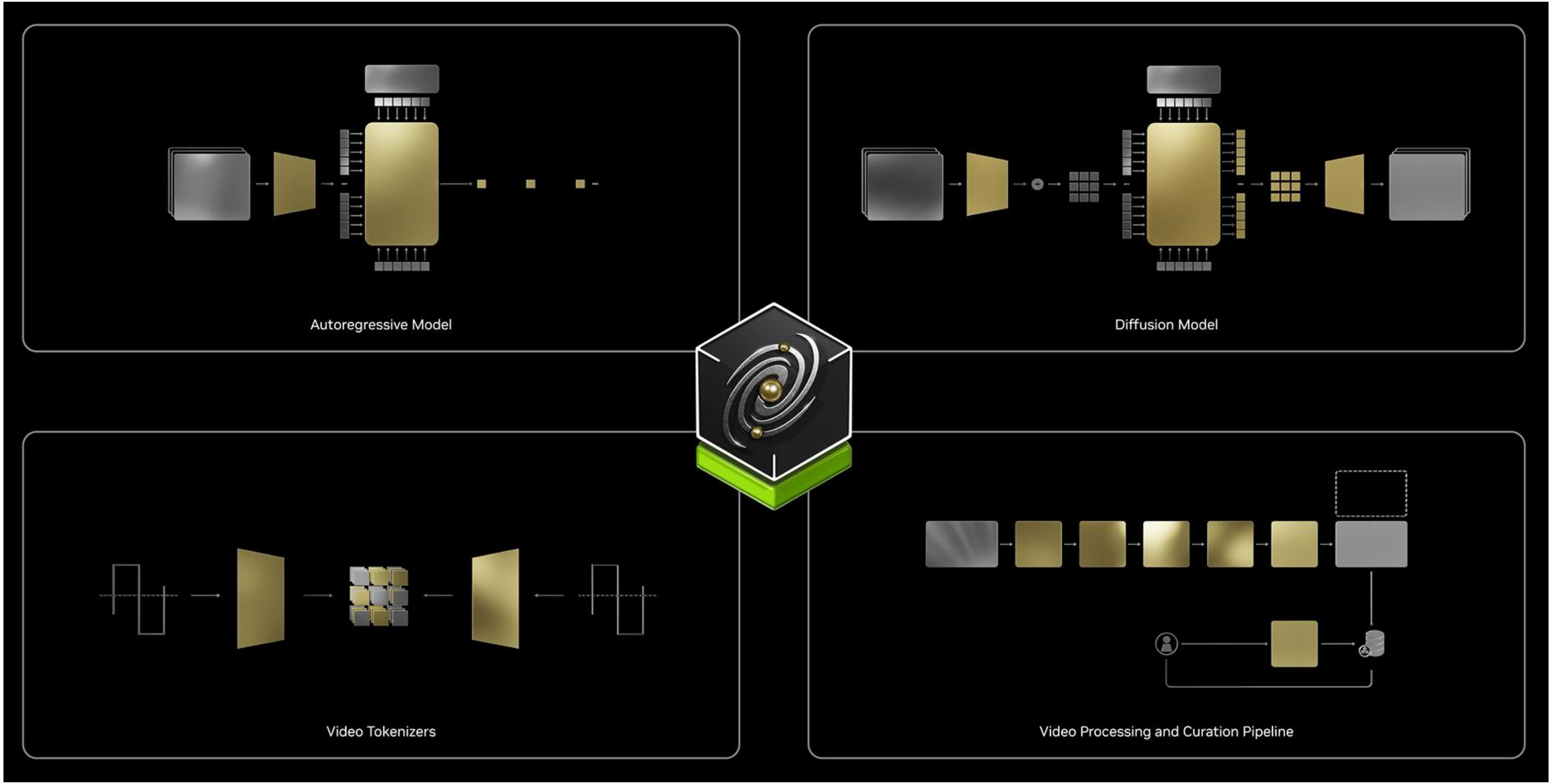
3.Cosmos - Pre-trained WFM



[Post-training task]
Camera Control
Robotic Manipulation
Autonomous Driving

The figure illustrates how pre-trained **World Foundation Models (WFMs)** - trained on large, diverse video datasets capturing real-world physics - serve as generalist models. These **WFMs can be efficiently adapted to specific Physical AI tasks through post-training** using smaller datasets of "prompt"-video pairs (e.g., commands, instructions, or trajectories). The dashed lines in the figure represent the data flow involved in this pre-training and post-training process.

3.Cosmos



https://developer.nvidia.com/blog/advancing-physical-ai-with-nvidia-cosmos-world-foundation-model-platform/

3.Cosmos - World Foundation Model Platform

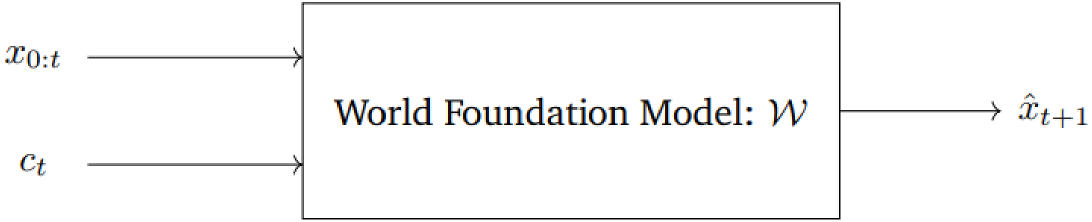
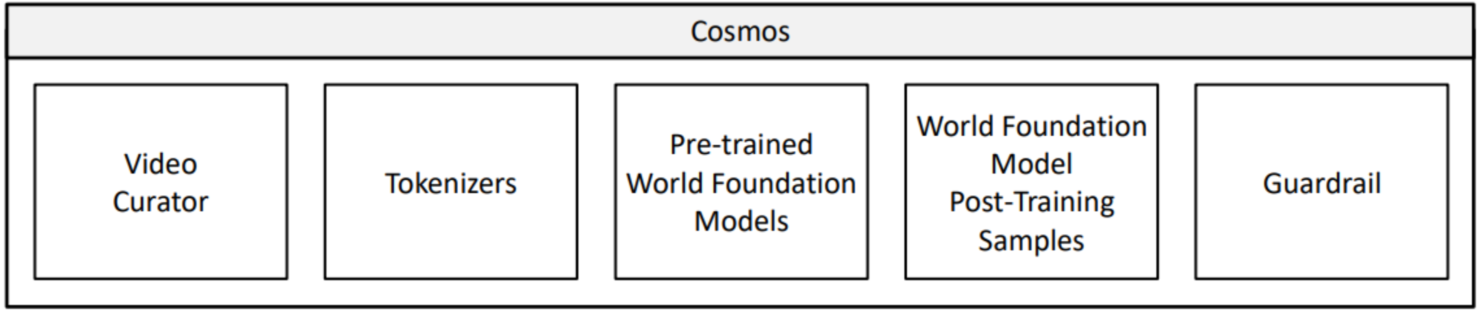


Figure 3: A world foundation model (WFM) $\mathcal{W}$ is a model that generates the future state of the world x_{t+1} based on the past observations $x_{0:t}$ and current perturbation c_t.

key uses of a **World Foundation Model (WFM)** in **Physical AI**:
1.Policy Evaluation: WFM allows testing and filtering of policy models in simulated environments, saving real-world costs and time.
2.Policy Initialization: WFMs help initialize policy models by providing learned world dynamics, addressing data scarcity issues.
3.Policy Training: WFMs, combined with reward models, enable reinforcement learning without needing direct real-world interaction.
4.Planning & Model-Predictive Control: WFMs simulate outcomes of different actions, aiding in decision-making and control by selecting the best action sequence.
5.Synthetic Data Generation: WFMs can generate labeled synthetic data for training and be fine-tuned for Sim2Real tasks using metadata like depth or semantics.

2.Cosmos consists of several major components



Cosmos

Video Curator

Tokenizers

Pre-trained World Foundation Models

World Foundation Model Post-Training Samples

Guardrail

Figure 4: Cosmos World Foundation Model Platform consists of several major components: video curator, video tokenizer, pre-trained world foundation model, world foundation model post-training samples, and guardrail.

Video curator : Select and prepare high-quality video clips for model training.

Video tokenization : causal tokenizers make the model more versatile and better suited for real-world, time-forward scenarios.

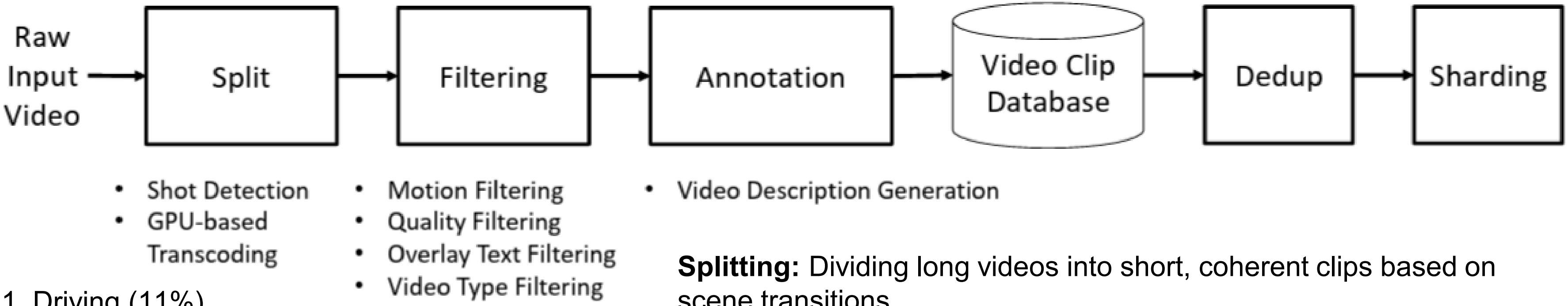
WFM pre-training : Train a general-purpose world model that understands physical dynamics.

World model post-training : Fine-tune the general WFM for specific Physical AI tasks.

Guardrail : Ensure the safe and responsible use of WFMs.

3.Cosmos - Data Curation

The team processes **20 million hours of raw video** (720p–4K, HD~UHD), much of which is redundant or unhelpful for learning physical concepts. To address this, they developed a modular data curation pipeline that selects the most valuable video segments and also incorporates image data to improve video generation quality and training speed. Using this pipeline, they generate approximately 100 million video clips for pre-training and **10 million for fine-tuning**.



1. Driving (11%),
2. Hand motion and object manipulation (16%),
3. Human motion and activity (10%),
4. Spatial awareness and navigation (16%),
5. First person point-of-view (8%),
6. Nature dynamics (20%),
7. Dynamic camera movements (8%),
8. Synthetically rendered (4%), and
9. Others (7%).

Splitting: Dividing long videos into short, coherent clips based on scene transitions.
Filtering: Removing static or low-quality videos, retaining only dynamic content beneficial for physical learning.
Annotation: Automatically generating textual descriptions for each video clip using a Vision-Language Model (VLM).
Deduplication: Eliminating semantically duplicated videos to enhance data diversity.
Sharding: Classifying clips by resolution and format to optimize data efficiency.

3.Cosmos - Tokenizer

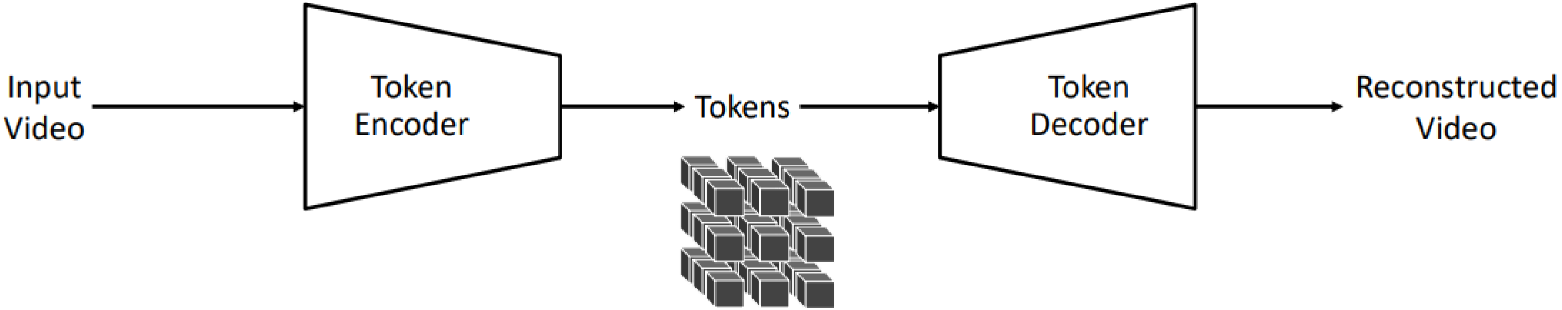
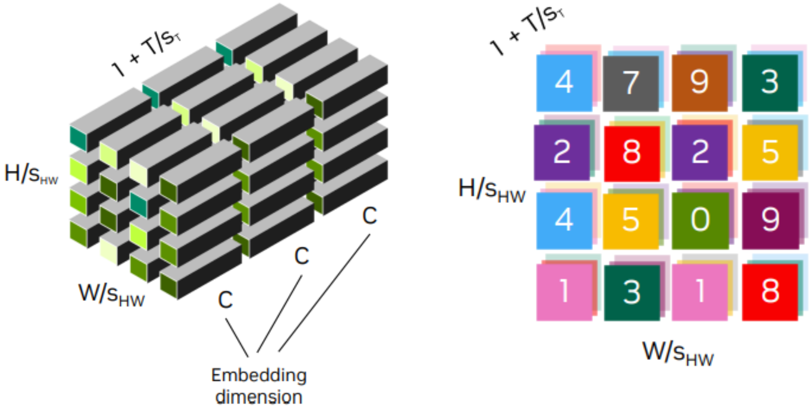


Figure 6: **Video tokenization pipeline.** An input video is encoded into tokens, which are usually much more compact than the input video. The decoder then reconstructs the input video from the tokens. Tokenizer training is about learning the encoder and decoder to maximally preserve the visual information in the tokens.



