

---

---

# Integrating Physical Property Reasoning into Vision-Language Models for Embodied Perception

Andreas Ilias Alexopoulos  
15.05.2025

---

---

---

# Content

Introduction

Models

Real World Gap

Outline & Conclusion

---

# The Challenge: “Fetch the Water”

---

# Meet RoboBERT

*“lift a full glass of water from the table and slot it safely onto a shelf”*



[5,6]

# What do we need to take into consideration?

Glass is fragile → may shatter under high grip force

Filled with water → spills if tilt  $> 12^\circ$

Condensation makes it slippery

It sits 2 cm from the edge → tiny nudge = disaster



[5]

---

---

# Research Question

What are different ways to integrate physical property into a VLM?

---

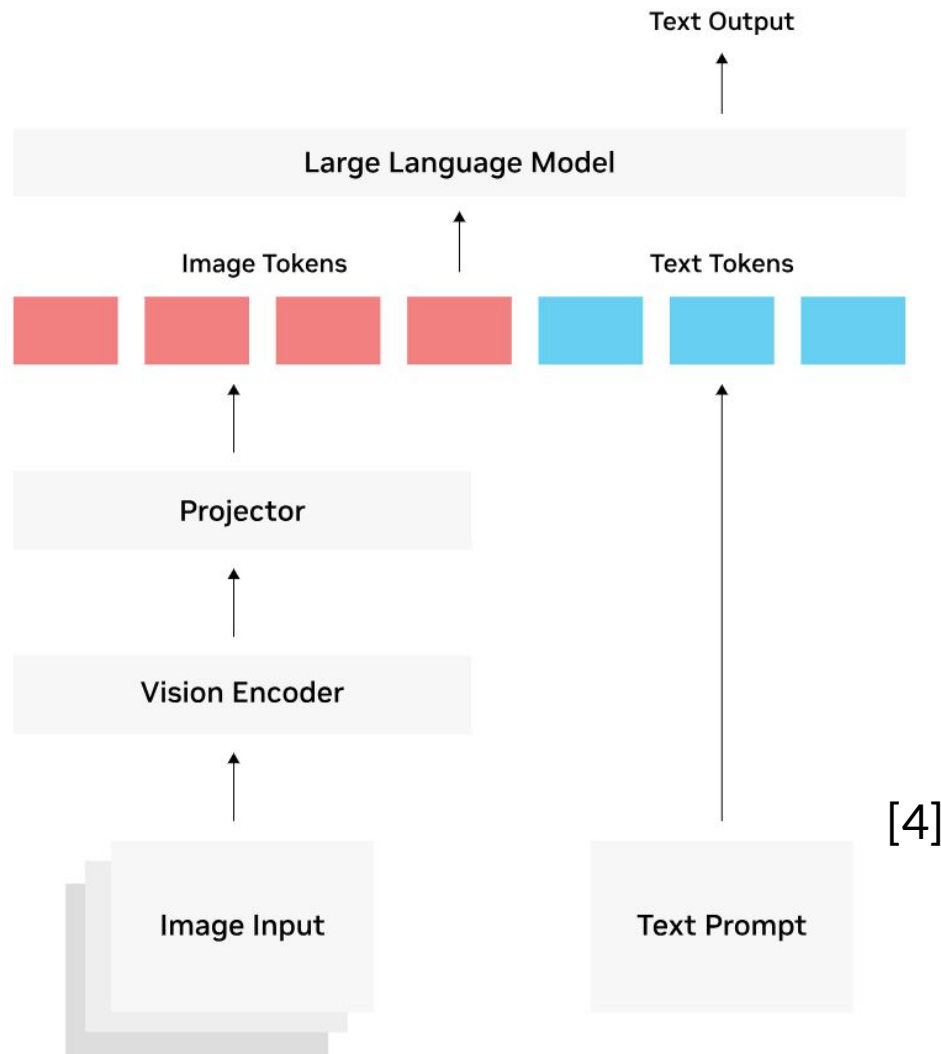
# Basics

---

# What is a VLM ?

Multi-Modal Input

Corresponding Text Output





---

# Model Explanation

---

## Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning



COGVIDEOX: TEXT-TO-VIDEO DIFFUSION MODELS WITH AN EXPERT TRANSFORMER

---

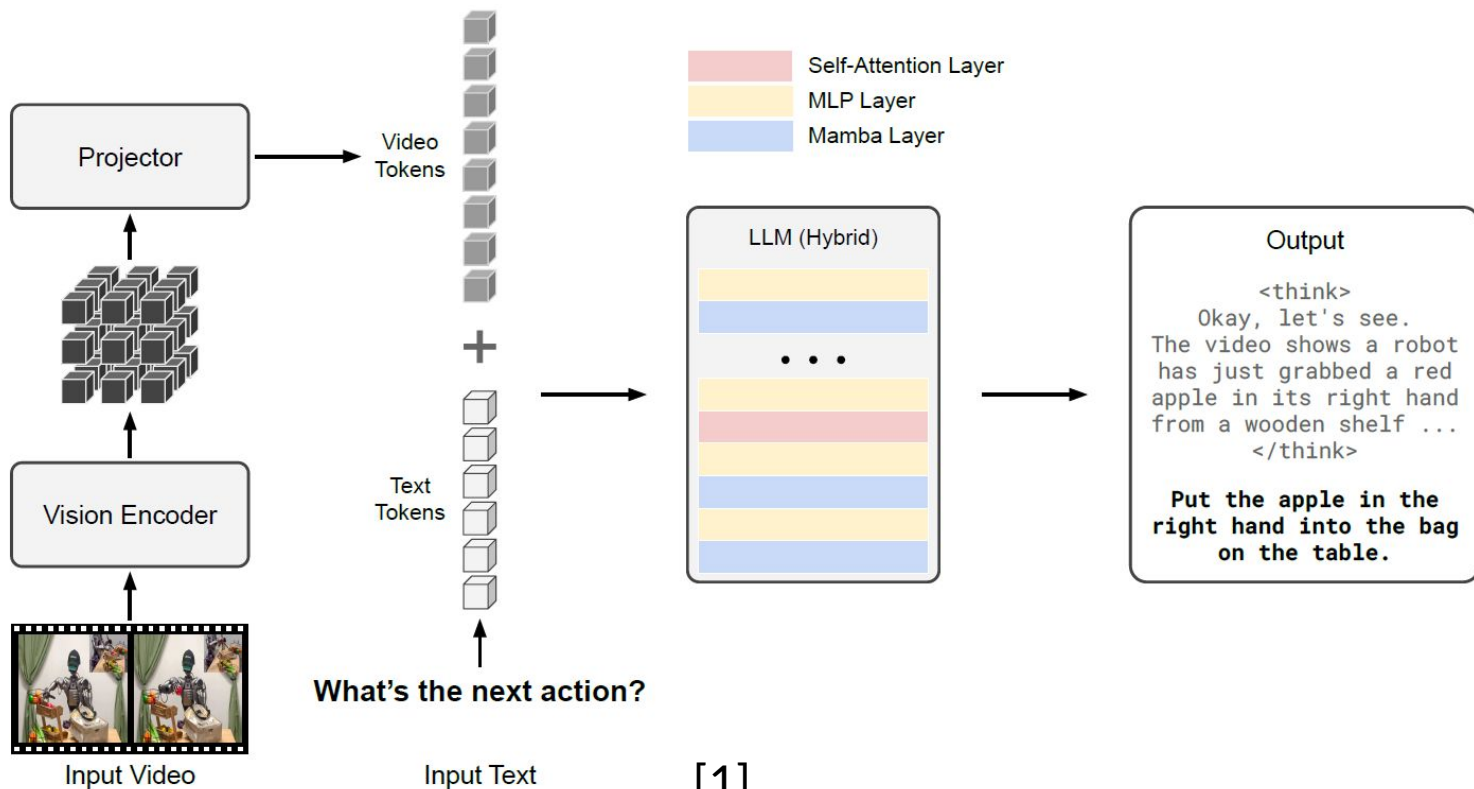
**VIDEOPHY: Evaluating Physical Commonsense for Video Generation**

