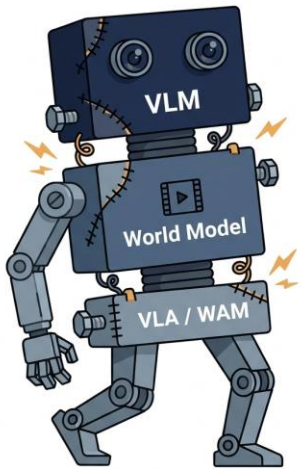
Cosmos 3: Omnimodal World Models for Physical AI

Technical report from Nvidia

Presented by: Anika Sharma

Motivation

- **Physical AI requires two coupled capabilities:** understanding and generation.
- **Prior work treats these capabilities in isolation:**
  - Perception & reasoning (VLMs)
  - World simulation (Video Generation, Forward Dynamics)
  - Action prediction (VLAs, WAMs)
- **Limitation:** These capabilities are interdependent, but existing approaches rely on fragmented, task-specific models.
- **Goal:** Unify understanding, simulation, and action generation within a single framework.



Gemini Notebook

Cosmos-3

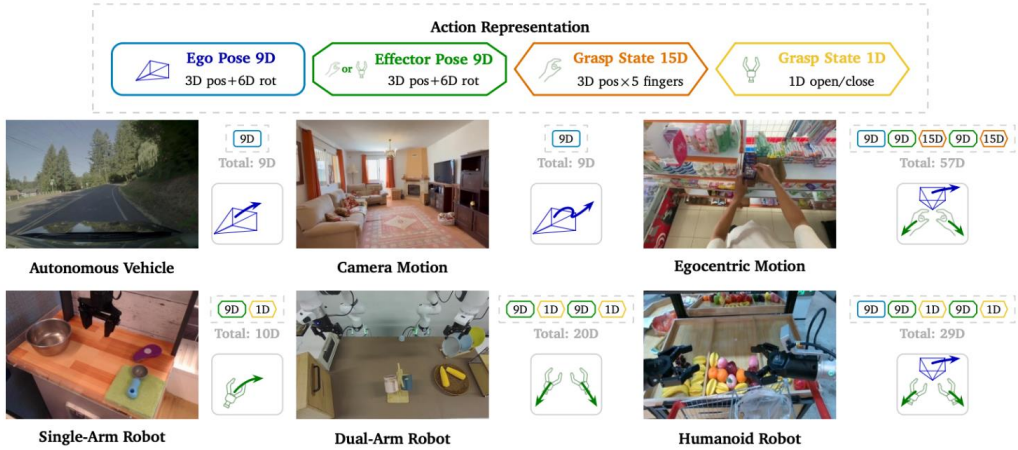


Figure 1: **Cosmos 3 serves as a general-purpose backbone for Physical AI.** By jointly modeling language, image, video, audio, and action for both understanding and generation, Cosmos 3 unifies a wide range of model classes within a single network architecture, including vision-language models, image generation models, audio-visual generation models, policy or world-action models, forward dynamics models, and inverse dynamics models.

Modality specific encoders: vision and audio

- **Image & Video Understanding:** Pretrained ViT encoder extracts visual features; jointly trained with the backbone.
- **Image & Video Generation:** Frozen video VAE compresses inputs temporally (4×) and spatially (32×32).
- **Audio Generation:** Frozen audio VAE encodes 48 kHz stereo audio into 25 tokens/second.
- **Unified Representation:** Modality-specific projections map encoded tokens into the transformer's hidden dimension.

Unified Action Representation and Tokenization

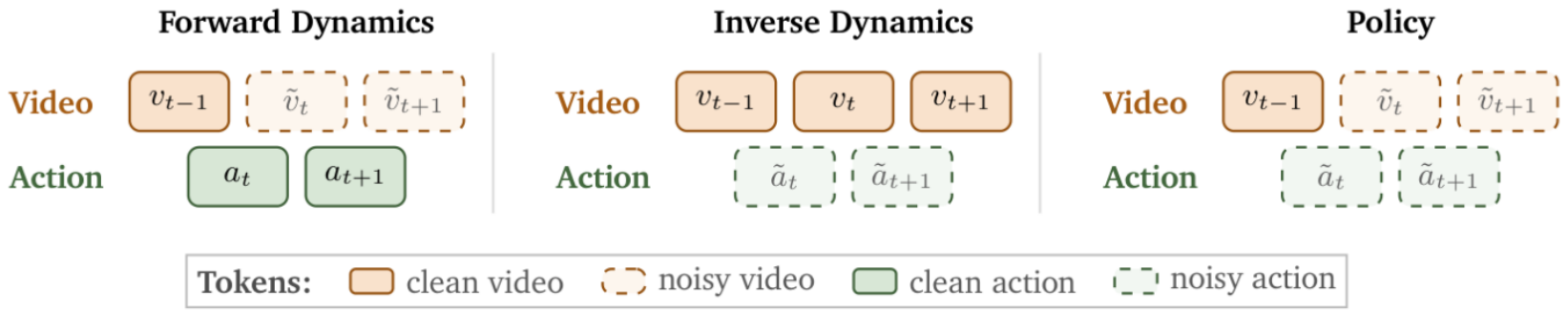


- **Goal:** Represent actions across different embodiments (robots, vehicles, cameras, human motion) in a shared latent space.
- **Three action components:**
  - **Ego pose:** Motion of the agent's observation frame.
  - **Effector pose:** Motion of manipulators (e.g., robot arms).
  - **Grasp state:** Current hand/gripper configuration.
- **Action tokenization:** Embodiment-specific linear projections map action vectors into a shared transformer representation and decode them back.