(a) Continuous tokens

(b) Discrete tokens

Since raw video data is challenging to process directly, it is converted into compressed **tokens**.

Two types of tokens are provided:

Continuous: Tokens represented as continuous vectors (suitable for **diffusion models**).

Discrete: Tokens represented as integer indices (suitable **for autoregressive models**).

These tokens serve as the fundamental input format for the WFM, enabling realistic simulations.

3.Cosmos - Tokenizer

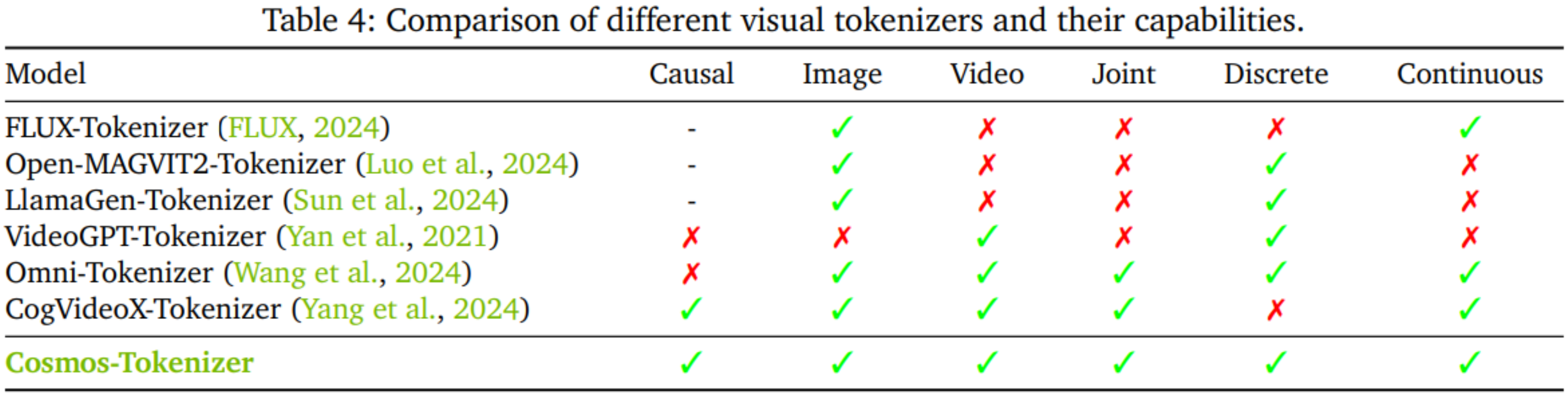


Table 4: Comparison of different visual tokenizers and their capabilities.

Model	Causal	Image	Video	Joint	Discrete	Continuous
FLUX-Tokenizer (FLUX, 2024)	-	✓	✗	✗	✗	✓
Open-MAGVIT2-Tokenizer (Luo et al., 2024)	-	✓	✗	✗	✓	✗
LlamaGen-Tokenizer (Sun et al., 2024)	-	✓	✗	✗	✓	✗
VideoGPT-Tokenizer (Yan et al., 2021)	✗	✗	✓	✗	✓	✗
Omni-Tokenizer (Wang et al., 2024)	✗	✓	✓	✓	✓	✓
CogVideoX-Tokenizer (Yang et al., 2024)	✓	✓	✓	✓	✗	✓
Cosmos-Tokenizer	✓	✓	✓	✓	✓	✓

Causal: Whether the model operates in a causal manner (i.e., generating tokens using **only past data**).

Image: Whether the model can process image data.

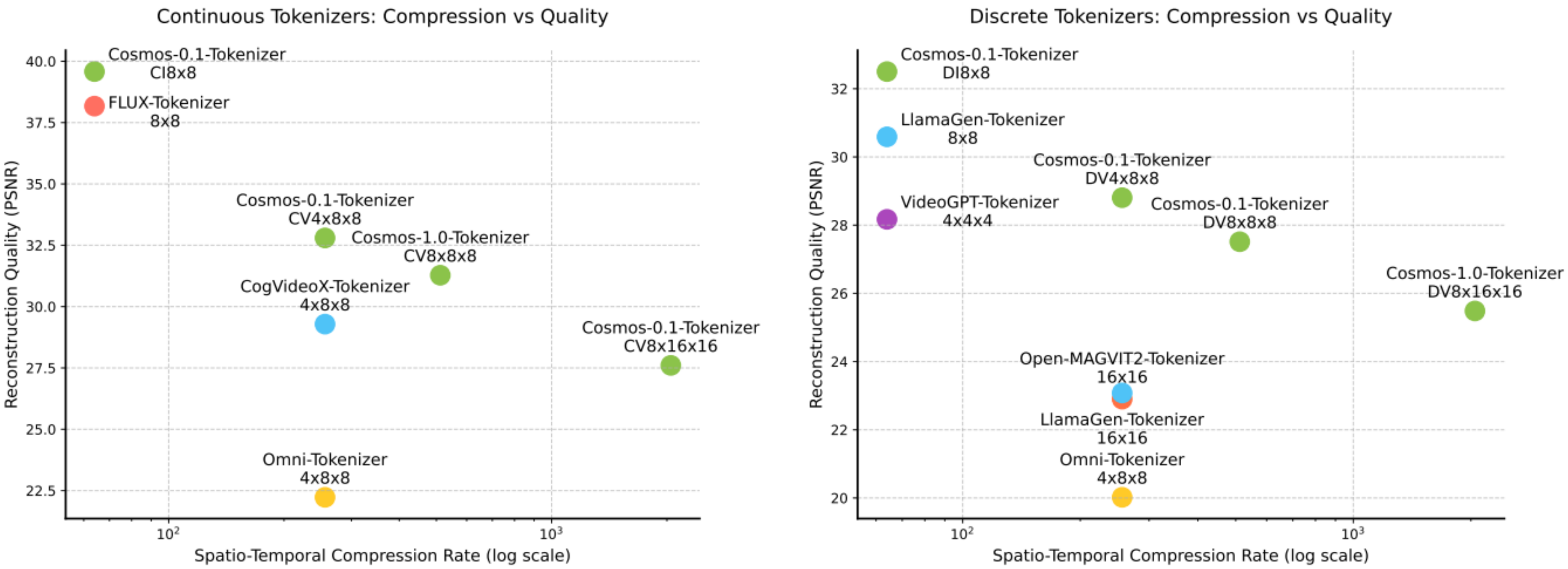
Video: Whether the model can process video data.

Joint: Whether the model can simultaneously handle both image and video data.

Discrete: Whether tokens are generated in a discrete form.

Continuous: Whether tokens are generated in a continuous form.

3.Cosmos - Tokenizer



(a) Continuous tokenizers

(b) Discrete tokenizers

The Cosmos-Tokenizer series visually demonstrates superior performance compared to existing tokenizers in terms of quality relative to compression ratio. In other words, Cosmos tokenizers deliver robust performance, maintaining high video reconstruction quality even at relatively high compression rates.

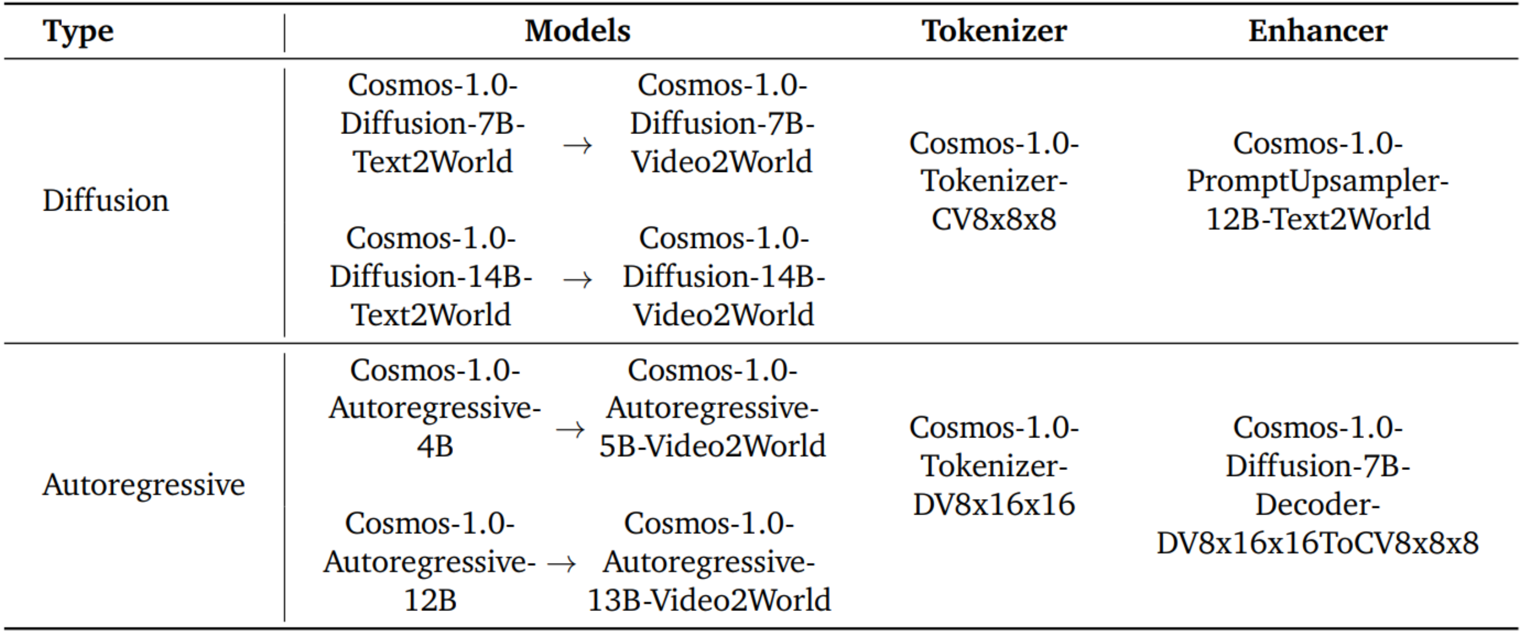
3.Cosmos - World Foundation Model Pre-training

Pre-trained **World Foundation Models (WFMs)** are general-purpose models that learn real-world physics and natural behaviors. Two deep learning paradigms are used to build them:

- Diffusion models: Break generation into a sequence of denoising steps.

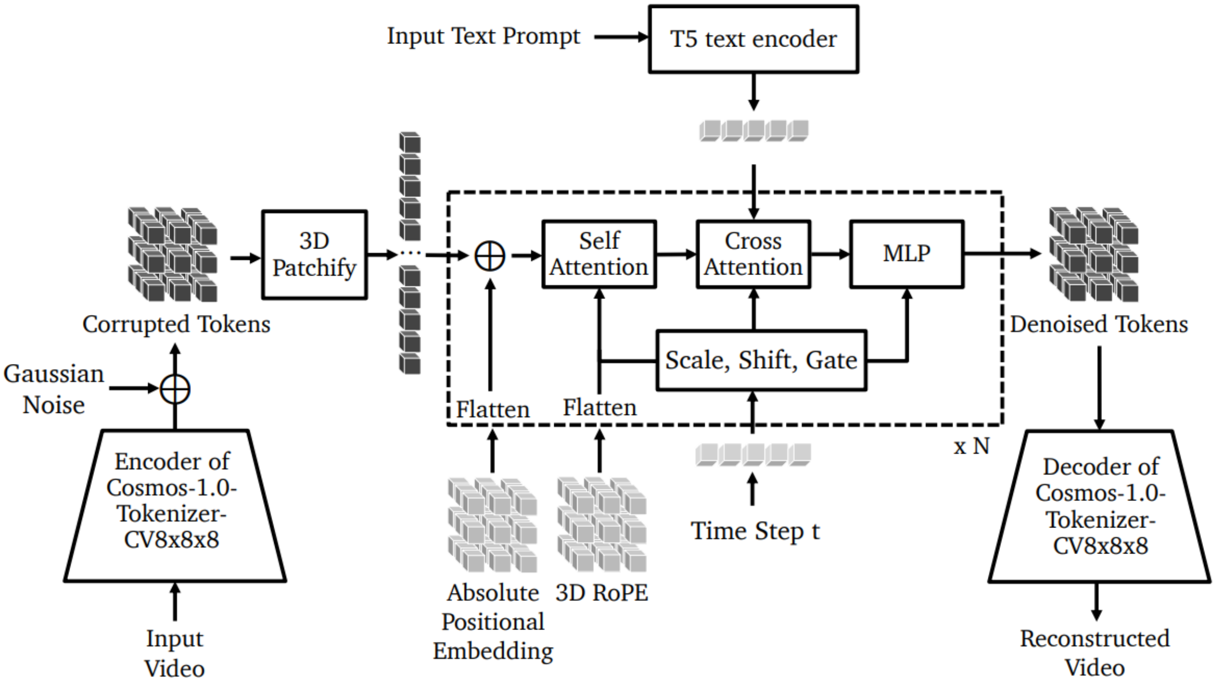
- Autoregressive models: Break generation into next-token prediction steps.