---

---

# COSMOS-Reason 1

# Overall Architecture

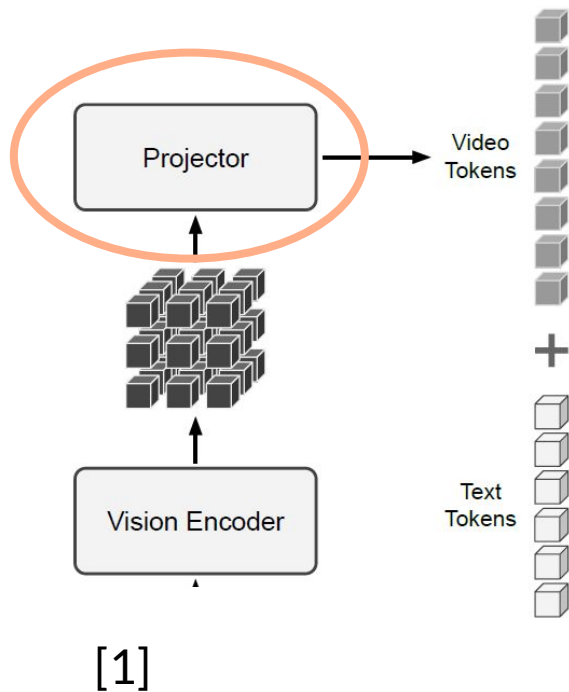


[1]

# Pipeline

1. Vision Pre Training
2. General SFT
3. Physical SFT

# Vision pre-training



ensures that the visual tokens are compatible with the text tokens

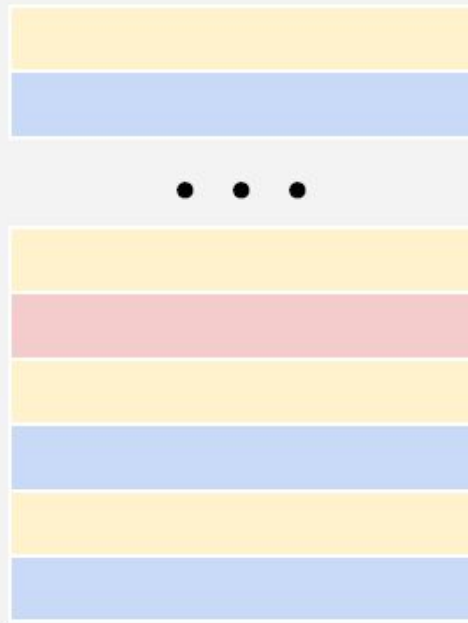
# General supervised fine-tuning

Projector

Vision Encoder

**Training WHAT is  
happening**

LLM (Hybrid)



# Training what is happening - Multi Modal Input

Dataset Type	Input	Expected Output	Purpose
Image-Text Pairs	Single images and corresponding text	Aligned visual and text tokens	Basic object and scene understanding
Cause and Effect	Initial physical state and action	Predicted outcome	Simple physical cause and effect
Video-Text Pairs	Video sequences and descriptive text	Aligned video frames and text tokens	context-aware reasoning



# Example

Image:

A single frame showing a robot arm reaching for an apple.

Text Prompt:

“A robot arm reaches for an apple on a table.”

# Training why something is happening

## General SFT:

- **Video:** A robot arm **picks up** an apple.
- **Text:** *"The robot arm picks up the apple."*

## Physical AI SFT:

- **Video:** A robot arm **pushes** a block, causing it to **fall** off the table.
- **Text:** *"The robot pushes the block, causing it to fall off the table and hit the ground."*