Token Arrangement and Generation modes

- **Unified token sequence:** All modalities are organized into two subsequences:
  - **Autoregressive (AR):** Language and ViT-encoded vision tokens for understanding and reasoning.
  - **Diffusion (DM):** VAE-encoded vision, audio, and action tokens for generation.
- **Token ordering:** AR tokens precede DM tokens; within DM, clean conditioning tokens precede noisy target tokens.
- **Task-specific generation:** Different input/output configurations enable:
  - **Text-to-Image / Text-to-Video:** Generate images or videos from language.
  - **Image-to-Video / Video-to-Video:** Generate video conditioned on initial frames.
  - **Video Transfer:** Generate RGB video conditioned on control inputs (e.g., depth or edges).

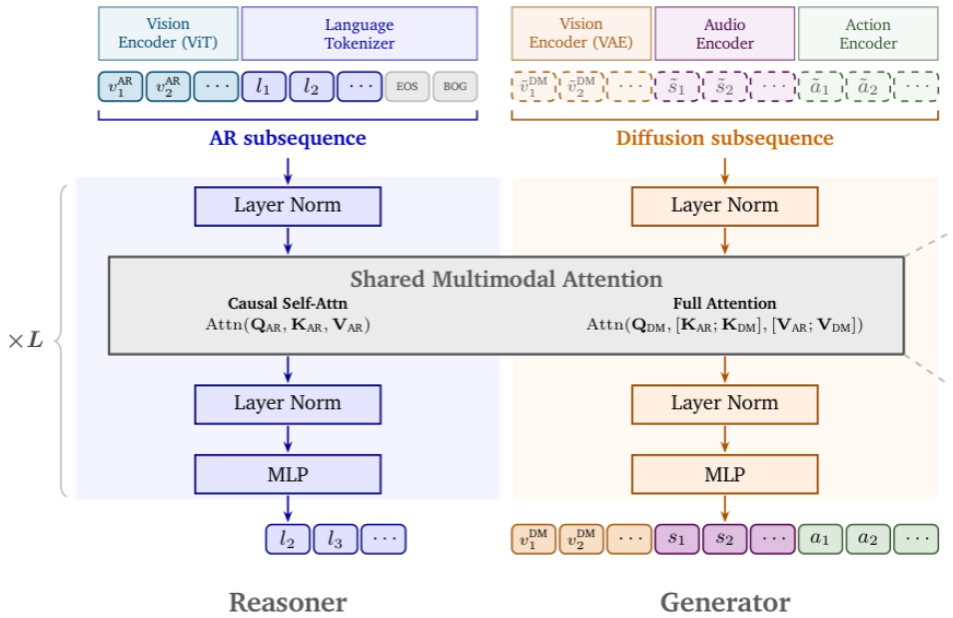
Action Generation Modes



- **Forward dynamics:** Predict future video given actions.
- **Inverse dynamics:** Infer actions from observed video transitions.
- **Policy:** Jointly predict future video and actions.

Source: Fig. 4 and Sec. 2.2.2 (PDF p. 10)

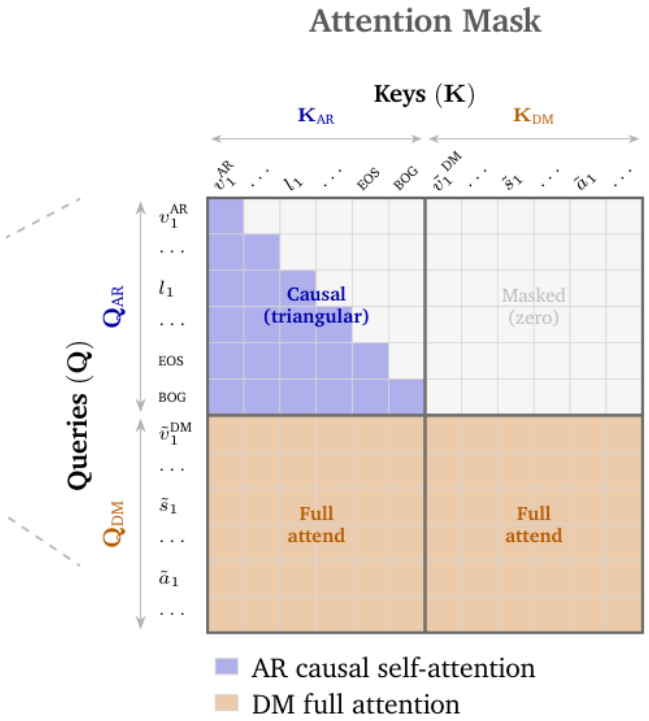
Mixture-of-Transformers (MoT) Architecture



- **Dual-tower design:** Reasoner (AR) and Generator (DM) use distinct parameters.
- **Shared attention:** Both streams interact through one attention operator.
- **Initialization:** Both towers begin from pretrained VLM weights.
- **Outputs:** Next-token predictions (AR) and denoised latents (DM).