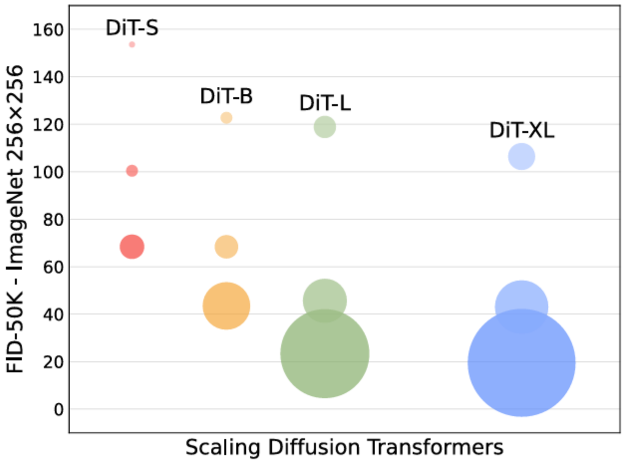
These scalable methods are optimized with GPU parallelization techniques. The WFMs discussed were trained **over three months** on a large-scale cluster of **10,000 NVIDIA H100 GPUs**.



Type	Models	Tokenizer	Enhancer
Diffusion	Cosmos-1.0-Diffusion-7B-Text2World → Cosmos-1.0-Diffusion-7B-Video2World Cosmos-1.0-Diffusion-14B-Text2World → Cosmos-1.0-Diffusion-14B-Video2World	Cosmos-1.0-Tokenizer-CV8x8x8	Cosmos-1.0-PromptUpsampler-12B-Text2World
Autoregressive	Cosmos-1.0-Autoregressive-4B → Cosmos-1.0-Autoregressive-5B-Video2World Cosmos-1.0-Autoregressive-12B → Cosmos-1.0-Autoregressive-13B-Video2World	Cosmos-1.0-Tokenizer-DV8x16x16	Cosmos-1.0-Diffusion-7B-Decoder-DV8x16x16ToCV8x8x8

3.Cosmos - Architecture : Diffusion World Foundation Model





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

$$\lambda(\sigma) = \left(\sigma^2 + \sigma_{\text{data}}^2\right) / \left(\sigma \cdot \sigma_{\text{data}}\right)^2,$$

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

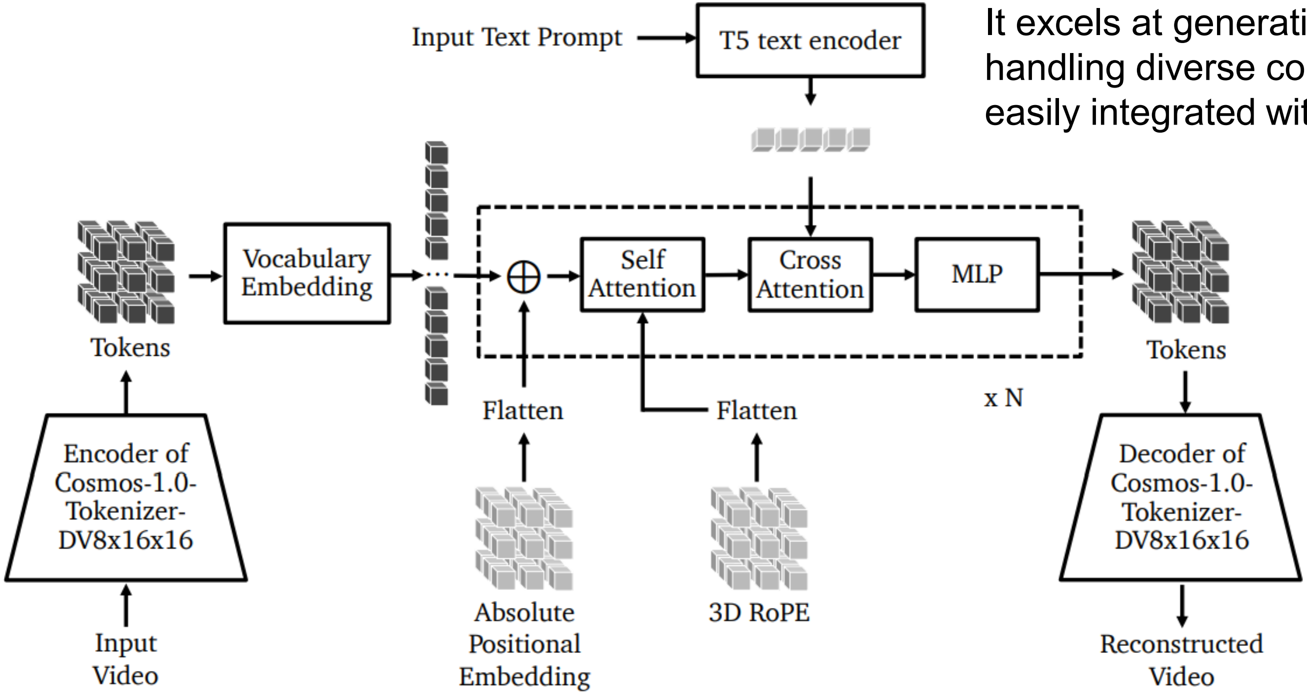
3D RoPE is used to encode relative positional information across the temporal, height, and width dimensions.

$$\mathcal{L}(D_\theta, \sigma) = \mathbb{E}_{\mathbf{x}_0,\mathbf{n}}\left[\left\|D_\theta(\mathbf{x}_0 + \mathbf{n}; \sigma) - \mathbf{x}_0\right\|_2^2\right],$$

Peebles, William, and Saining Xie. "Scalable diffusion models with transformers.“ ICCV 2023.

3.Cosmos - Architecture : Autoregressive-Video2World Model

A GPT-based Transformer model trained to predict the next token in a video sequence.
It excels at generating long-form videos and handling diverse control conditions and can be easily integrated with language model technologies.



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

3.Cosmos - Autoregressive vs Diffusion WFMs

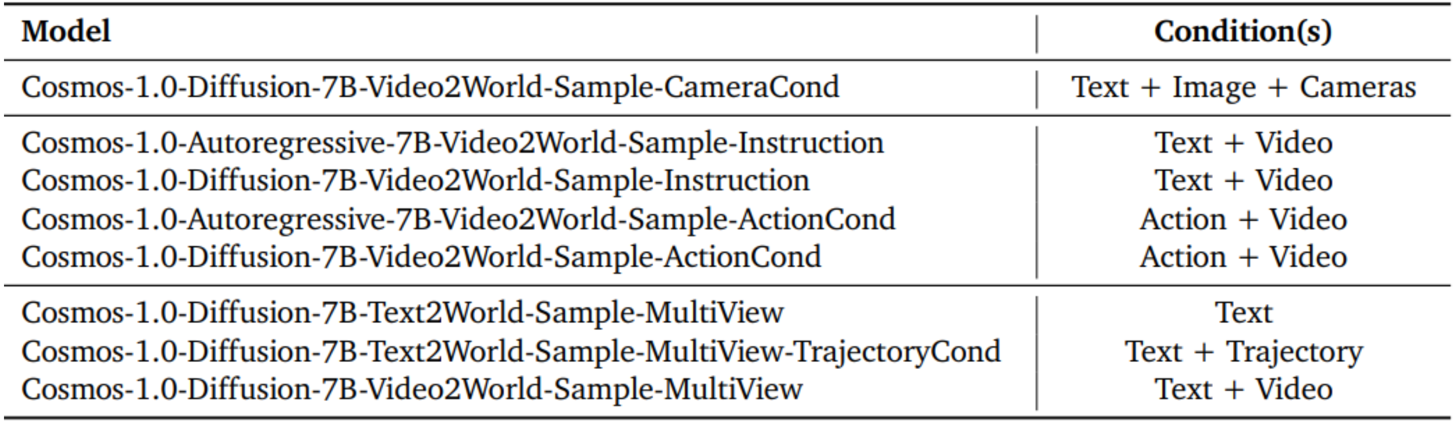
Feature	Diffusion WFMs	Autoregressive WFMs
Current Performance	Better 3D and video generation	Not yet on par
Control Signal Integration	**Strong (camera, end-effector, trajectories)**	Emerging potential
Output Flexibility	Multi-view videos, novel formats	Not highlighted yet
Inference Speed	Slower	Potentially **faster via causal optimization**
Pretrained Knowledge Transfer	Not emphasized	**Can leverage LLMs for richer world knowledge**
Hybrid Capability	Can distill into causal models	Can add bidirectional features and diffusion
Suitability for Real-Time/Planning	Less ideal currently	More promising with further development

Diffusion models currently outperform autoregressive models in 3D consistency and robotics video generation, **but hybrid approaches are increasingly promising**. The boundary between these models is blurring, as diffusion models can be distilled into causal transformers for efficiency, and autoregressive models can adopt bidirectional attention and diffusion-based generation heads.

3.Cosmos - Post-trained World Foundation Model

A pretrained **universal WFM is fine-tuned and applied** to various physical AI application domains.

[Major use cases include]
Camera Control: Provide a virtual environment where videos change in real-time according to desired camera positions.
Robotic Manipulation: Offer models capable of accurately predicting future states conditioned on robot actions or commands.
Autonomous Driving: Build autonomous driving simulators that generate realistic scenarios based on multi-camera views and vehicle trajectories.
These models can be efficiently developed **using small amounts of domain-specific data** based on a universal WFM.



Model	**Condition(s)**
Cosmos-1.0-Diffusion-7B-Video2World-Sample-CameraCond	Text + Image + Cameras
Cosmos-1.0-Autoregressive-7B-Video2World-Sample-Instruction	Text + Video
Cosmos-1.0-Diffusion-7B-Video2World-Sample-Instruction	Text + Video
Cosmos-1.0-Autoregressive-7B-Video2World-Sample-ActionCond	Action + Video
Cosmos-1.0-Diffusion-7B-Video2World-Sample-ActionCond	Action + Video
Cosmos-1.0-Diffusion-7B-Text2World-Sample-MultiView	Text
Cosmos-1.0-Diffusion-7B-Text2World-Sample-MultiView-TrajectoryCond	Text + Trajectory
Cosmos-1.0-Diffusion-7B-Video2World-Sample-MultiView	Text + Video

3.Cosmos - Post-training WFM for Camera Control

The model **Cosmos-1.0-Diffusion-7B-Video2World** is enhanced through **camera pose conditioning**, enabling **camera control** for 3D world simulation. The resulting **post-trained model** is named **Cosmos-1.0-Diffusion-7B-Video2World-Sample-CameraCond**.

- **Input**: A single reference image
- **Output**: A **temporally coherent**, **3D-consistent video** simulating different camera trajectories
- The simulation ensures **perspective changes align** with the true 3D structure of the scen

1.DL3DV-10K (Ling et al., CVPR 2024)

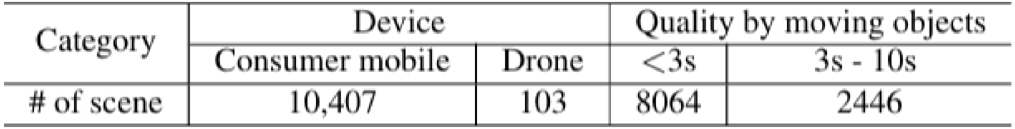
- A large-scale video dataset of static scenes(10,510). Videos are chunked into clips of 256 frames each (51,200K) , Used as input data for tasks involving scene understanding and modeling.

2.GLOMAP (Pan et al., ECCV 2024)

A Structure-from-Motion (SfM) method used to generate dense camera pose annotations for every frame within a clip.

Computes relative camera poses with the first frame set as the identity transformation.

Additionally, a proprietary Vision-Language Model (VLM) is employed to generate text prompts (captions) describing each clip as a static scene.



Category	Device		Quality by moving objects	
	Consumer mobile	Drone	<3s	3s - 10s
# of scene	10,407	103	8064	2446

Table 2. Number of scenes by devices and level of quality.