# Physical-AI supervised fine-tuning

Physical common-sense SFT

Embodied reasoning SFT

# Physical common-sense SFT

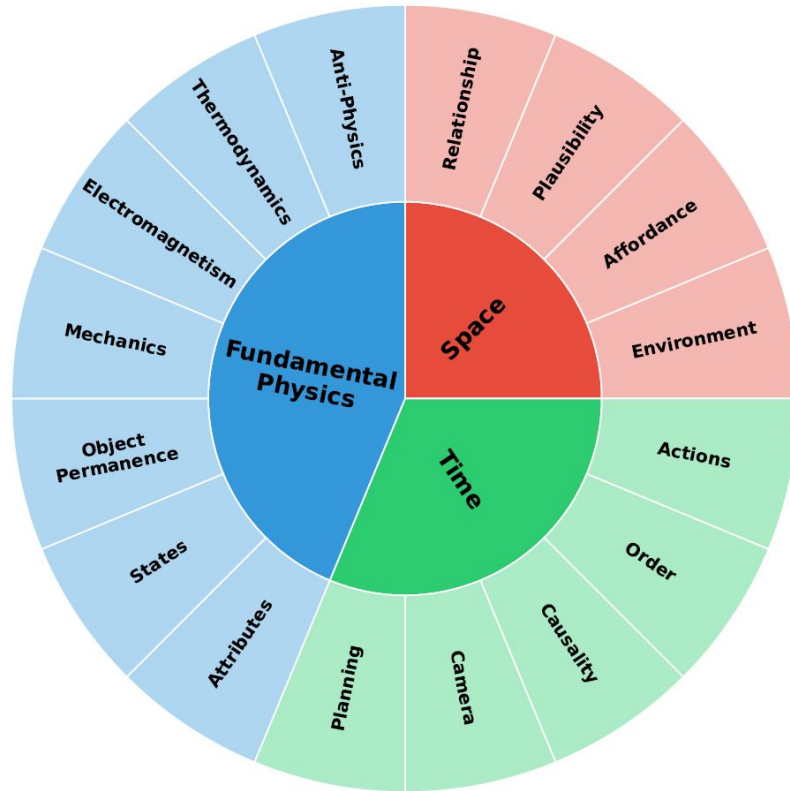
Understanding the basic rules of the physical world.

Reasoning about what will happen in a scene based on common sense

## Object Permanence Test:

### Training Data:

- **Video:** A ball rolling behind a wall and coming out the other side.
- **Question:** *“Where is the ball when it is not visible?”*
- **Answer:** *“The ball is behind the wall.”*



[1]

Space: Relationship	Determine the spatial relationship of objects in a scene. Perspective is important; for example, an object is left to the person or left to the camera view.
Space: Plausibility	Determine if a possible spatial relationship is feasible.
Space: Affordance	Understand object interaction with subjects such as humans, animals, robots, <i>etc.</i>
Space: Environment	Understand the scene or the surrounding environment.

# Embodied reasoning SFT

Understanding the intentions behind actions.

Reasoning about multi-step interactions and long-term consequences.

## Task-Oriented Sequence

- **Video:** A sequence showing a robot arm grasping, lifting, and stacking blocks.
- **CoT Reasoning Trace:**
  1. *“The robot arm identifies the first block.”*
  2. *“The robot arm grasps the block.”*
  3. *“The robot arm lifts the block.”*
  4. *“The robot arm positions the block above the second block.”*
  5. *“The robot arm releases the block, completing the stack.”*