Source: Fig. 5 and Sec. 2.3 (PDF p. 11)

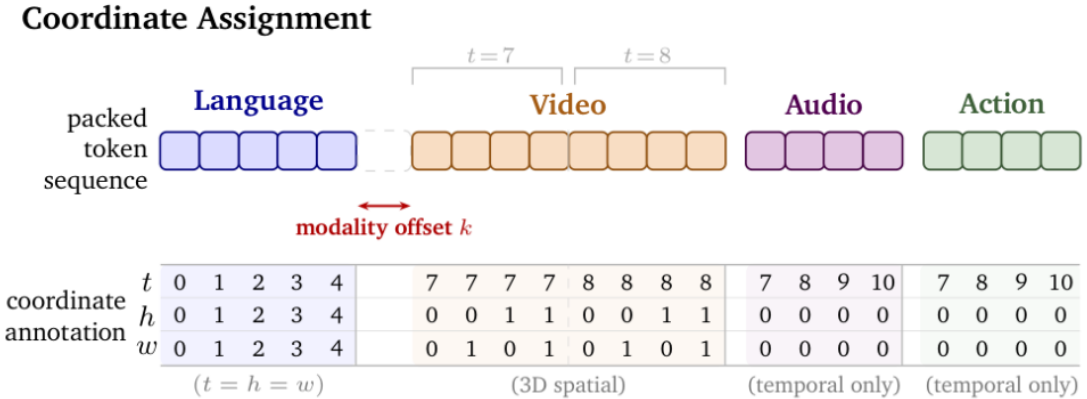
Dual-Stream Joint Attention

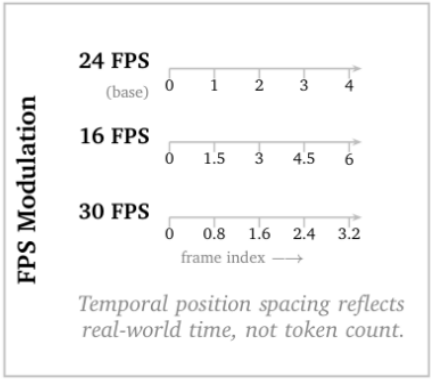


Reasoner: causal self-attention
Generator: full attention to both streams
DM can attend to AR; AR cannot attend to DM.

Source: Eqs. (7)–(8), Fig. 5 and Sec. 2.3.2 (PDF pp. 11–12)

Multimodal Positional Embeddings





- **3D MRoPE:** Assign temporal, height, and width coordinates to tokens.
- **Temporal alignment:** Video, audio, and action use a shared physical-time axis.
- **FPS modulation:** Rescales time steps when modalities have different rates.

Source: Fig. 6, Eq. (9) and Sec. 2.4 (PDF pp. 12–14)

Training Objectives: Reasoner vs. Generator

REASONER (AR)

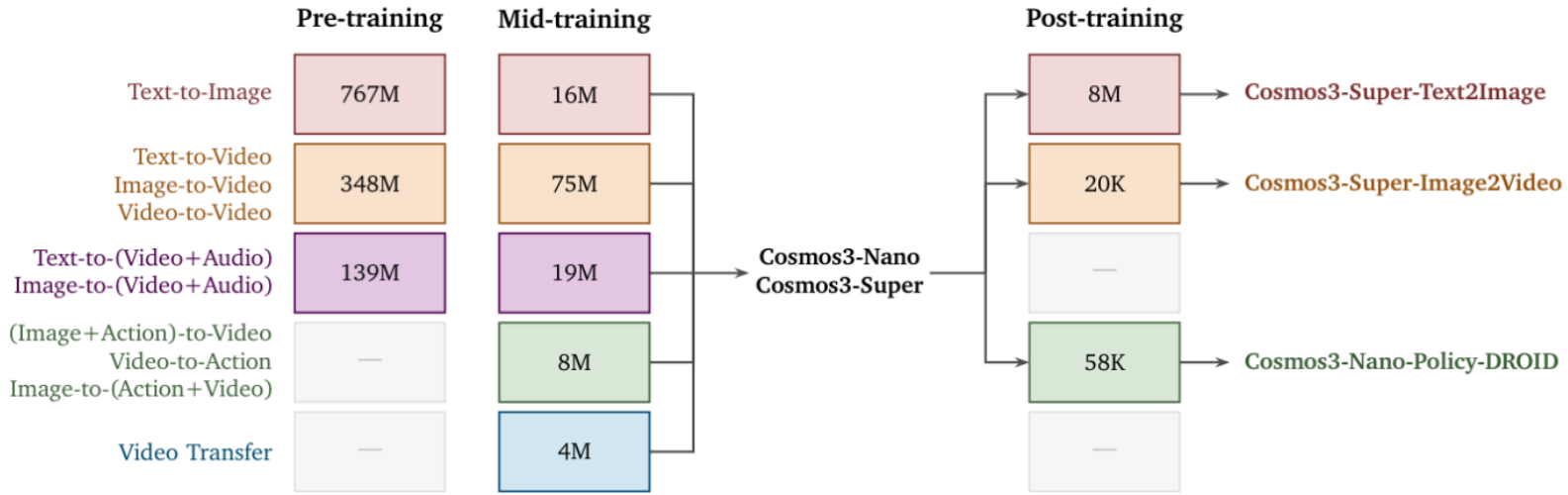
- **Objective:** Next-token prediction.
- **Data:** Vision–language pairs.
- **Curriculum:** Broad pre-training, then Physical AI SFT.