3.Cosmos - Post-training WFM for Camera Control

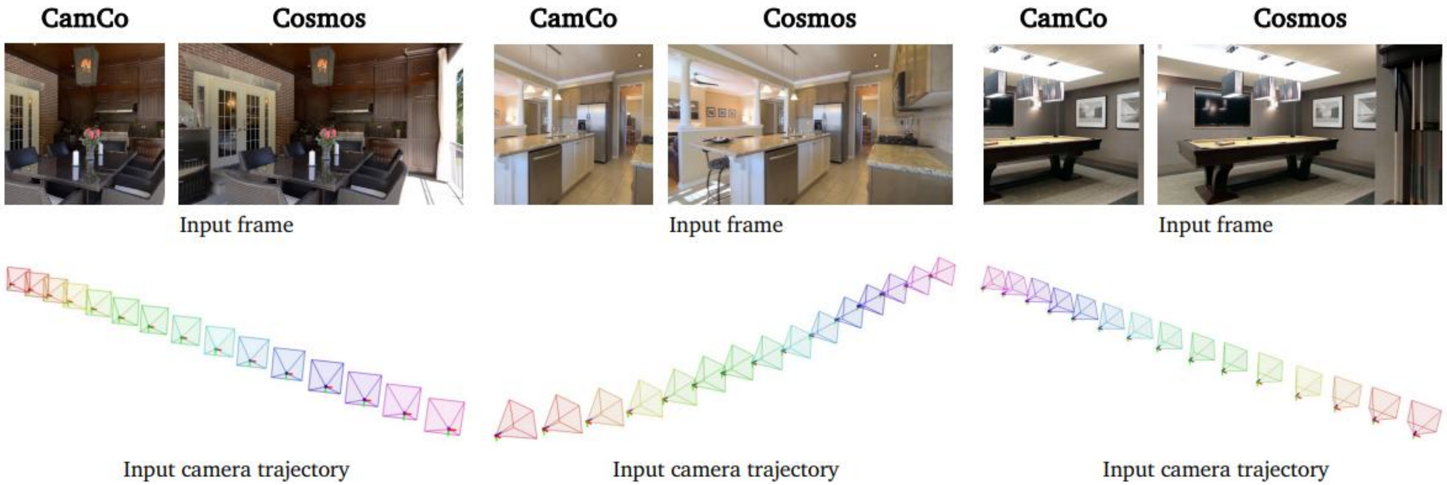
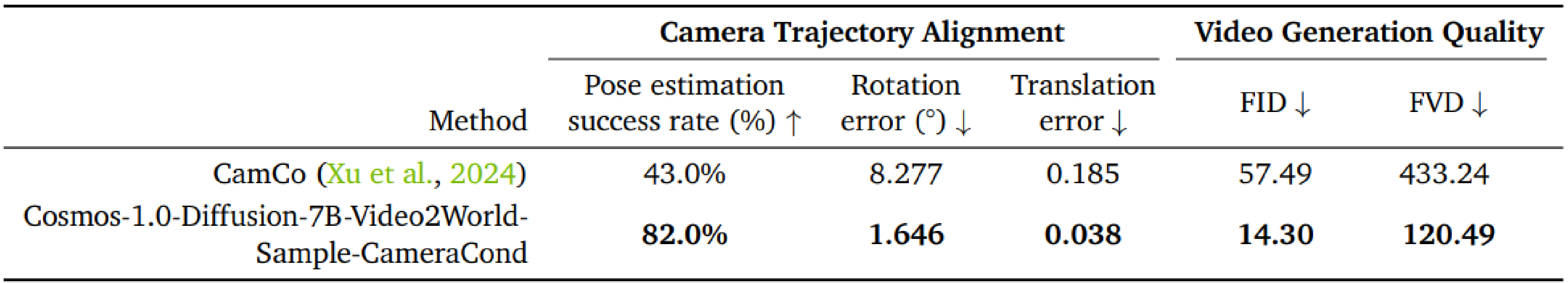
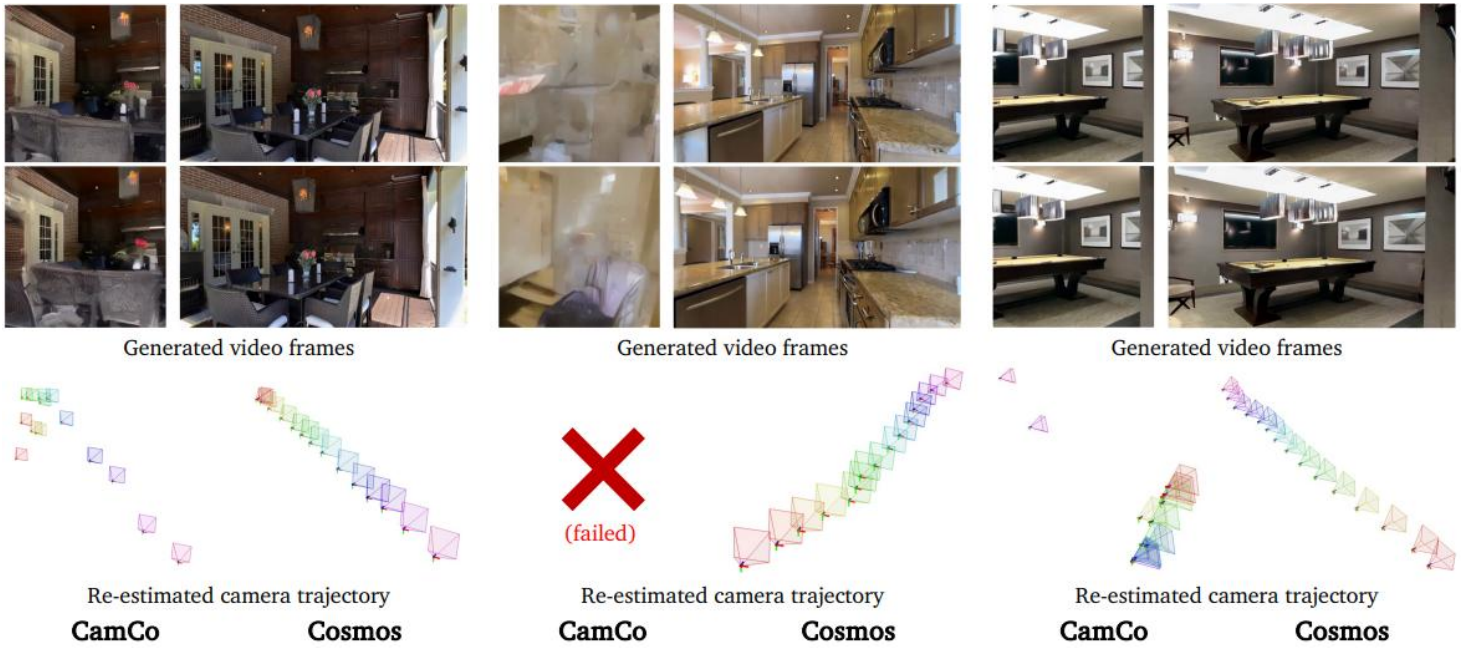


Table 22: Quantitative comparison of post-trained WFM with camera control.



	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-1.0-Diffusion-7B-Video2World-Sample-CameraCond	**82.0%**	**1.646**	**0.038**	**14.30**	**120.49**

3.Cosmos - Post-training WFM for Camera Control



3.Cosmos - Post-training WFM for Robotic Manipulation

A **pre-trained World Foundation Model (WFM)** is fine-tuned for two robotic manipulation tasks:

1.Instruction-Based Video Prediction

1. **Input**: Current video frame + a **text instruction**
2. **Output**: A video predicting the robot's future actions based on the instruction

2.Action-Based Next-Frame Prediction

1. **Input**: Current video frame + an **action vector**
2. **Output**: The **next frame**, showing the result of the action

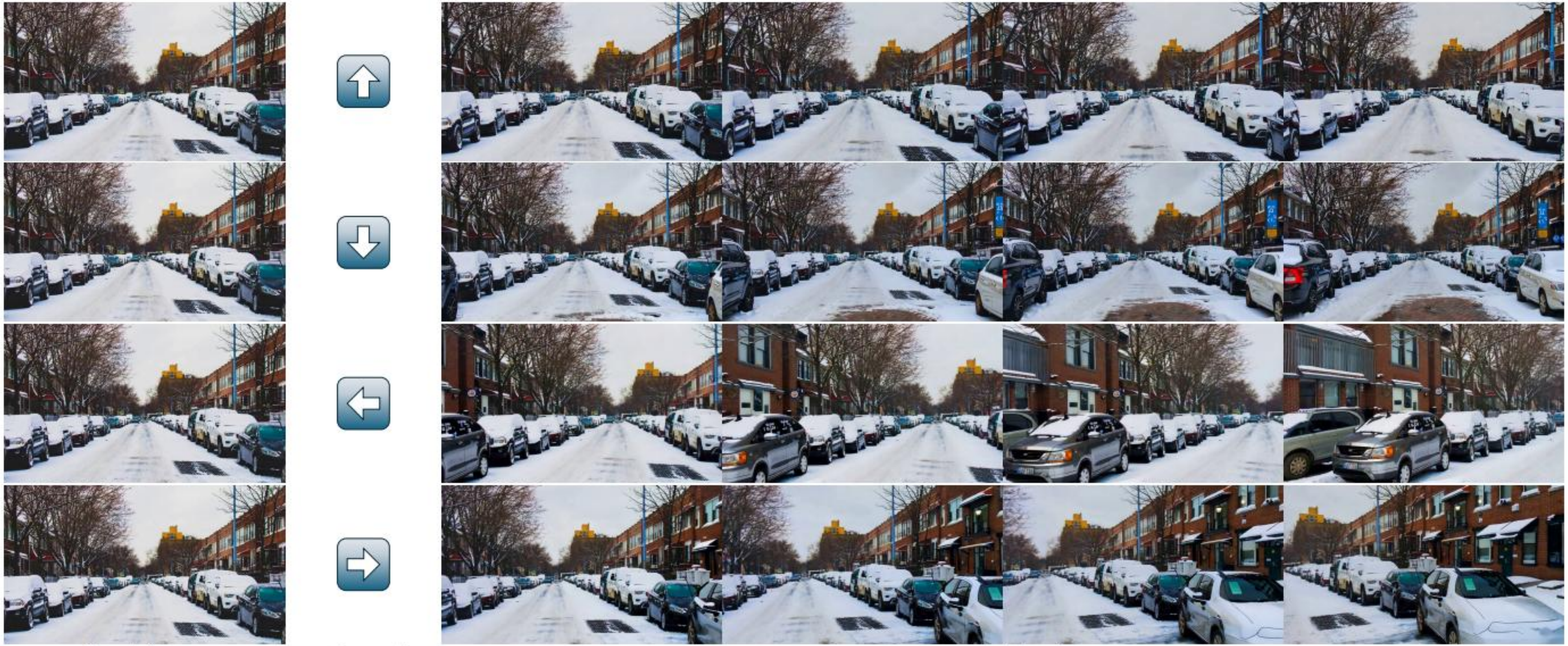
The model can also be run **autoregressively** over a sequence of actions to generate a full predicted video of the robot performing a task.

Cosmos-1X (Internal Dataset) : Instruction-based video prediction (30 FPS, 512×512 resolution)
Source: ~200 hours of egocentric videos from EVE, a humanoid robot by 1x.Tech
Content: ~12,000 labeled episodes (1–9 seconds) showing tasks like navigation, folding clothes, cleaning, etc.
Annotations: One-sentence instructions, upsampled via a proprietary VLM

Bridge (Ebert et al., RSS 2022) - Action-based next-frame generation (5 FPS, 320×256 resolution)
Content: ~20,000 episodes of third-person views of a robot arm in a kitchen setting
Annotations: Per-frame 7D action vectors in gripper space: (Δx, Δy, Δz, $\Delta \theta r$, $\Delta \theta p$, $\Delta \theta y$, ΔGripper)

3.Cosmos - Post-training WFM for Robotic Manipulation

joystick-like control input on the camera

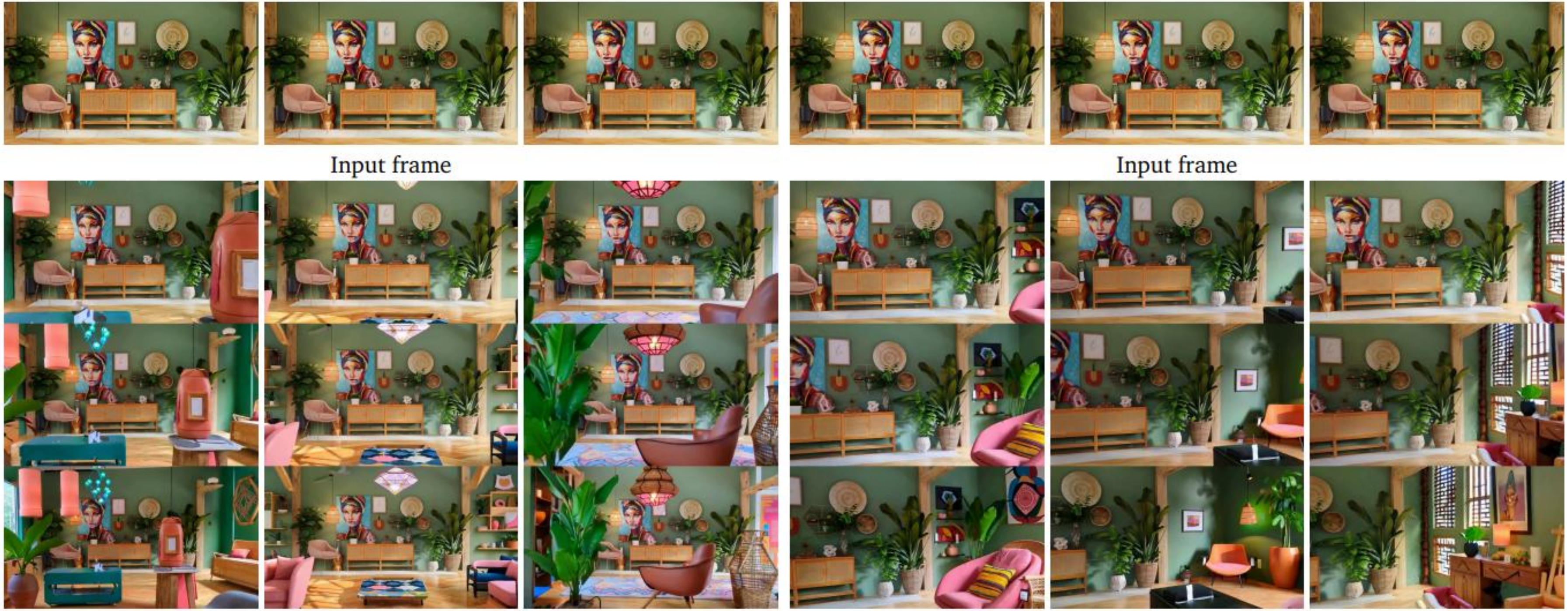


Input frame Control Generated video frames

The model supports joystick-like camera control (e.g., move forward/backward, rotate left/right), enabling interactive navigation of simulated 3D worlds. This allows users—or a Physical AI agent—to generate and predict future video frames based on different camera movements or scenarios.