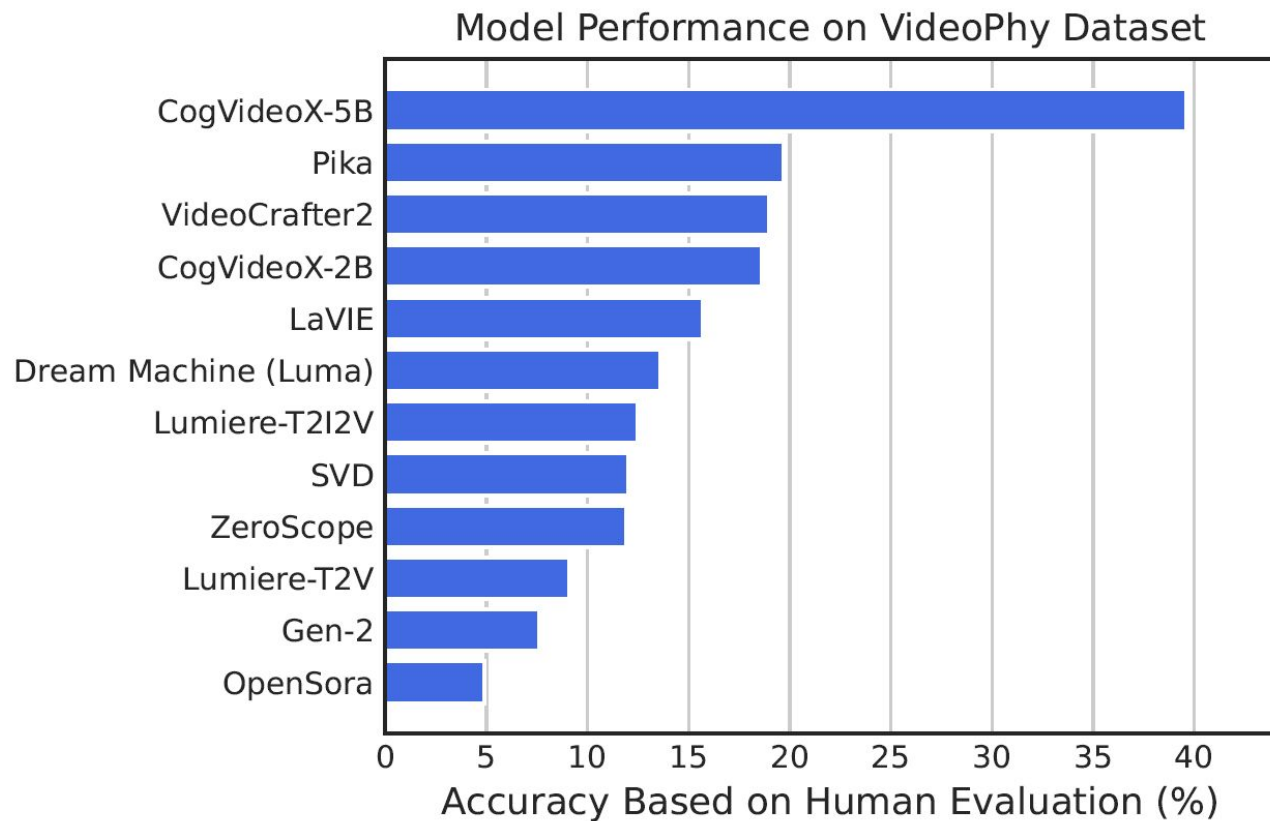
---

# CogVideoX

# What is CogVidoeX ?

a text-to-video generation model based  
on a diffusion transformer