GENERATOR (DM)

- **Objective:** Rectified flow matching.
- **Data:** Vision, audio and action latents.
- **Training target:** Predict velocity from noise toward clean data.

Source: Secs. 4.1–4.2 (PDF pp. 26–27). Flow-matching loss excludes clean conditioning tokens.

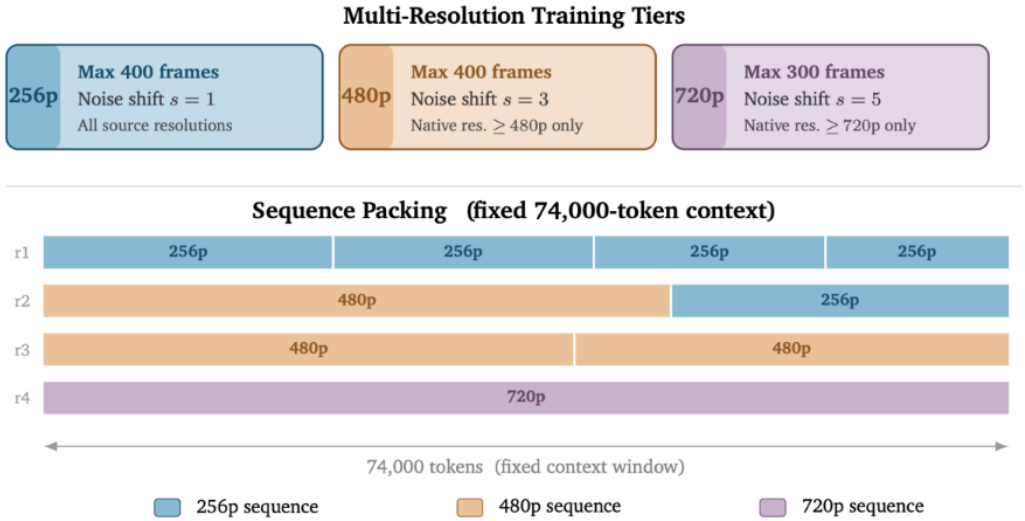
Progressive Multimodal Training

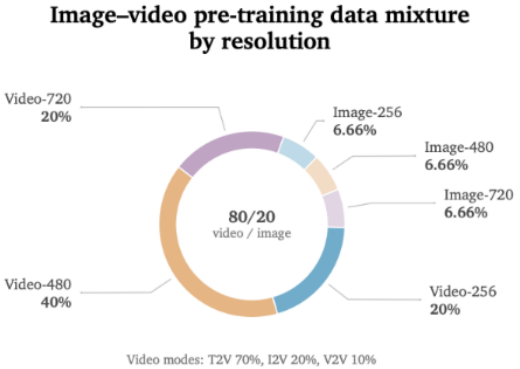


- **Pre-training:** Image, video and audio generation.
- **Mid-training:** Add action modes and control-conditioned video transfer.
- **Post-training:** Specialize T2I, I2V and robot policy on targeted data.

Source: Fig. 8 and Sec. 3.2 (PDF p. 20)

Multi-Resolution Generator Training

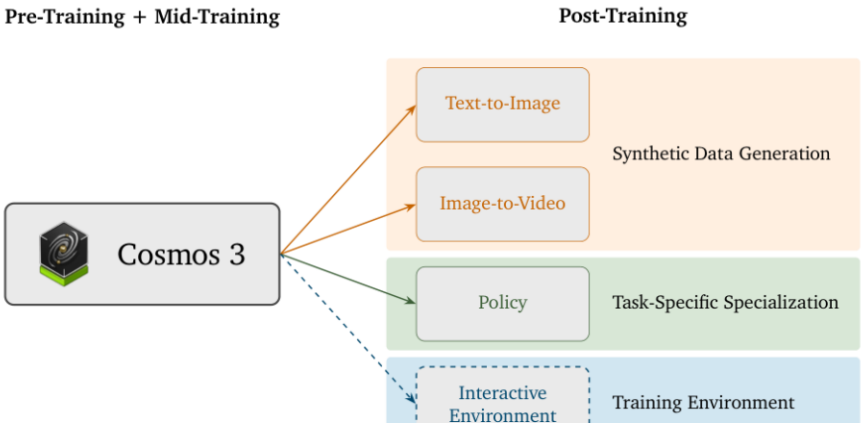




- **Resolution tiers:** 256p, 480p and 720p with variable frame counts.
- **Sequence packing:** Fit different sequences into a 74,000-token budget.