3.Cosmos - Post-training WFM for Robotic Manipulation

Generation results from the same input image and camera control with different random seeds



Input frame

Input frame

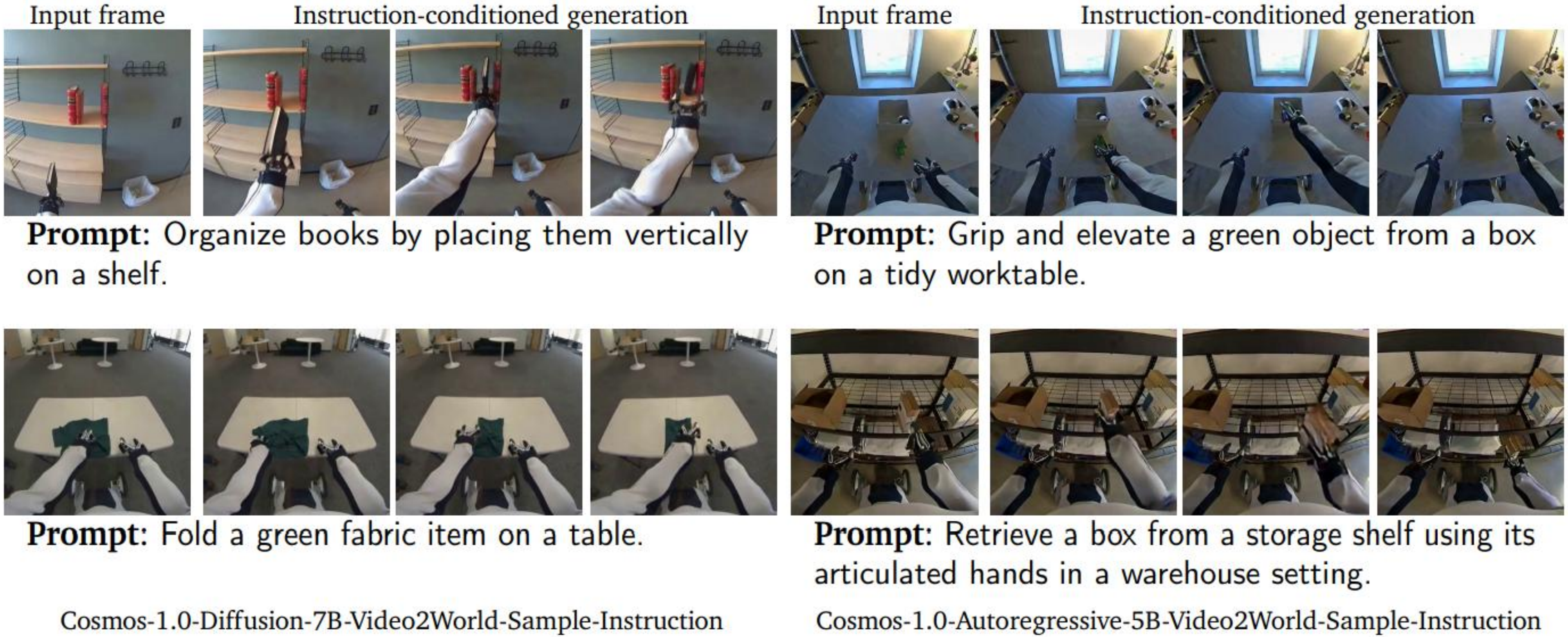
Generated frames with various seeds (moving backward)

Generated frames with various seeds (rotating right)

To demonstrate diversity, multiple videos are generated from the same input image and camera control using different random seeds.Cosmos-1.0-Diffusion-7B-Video2World-Sample-CameraCond produces varied yet 3D-consistent and temporally coherent simulations, enabling exploration of multiple possible futures from a single starting point.

3.Cosmos - Post-training WFM for Robotic Manipulation

Instruction-based video prediction samples on the Cosmos-1X dataset


Input frame
Instruction-conditioned generation
Input frame
Instruction-conditioned generation

Prompt: Organize books by placing them vertically on a shelf.

Prompt: Grip and elevate a green object from a box on a tidy worktable.

Prompt: Fold a green fabric item on a table.

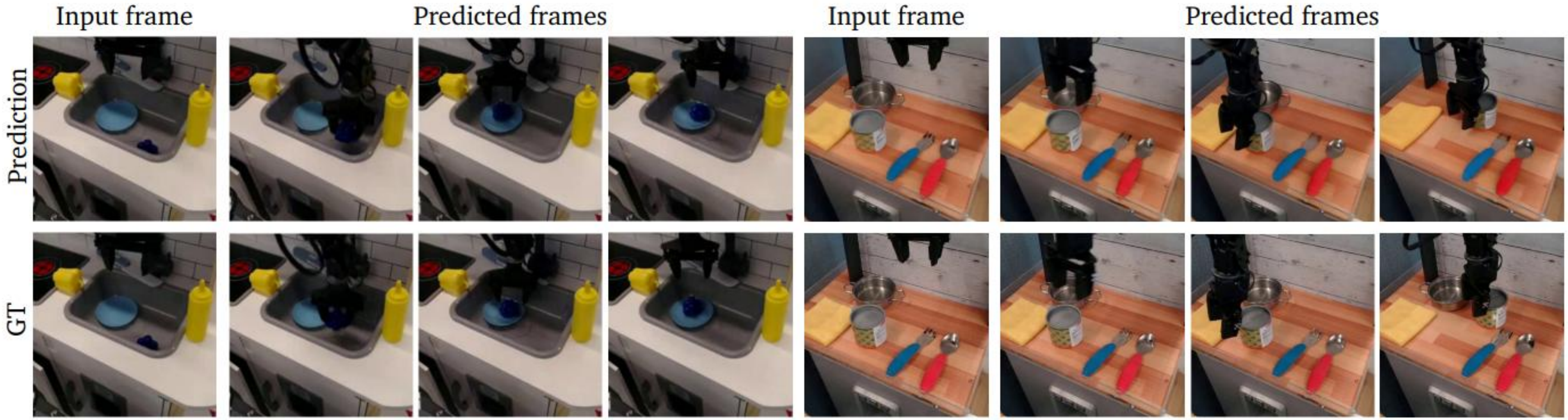
Prompt: Retrieve a box from a storage shelf using its articulated hands in a warehouse setting.

Cosmos-1.0-Diffusion-7B-Video2World-Sample-Instruction

Cosmos-1.0-Autoregressive-5B-Video2World-Sample-Instruction

3.Cosmos - Post-training WFM for Robotic Manipulation

Action-based next-frame prediction samples on the Bridge dataset



Cosmos-1.0-Diffusion-7B-Video2World-Sample-ActionCond

Cosmos-1.0-Autoregressive-5B-Video2World-Sample-ActionCond

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-1.0-Autoregressive-5B-Video2World-Sample-ActionCond	19.95	0.80	0.36	434
Cosmos-1.0-Diffusion-7B-Video2World-Sample-ActionCond	**21.14**	**0.82**	**0.32**	**190**

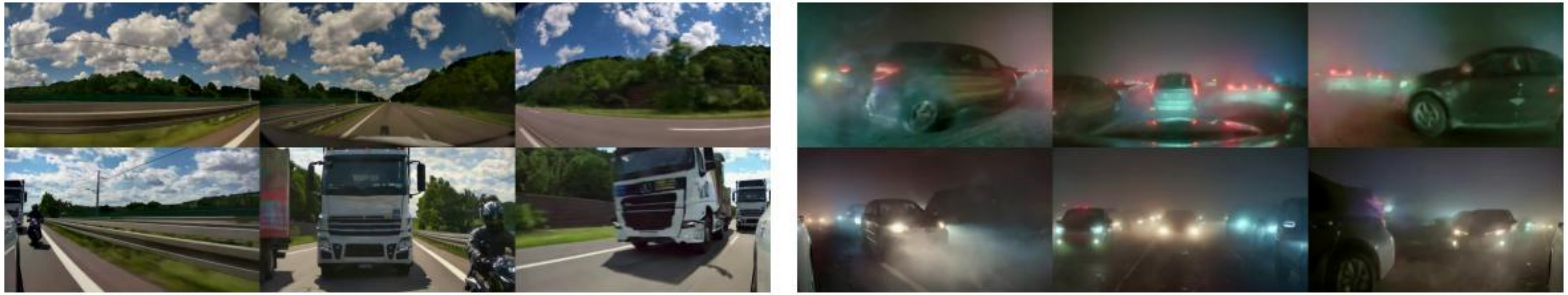
3.Cosmos - Post-training WFM for Autonomous Driving

A multi-view world model is developed by fine-tuning a pre-trained World Foundation Model (WFM) for autonomous driving.It simulates in-the-wild driving scenes using inputs from multiple camera views, matching the sensor setup of autonomous vehicles, making it suitable for training driving agents in realistic conditions.

Real Driving Scene (RDS) dataset:
•**Source**: Curated internally by NVIDIA; ~3.6 million **20 second video clips** (~20,000 hours).
•**Views**: 6 synchronized camera angles - **surround-view** (front, left, right, rear, rear-left, rear-right)
•**Extras**: Includes **ego-motion data** for trajectory reconstruction.
•**Timestamps**: Front camera timestamps used to align all views.

Vehicle density: none, low, medium, high
Weather: clear, rain, snow, fog
Lighting: day, night
Speed: standing, low, local, highway
Driving behavior: trajectory curvature and acceleration (high/medium/low)
Road type: rural, residential, urban (based on OpenStreetMap)
Augmented to include rare road structures (e.g., tollbooths, tunnels, bridges).
Captions: Each view is annotated with a template-based description of the camera's position and orientation.

3.Cosmos - Post-training WFM for Autonomous Driving

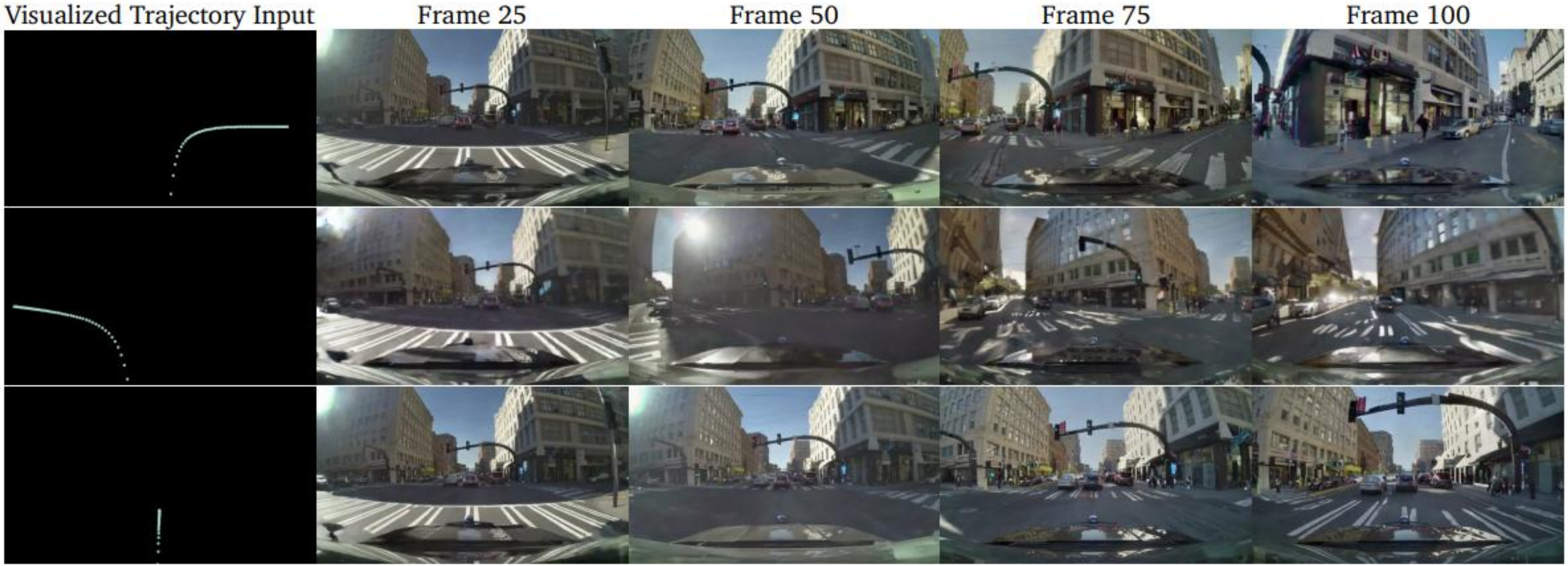


Prompt: The video captures a highway scene with a white truck in the foreground, moving towards the camera. The truck has a large cargo area and is followed by a motorcyclist wearing a full-face helmet. The road is marked with white lines and has a metal guardrail on the right side. The sky is partly 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.

Text-conditioned samples generated by Cosmos-1.0-Diffusion-7B-Text2World-SampleMultiView, extended to 8 seconds by Cosmos-1.0-Diffusion-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.