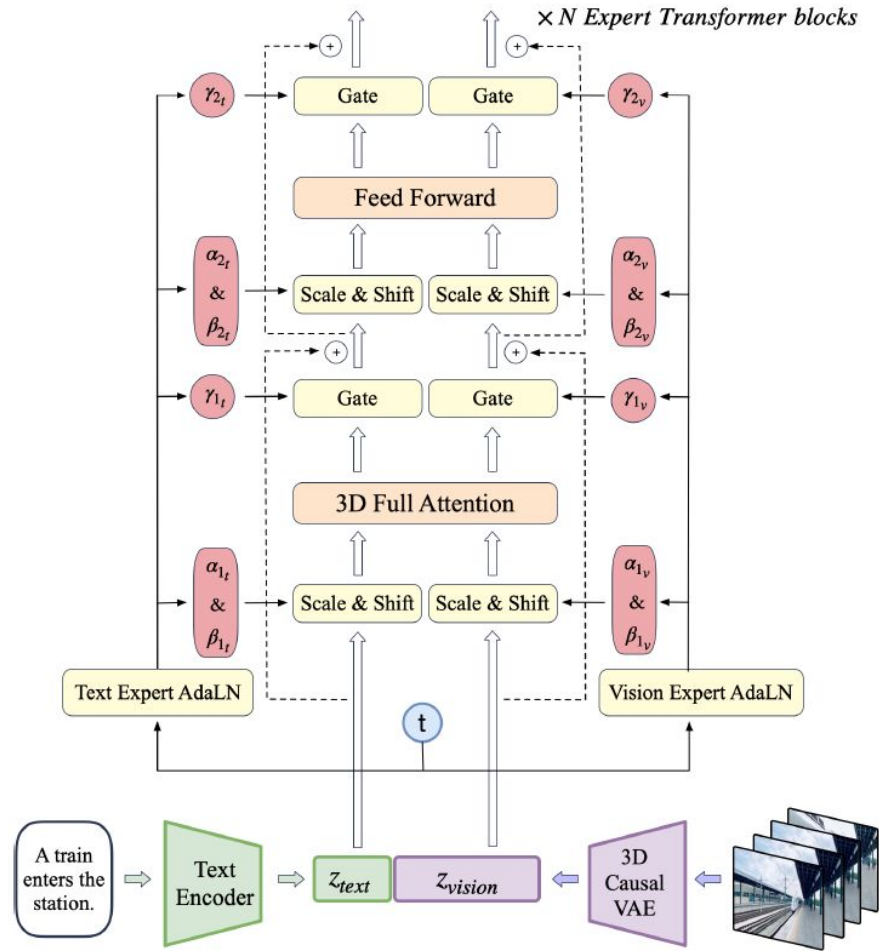
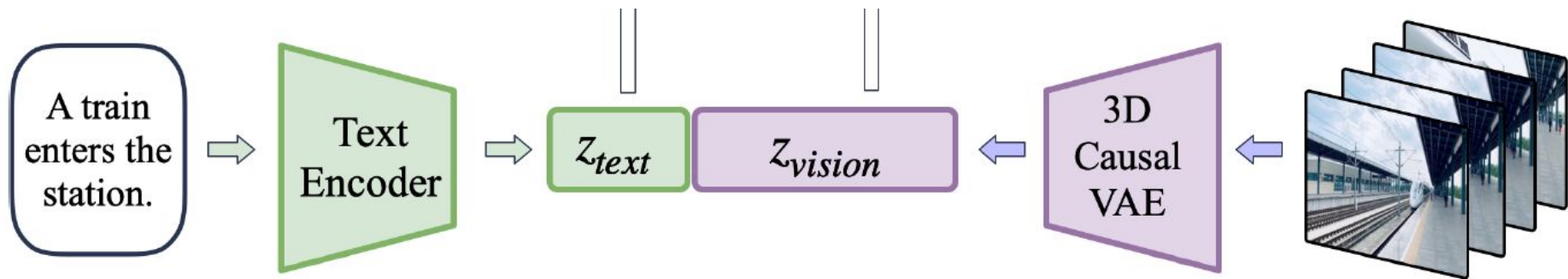


Figure 3: The overall architecture of CogVideoX. [3]