Source: Fig. 10 and Sec. 4.2.1 (PDF pp. 27–28)

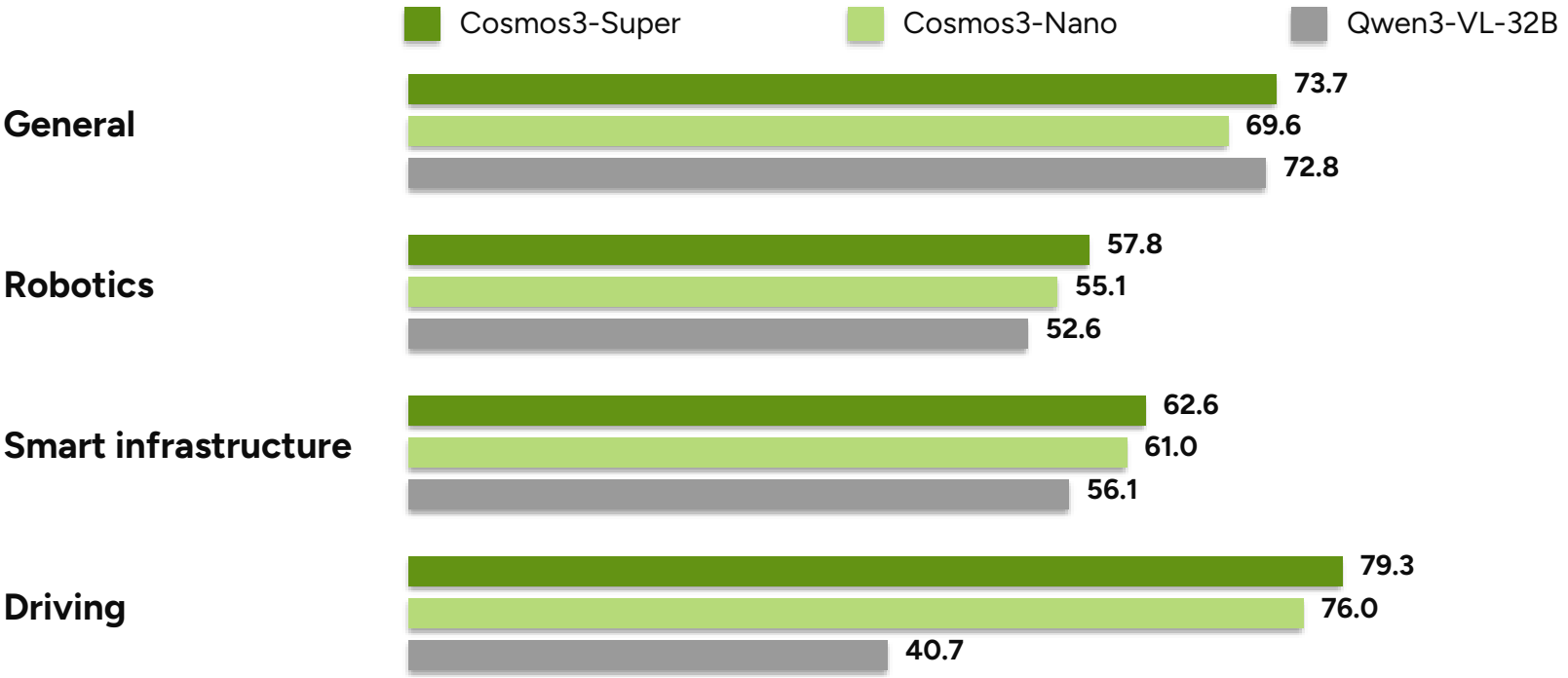
Post-Training for Downstream Tasks



- **Text-to-Image:** Super model refined for prompt adherence and visual quality.
- **Image-to-Video:** Super model specialized for ~8-second, 480p continuations.
- **Robot policy:** Nano model adapted on DROID for closed-loop actions.