3.Cosmos - Post-training WFM for Autonomous Driving



Trajectory-conditioned generated samples from Cosmos-1.0-Diffusion-7B-Text2World-SampleMultiView-TrajectoryCond.
Given the trajectory inputs on the left-most column, we generate multi-view videos that follow the given trajectory. We visualize the front camera view in this figure.

3.Cosmos - Post-training WFM for Autonomous Driving

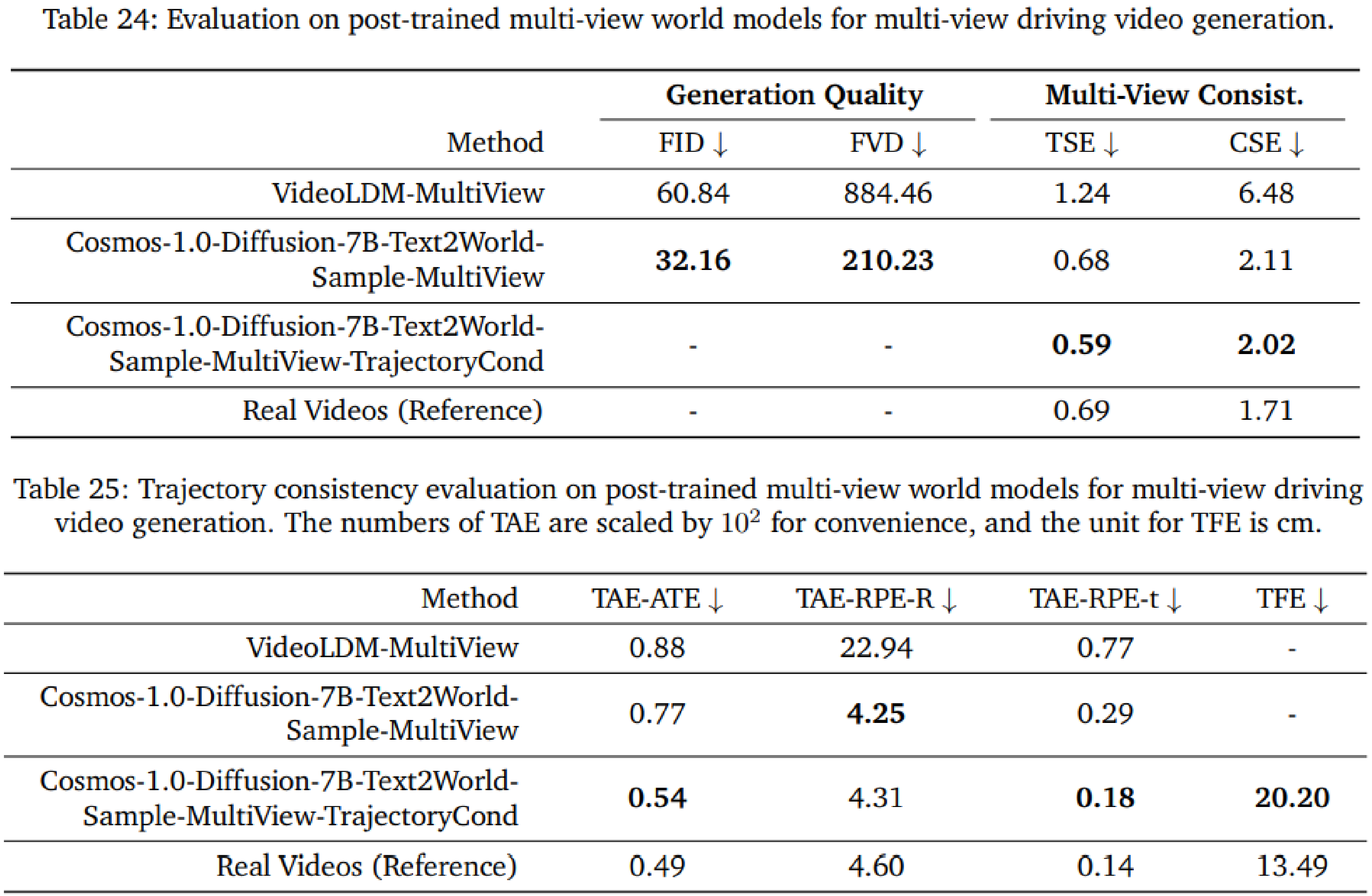


Table 24: Evaluation on post-trained multi-view world models for multi-view driving video generation.

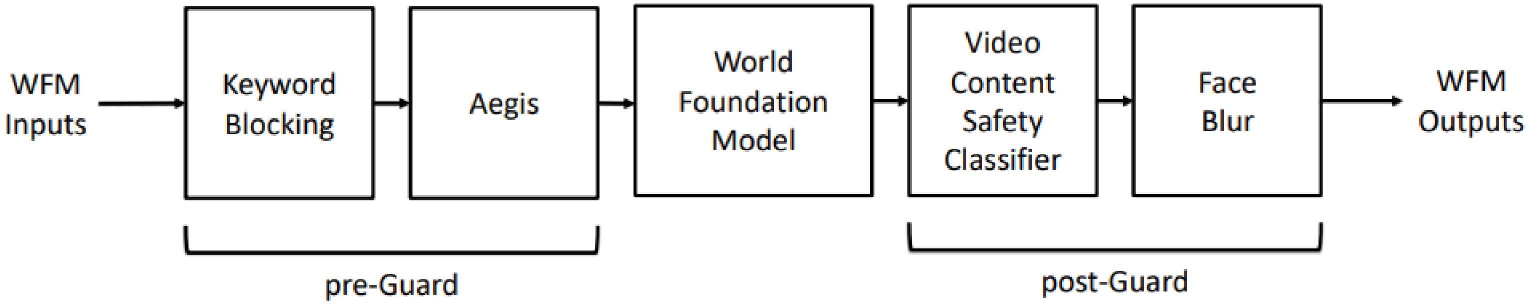
	Generation Quality		Multi-View Consist.	
Method	FID ↓	FVD ↓	TSE ↓	CSE ↓
VideoLDM-MultiView	60.84	884.46	1.24	6.48
Cosmos-1.0-Diffusion-7B-Text2World-Sample-MultiView	**32.16**	**210.23**	0.68	2.11
Cosmos-1.0-Diffusion-7B-Text2World-Sample-MultiView-TrajectoryCond	-	-	**0.59**	**2.02**
Real Videos (Reference)	-	-	0.69	1.71

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-1.0-Diffusion-7B-Text2World-Sample-MultiView	0.77	**4.25**	0.29	-
Cosmos-1.0-Diffusion-7B-Text2World-Sample-MultiView-TrajectoryCond	**0.54**	4.31	**0.18**	**20.20**
Real Videos (Reference)	0.49	4.60	0.14	13.49

3.Cosmos - Guardrails

the safe use of WFMs



The Cosmos platform implements a two-stage Guardrail system to ensure safety

Pre-Guard (Input Safety Mechanism):
A system that blocks inappropriate or harmful content from input prompts, employing composite filtering based on keywords and a language model (Aegis).
Post-Guard (Output Safety Mechanism):
A system that automatically analyzes generated video outputs to block or modify harmful or inappropriate content, utilizing tools such as video content safety classifiers and face-blurring technologies.

4.Conclusion & Limitation

The Cosmos is a key step toward building general-purpose simulators **for the physical world**.

The work presents a full-stack approach—covering **data curation**, **tokenizer design**, **diffusion and autoregressive model architectures**, and **fine-tuning** for tasks in **3D navigation**, **robotic manipulation**, and **autonomous driving**, all requiring **3D consistency and action control**.

Limitations:

•**Early-stage models** still struggle with:

- Lack of **object permanence**
- Poor handling of **contact-rich dynamics**
- **Instruction-following inconsistencies**
- Limited realism in simulating **physical laws** (e.g., gravity, light, fluids)

•**Evaluation challenges**:

- Human judgments are **subjective and biased**
- Misalignment between **human evaluation** and **task-specific metrics**

Future Directions: Develop **automated evaluators** using **multi-modal LLMs**

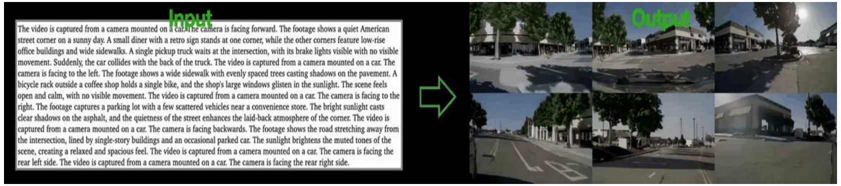
•Use existing physical simulators for **reproducible, interactive evaluations** to **reduce reliance on human judgment.**

Cosmos enables safe, realistic digital training and testing for physical AI agents, avoiding real-world risks and costs. It thus significantly accelerates physical AI development and offers a scalable solution for previously difficult challenges.

5.Cosmos World Foundation Models

Cosmos-Predict

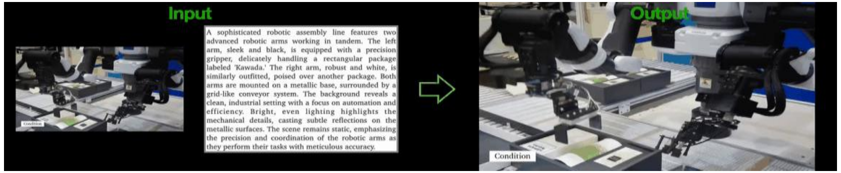
Text2World



The video is captured from a camera mounted on a car. The camera is facing forward. The footage shows a quiet American street corner on a sunny day. A small diner with a retro sign stands at one corner, while the other corners feature low-rise office buildings and wide sidewalks. A single pickup truck waits at the intersection, with its brake lights visible with no visible movement. Suddenly, the car collides with the back of the truck. The video is captured from a camera mounted on a car. The camera is facing to the left. The footage shows a wide sidewalk with evenly spaced trees casting shadows on the pavement. A bicycle rack outside a coffee shop holds a single bike, and the shop's large windows glisten in the sunlight. The scene feels open and calm, with no visible movement. The video is captured from a camera mounted on a car. The camera is facing to the right. The footage captures a parking lot with a few scattered vehicles near a convenience store. The bright sunlight casts clear shadows on the asphalt, and the quietness of the street enhances the laid-back atmosphere of the corner. The video is captured from a camera mounted on a car. The camera is facing backwards. The footage shows the road stretching away from the intersection, lined by single-story buildings and an occasional parked car. The sunlight brightens the muted tones of the scene, creating a relaxed and spacious feel. The video is captured from a camera mounted on a car. The camera is facing the rear left side. The video is captured from a camera mounted on a car. The camera is facing the rear right side.

Text Prompt → Video Generation

Video2World

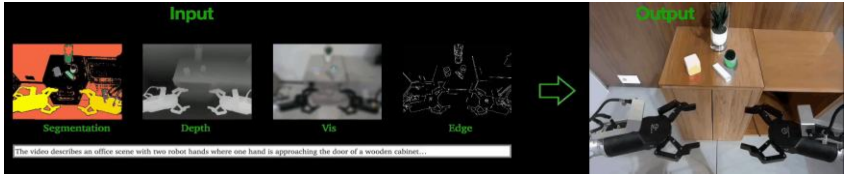


A sophisticated robotic assembly line features two advanced robotic arms working in tandem. The left arm, sleek and black, is equipped with a precision gripper, delicately handling a rectangular package labeled 'Kawada.' The right arm, robust and white, is similarly outfitted, poised over another package. Both arms are mounted on a metallic base, surrounded by a grid-like conveyor system. The background reveals a clean, industrial setting with a focus on automation and efficiency. Bright, even lighting highlights the mechanical details, casting subtle reflections on the metallic surfaces. The scene remains static, emphasizing the precision and coordination of the robotic arms as they perform their tasks with meticulous accuracy.

Text Prompt + image/video input → Video Generation

Cosmos-Transfer

World2World



The video describes an office scene with two robot hands where one hand is approaching the door of a wooden cabinet...

Text Prompt + Seg/Depth/Vis/Edge Video input → Video Generation

Cosmos-Reason

User prompt:
Is it safe to
turn right?

Based on the video,
turning right may not
be entirely safe due to
the following
factors: ...

Text Prompt + Video input → **Multimodal LLM based input Text** Generation

5.Cosmos World Foundation Models

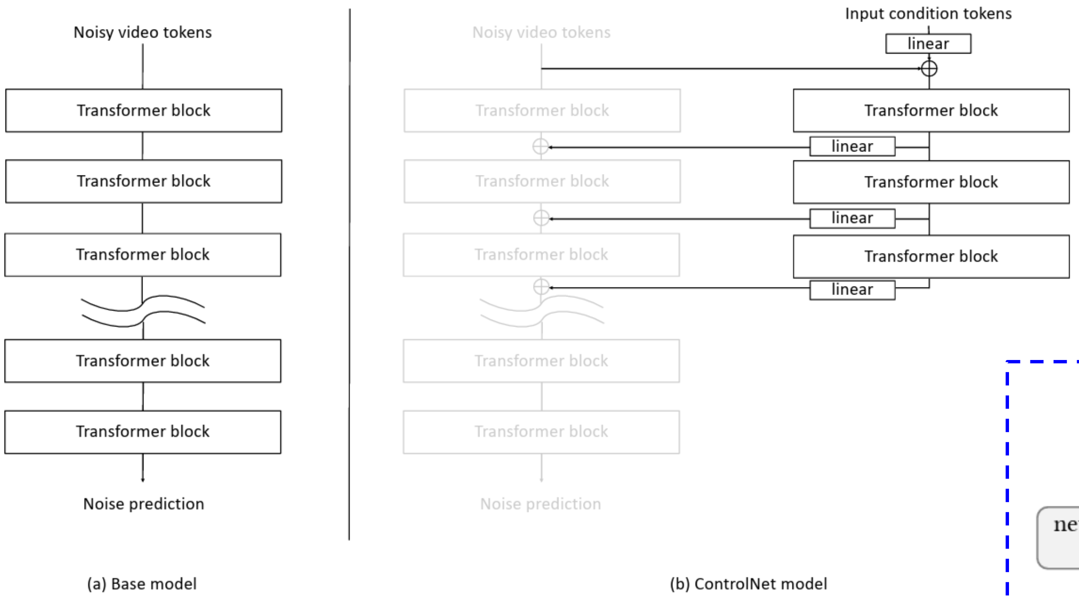
Cosmos-Transfer1: Conditional World Generation with Adaptive Multimodal Control

Cosmos-Transfer1 is a **conditional world generation model** that generates simulations based on **multi-modal spatial inputs** such as **segmentation**, **depth**, and **edges**.
It features an **adaptive spatial conditioning scheme**, allowing different inputs to be weighted **differently across spatial locations**, enabling **highly controllable world generation**.

Key Highlights:
•Supports **world-to-world transfer** tasks, including **Sim2Real**.
•Evaluated for applications in **Physical AI**, such as **robotics** and **autonomous vehicle data enrichment**.
•Includes an **inference scaling strategy** for **real-time generation** using NVIDIA GB200 NVL72(72 GPU).
•**Models and code are open-sourced** to support community research.

Related Works
Visual Domain Transfer: Converting abstract inputs like segmentation maps into photorealistic images. This has extended to video synthesis, enabling dynamic and temporally coherent generation.
Spatial Control in Diffusion Models: ControlNet and its extensions, including adaptations for transformers and video generation, offer improved precision in spatially guided outputs.
Enhancing Simulation with Generative Models: Diffusion-based models with spatial control outperform GANs in sim-to-real transfer, supporting safer and more diverse training for Physical AI systems.

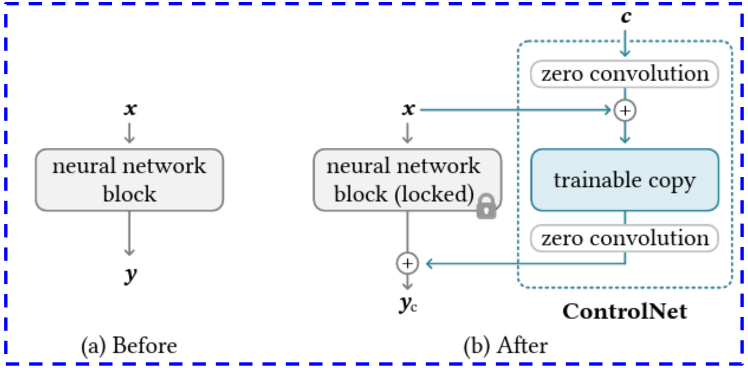
5.Cosmos World Foundation Models - ControlNet (ICCV 2023)



(a) Base model

(b) ControlNet model

Input Canny edge
Default
"masterpiece of fairy tale, giant deer, golden antlers"
"..., quaint city Galic"
Input human pose
Default
"chef in kitchen"
"Lincoln statue"



(a) Before

(b) After

(a) Base model is the base DiT-based diffusion model. It consists of a sequence of transformer blocks and learns to predict the added noise in the input noisy tokens.

(b) ControlNet extends the base model to a conditional diffusion model.

Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." ICCV 2023.

5.Cosmos World Foundation Models

Modality and Training
Each variant of Cosmos-Transfer1-7B allows targeted control over different visual attributes, enhancing flexibility and realism in simulation tasks.

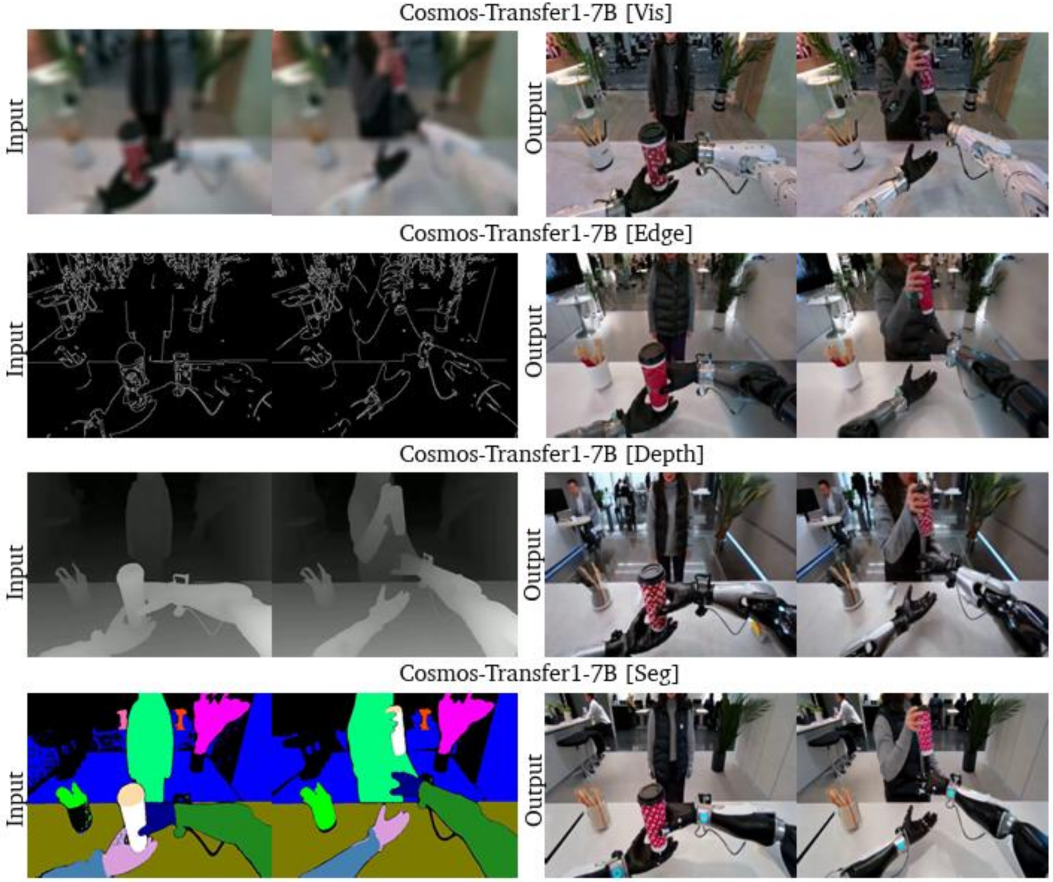
[Vis] — Bilateral Blur Input Input: Blurry video using bilateral blur
Use: Preserve color and rough shapes while modifying texture details (e.g., for CG-to-real transfer)
Training: Blur parameters randomized as data augmentation

[Edge] — Canny Edge Input Input: Canny edges, frame-by-frame
Use: Preserve scene structure while allowing creative freedom in texture and detail
Training: Edge thresholds randomized

[Depth] — Depth Map Input Input: Normalized depth maps from DepthAnything2
Use: Maintain accurate 3D geometry of the scene

[Seg] — Segmentation Mask Input Input: Object segmentation masks using GroundingDino + SAM2
Use: Preserve segmentation layout while enabling free generation
Note: Segmentation colors are randomized and non-semantic

5.Cosmos World Foundation Models



Cosmos-Transfer1-7B [Vis]

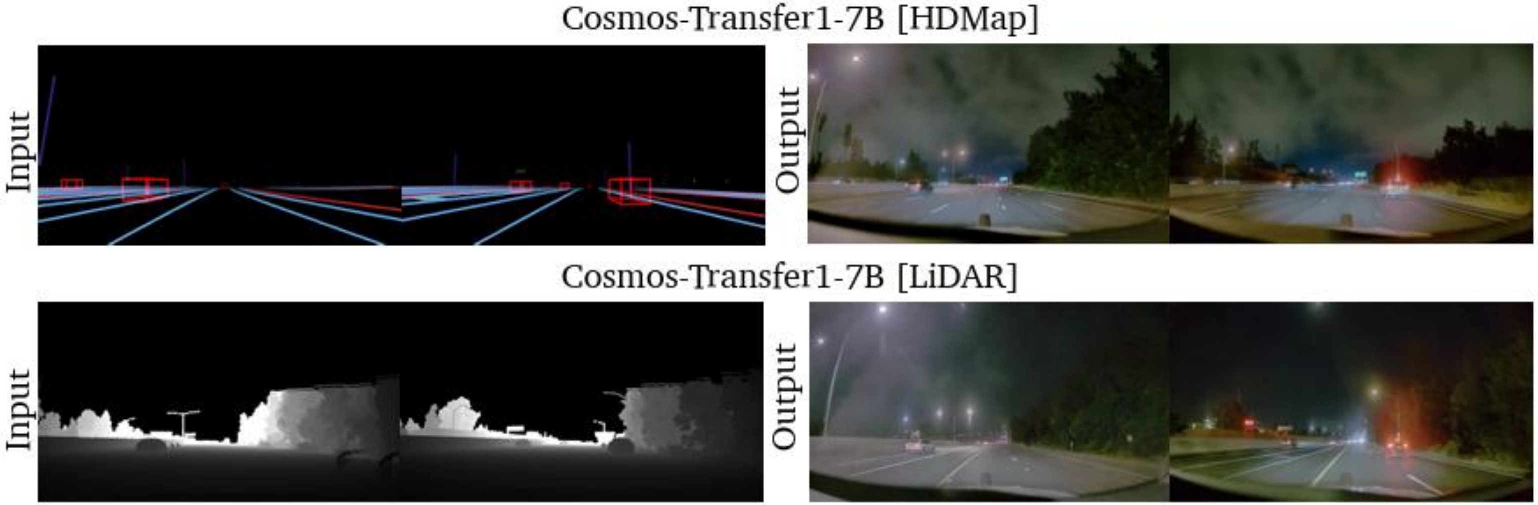
Cosmos-Transfer1-7B [Edge]

Cosmos-Transfer1-7B [Depth]

Cosmos-Transfer1-7B [Seg]

"In a modern office, sleek black robotic arms with articulated joints interact with a woman at a white counter. She wears a dark vest over a gray long-sleeve shirt. The arms smoothly hand her a red and white patterned coffee cup with a black lid."

5.Cosmos World Foundation Models



Cosmos-Transfer1-7B [HDMap]

Cosmos-Transfer1-7B [LiDAR]

"The video is captured from a camera mounted on a car. The camera is facing forward. The video captures a nighttime drive on a multi-lane highway. The road is illuminated by streetlights, and several vehicles are visible, including a white car on the left and a black car on the right. The scene is relatively quiet, with minimal traffic. The surrounding area is dark, and there are trees and bushes on the right side of the road. The weather appears to be clear, with no visible signs of rain or fog. The time of day is nighttime, as indicated by the darkness and artificial lighting."

Figure 4: **Input and generated videos from Cosmos-Transfer1-7B-Sample-AV operating on individual modality settings.** Cosmos-Transfer1-7B-Sample-AV [HDMap] preserves the original road layout of a driving scene while Cosmos-Transfer1-7B-Sample-AV [LiDAR] preserves the input semantic details.