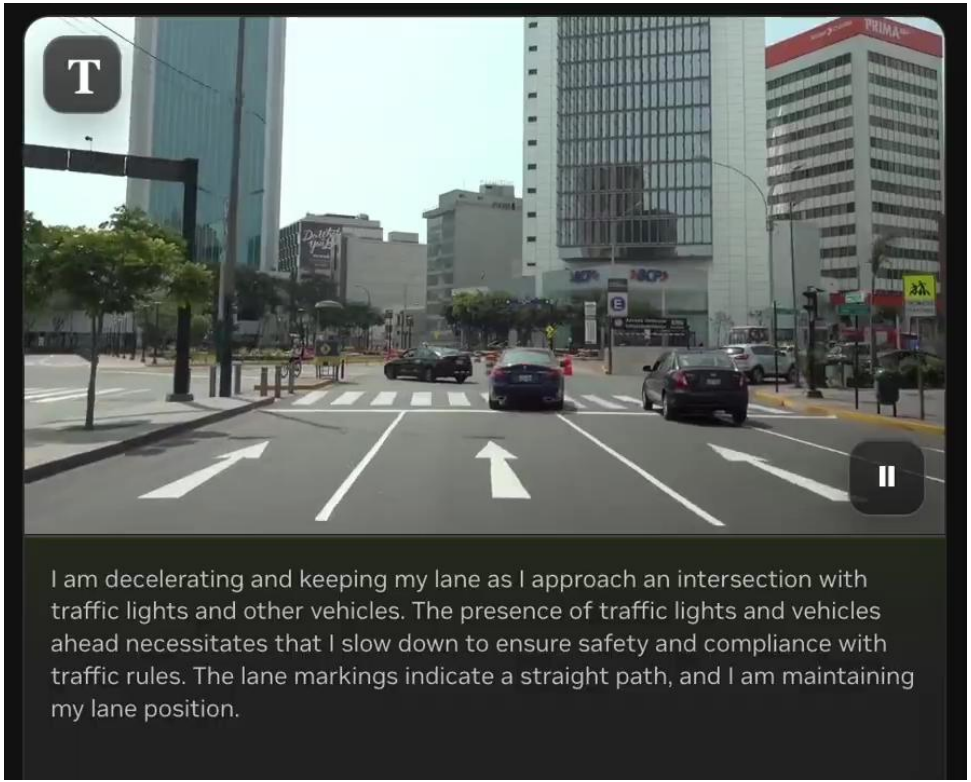
Source: Fig. 2 and Secs. 4.2.3–4.2.5 (PDF pp. 6, 30–32)

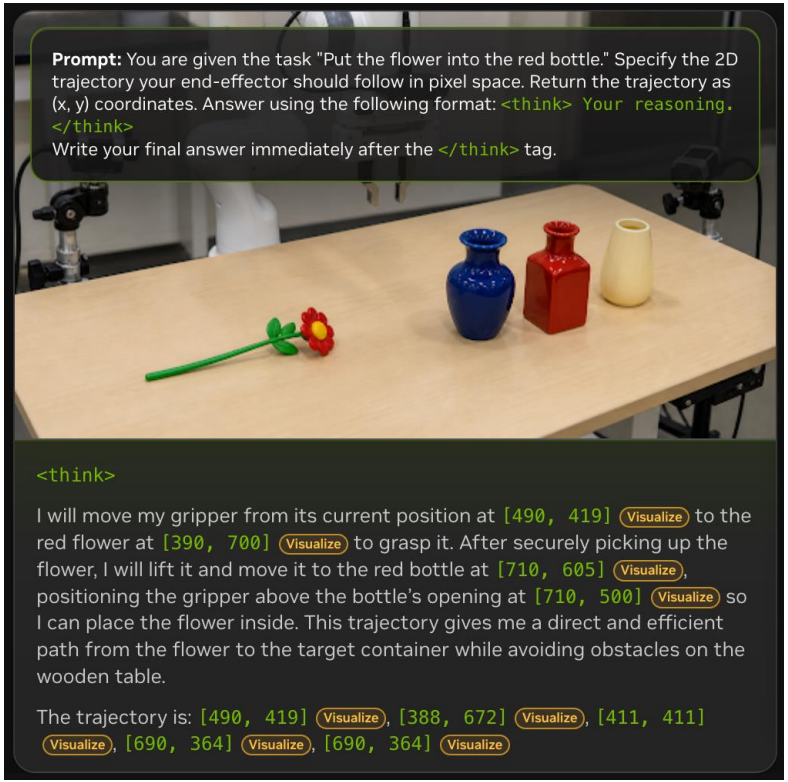
Results: Multimodal Understanding



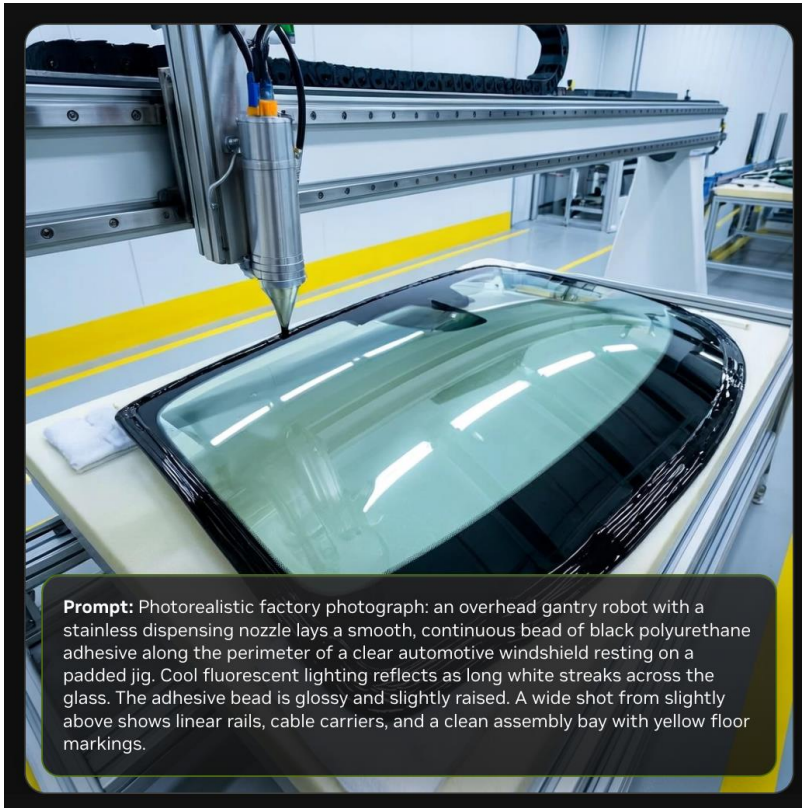
Source: Table 10 (PDF p. 51). Category averages over different benchmark sets; higher is better.