5.Cosmos World Foundation Models

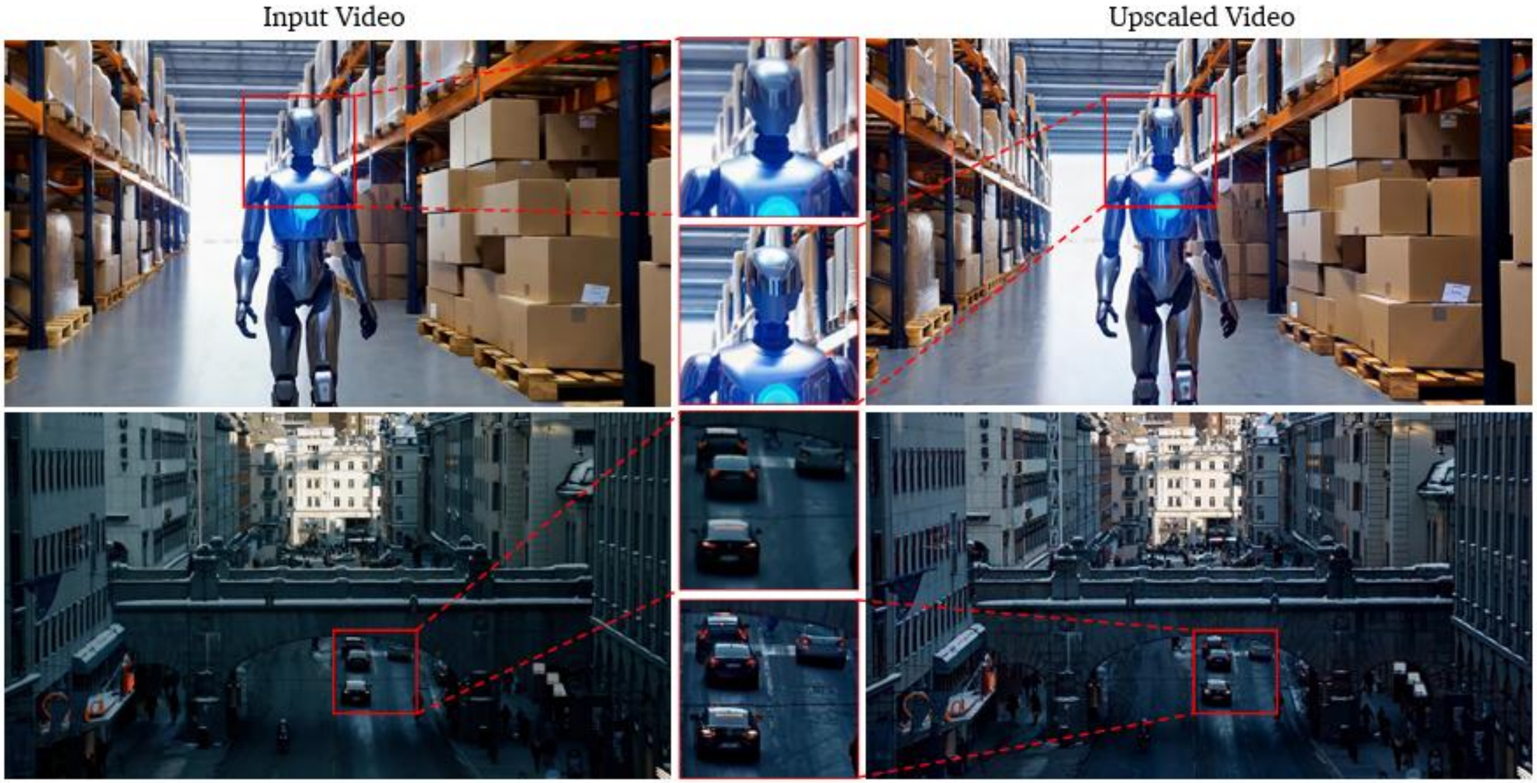
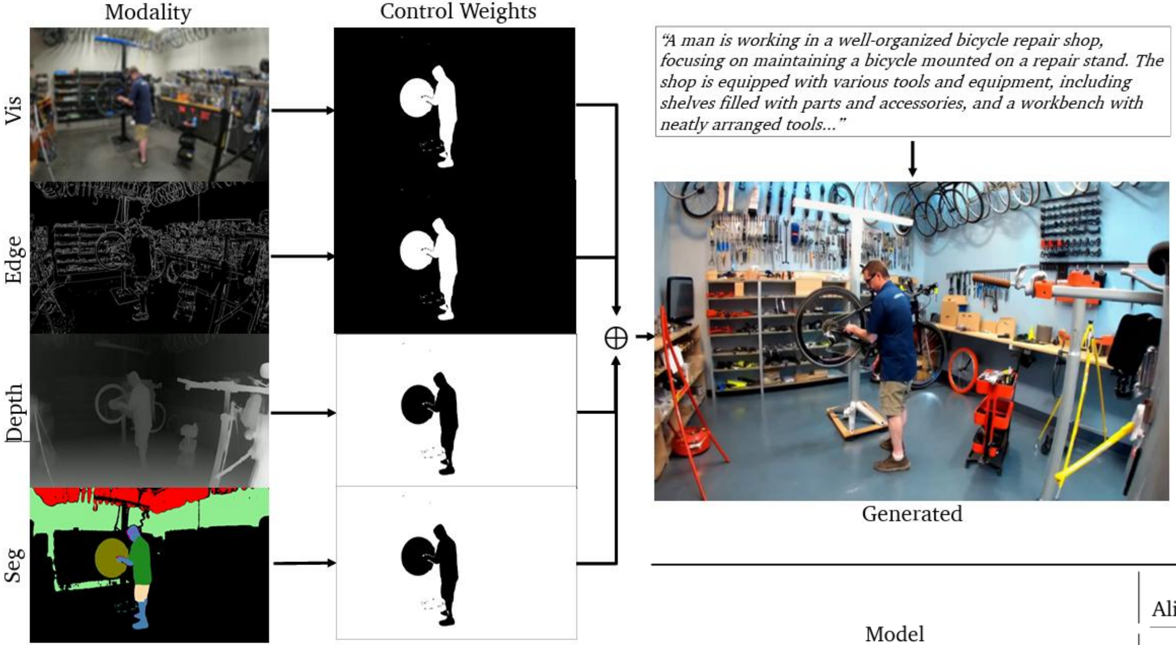


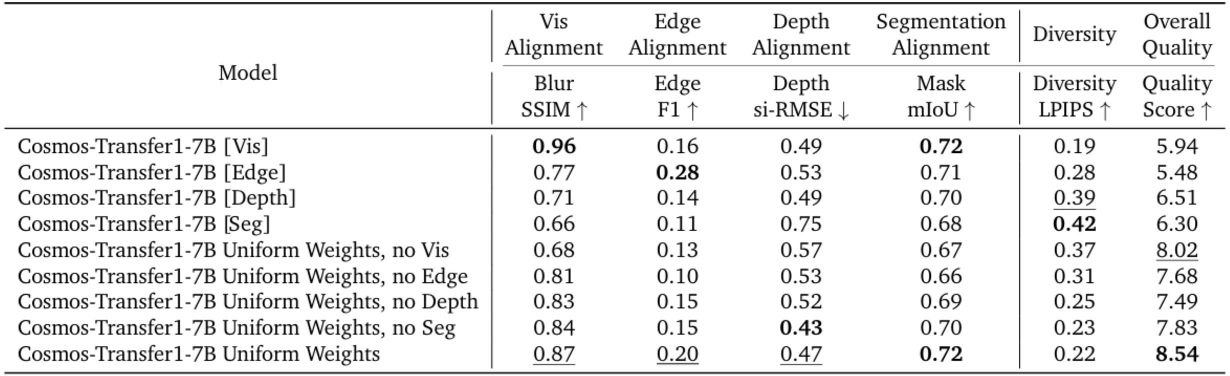
Figure 5: **Cosmos-Transfer1-7B-4KUpscaler upscales videos from 720p to 4k resolution.** The input video in the first row is a generated video, while the second row is a real video. Note how the model adds realistic reflections and sharpens the textures in the input.

5.Cosmos World Foundation Models



"A man is working in a well-organized bicycle repair shop, focusing on maintaining a bicycle mounted on a repair stand. The shop is equipped with various tools and equipment, including shelves filled with parts and accessories, and a workbench with neatly arranged tools..."

the impact of different control modalities



Model	Vis Alignment	Edge Alignment	Depth Alignment	Segmentation Alignment	Diversity	Overall Quality
	Blur SSIM ↑	Edge F1 ↑	Depth si-RMSE ↓	Mask mIoU ↑	Diversity LPIPS ↑	Quality Score ↑
Cosmos-Transfer1-7B [Vis]	**0.96**	0.16	0.49	**0.72**	0.19	5.94
Cosmos-Transfer1-7B [Edge]	0.77	**0.28**	0.53	0.71	0.28	5.48
Cosmos-Transfer1-7B [Depth]	0.71	0.14	0.49	0.70	0.39	6.51
Cosmos-Transfer1-7B [Seg]	0.66	0.11	0.75	0.68	**0.42**	6.30
Cosmos-Transfer1-7B Uniform Weights, no Vis	0.68	0.13	0.57	0.67	0.37	8.02
Cosmos-Transfer1-7B Uniform Weights, no Edge	0.81	0.10	0.53	0.66	0.31	7.68
Cosmos-Transfer1-7B Uniform Weights, no Depth	0.83	0.15	0.52	0.69	0.25	7.49
Cosmos-Transfer1-7B Uniform Weights, no Seg	0.84	0.15	**0.43**	0.70	0.23	7.83
Cosmos-Transfer1-7B Uniform Weights	0.87	0.20	0.47	**0.72**	0.22	**8.54**

5.Cosmos World Foundation Models

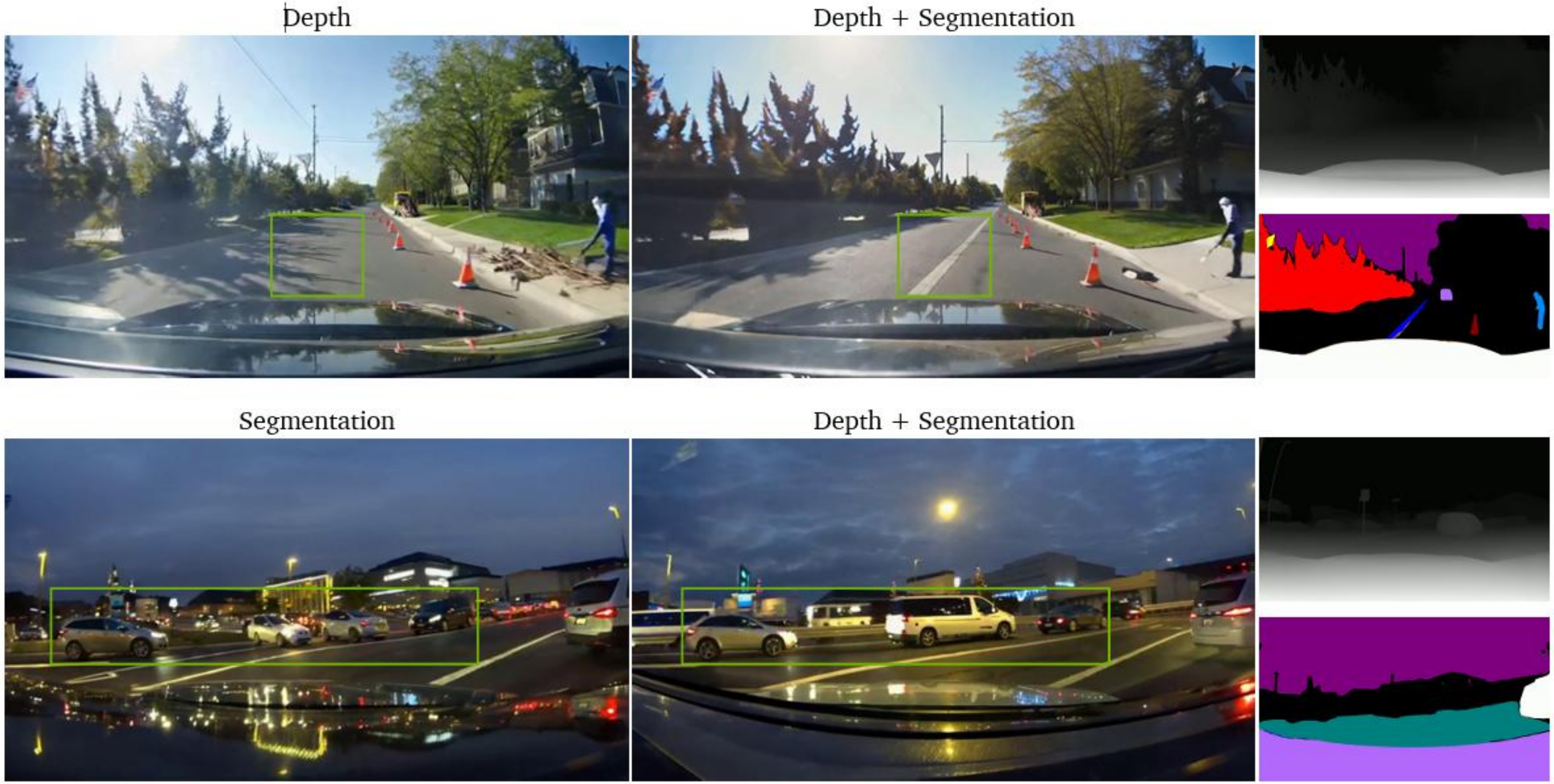


Figure 9: **Comparison of the generation results conditioned on depth and segmentation of Cosmos-Transfer1-7B**. In each example, the highlighted regions illustrate the enhancements achieved by incorporating multiple control signals over relying on a single one.

Cosmos (Original) vs Cosmos-Transfer1 (New)

Aspect	Cosmos (Original)	Cosmos-Transfer1 (New)
Model Architecture	Dual model: Diffusion + Autoregressive transformers	Diffusion-only model with multi-branch ControlNet architecture
Conditioning Inputs	Text prompts, past video frames	Multiple spatial inputs: segmentation, depth, edge, HDMap, LiDAR
Control Mechanism	Basic text or action conditioning	Spatiotemporal adaptive control maps per modality
Training Strategy	Full pre-training (10K H100 GPUs, 3 months), then fine-tuning	ControlNet modules trained separately on frozen backbone (Cosmos-Predict1)
Tokenization	Introduced Cosmos Tokenizer (Continuous & Discrete, causal, high compression quality)	Reuses Cosmos Tokenizer; no changes
Dataset Strategy	Massive scale: 100M+ clips from internet videos	Uses Cosmos fine-tuning data + new paired datasets (e.g. RDS-HQ for driving)
Evaluation Metrics	Visual quality (FVD, PSNR, SSIM)	Adds alignment metrics (Blur SSIM, Edge F1, Depth RMSE, Seg mIoU), Diversity, Quality
Robotics Application	Forecasting future video given robot actions	Sim2Real video synthesis from segmentation/depth input in simulation
Autonomous Driving	Mentioned as future direction	Full HDMap & LiDAR-to-Video generation with spatial realism
Real-Time Inference	Not addressed	Real-time generation achieved on GB200 NVL72 (e.g., 5s video in 4.2s)
Deployment Focus	General world modeling for Physical AI	Controllable Sim2Real generation for robotics and AV scenarios

Thanks

Any Questions?

You can send mail to
Susang Kim(healess1@gmail.com)