Results: Vision-Language Reasoning



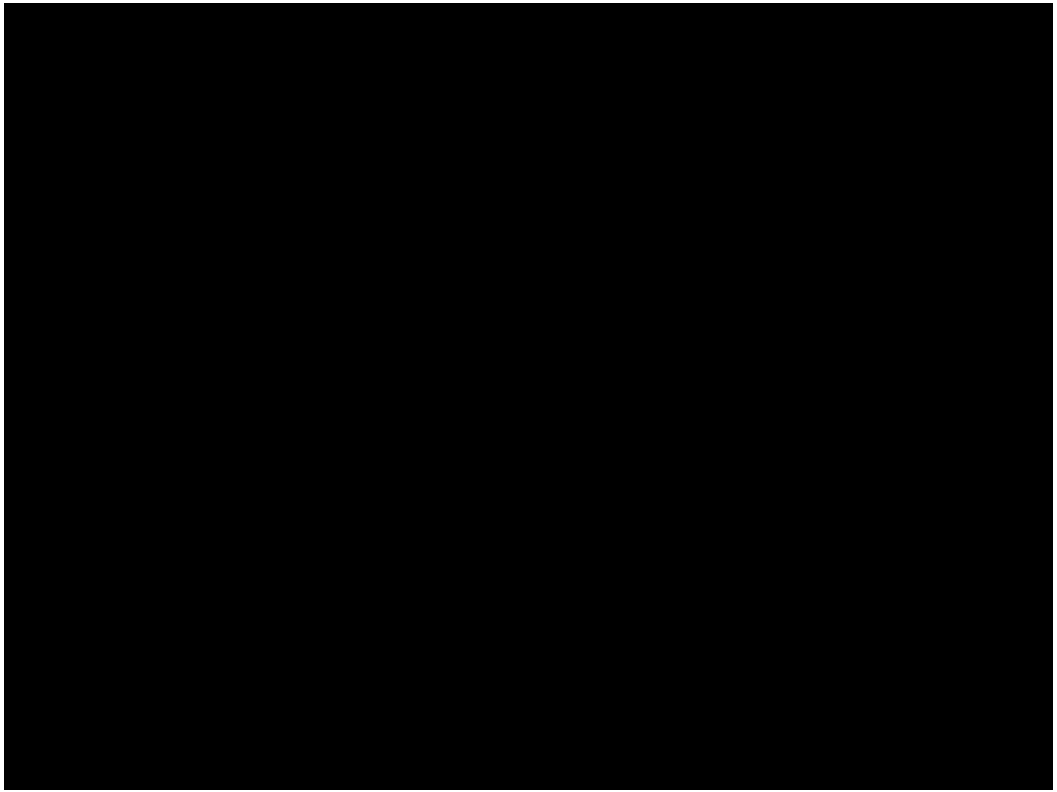


Results: Text-to-Image Generation



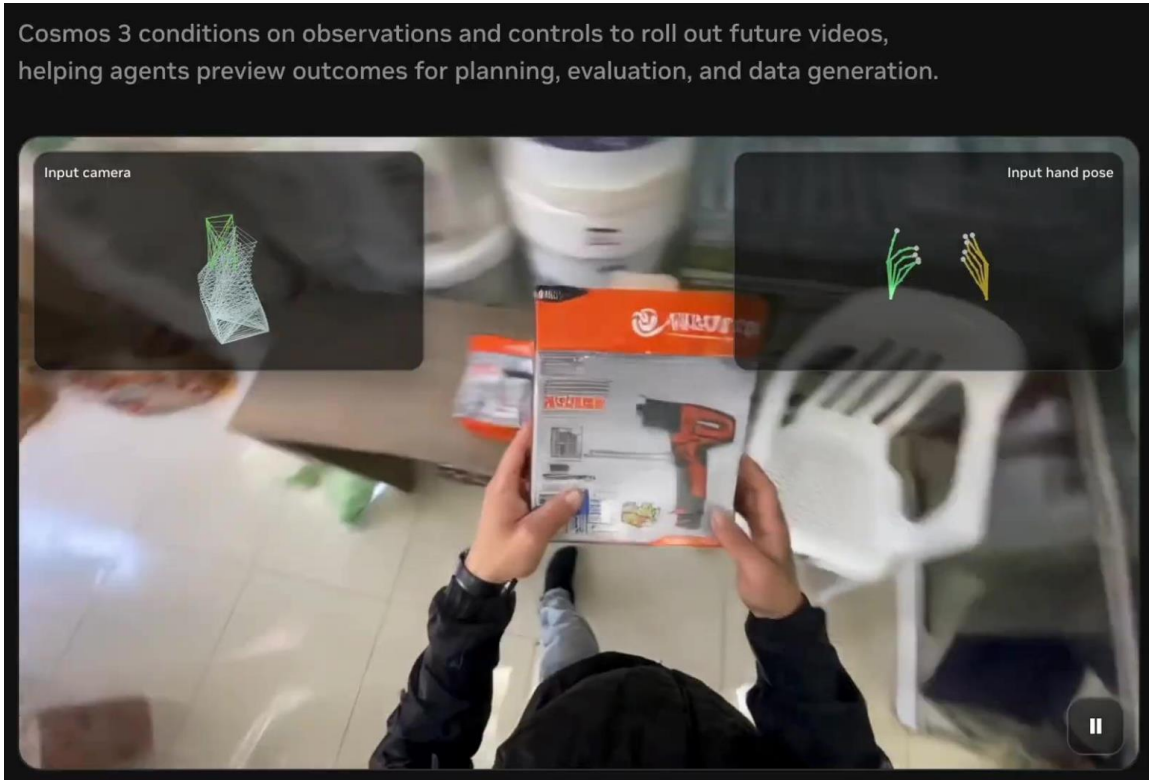


Results: Audio-Video & Video Transfer



Source: Figs. 20–21, Tables 15–16, Secs. 6.2.3–6.2.4 (PDF pp. 61–64)

Results: Forward & Inverse Dynamics



Forward dynamics

Source: Fig. 24, Table 18 and Sec. 6.2.5 (PDF pp. 65–68); MT-init and PT-init use matched downstream recipes.

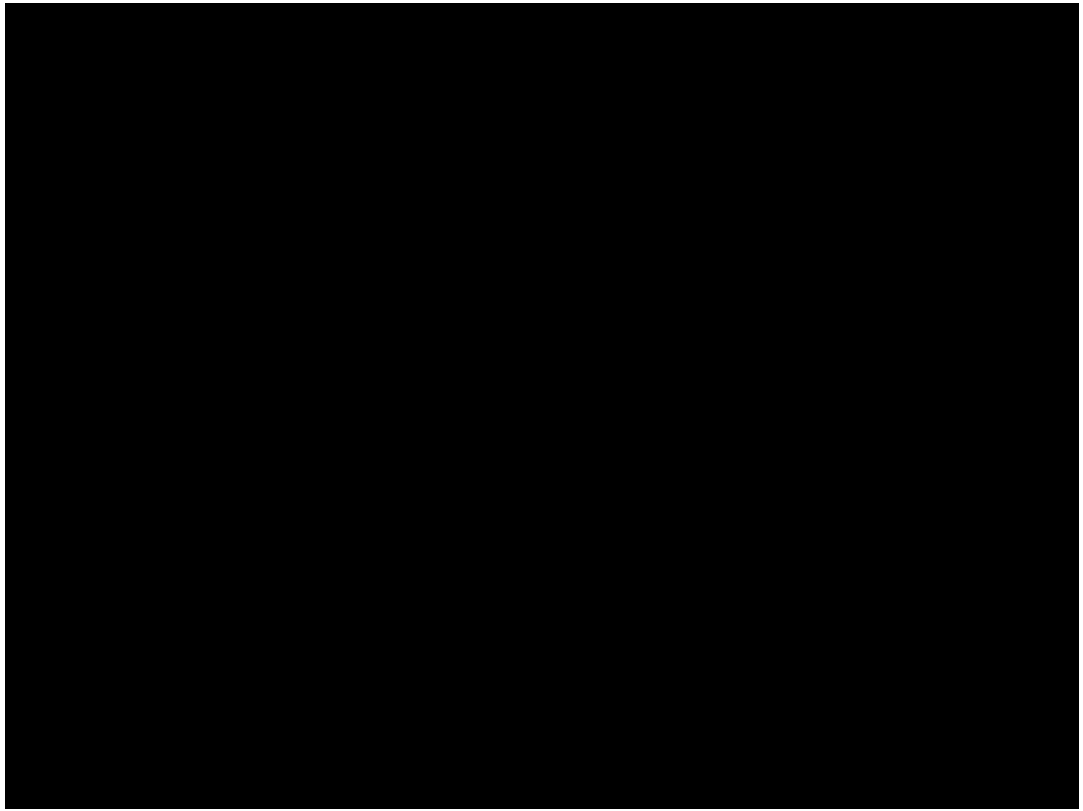
Results: Forward & Inverse Dynamics



Inverse dynamics

Source: Fig. 24, Table 18 and Sec. 6.2.5 (PDF pp. 65–68); MT-init and PT-init use matched downstream recipes.

Results: Robot Policy



Source: Fig. 24, Table 18 and Sec. 6.2.5 (PDF pp. 65–68); MT-init and PT-init use matched downstream recipes.

The paper still had internal separation between its reasoner and its generator: how can you foresee those boundaries being blurred/broken down even further?

As understanding and generation become more closely connected, do you think future models will move toward a fully shared representation, or will separate specialized pathways remain more effective?