# Input Encoding

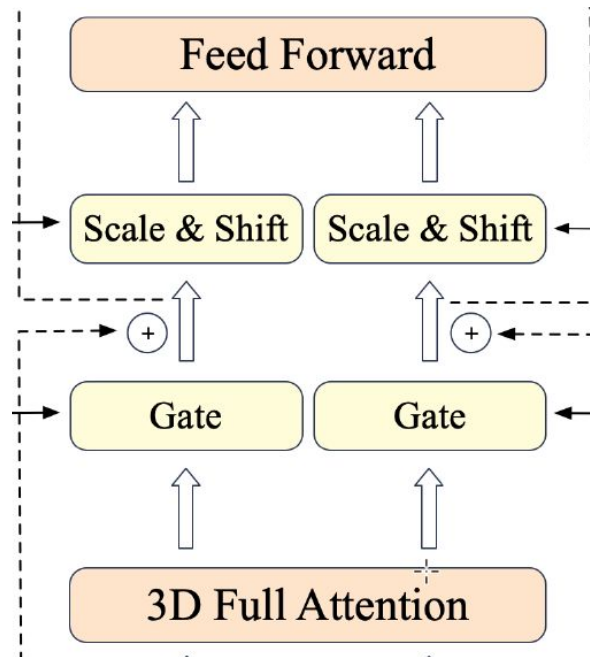


$(300, 256, 256, 3) \rightarrow (38, 32, 32, 128)$

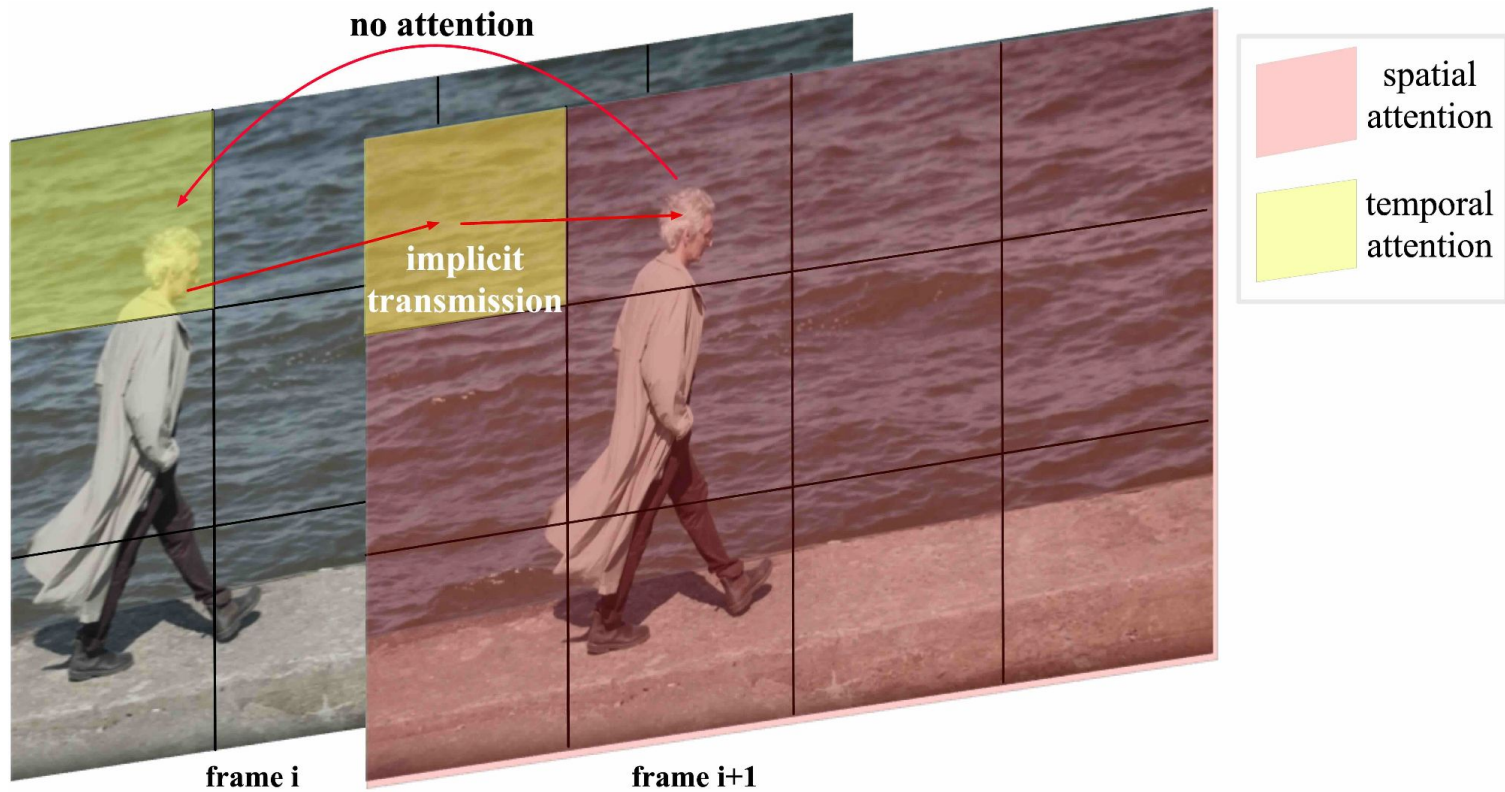
capturing spatial structure and  
temporal motion

# Expert Transformer Blocks

Understand deeper physical connections between frames

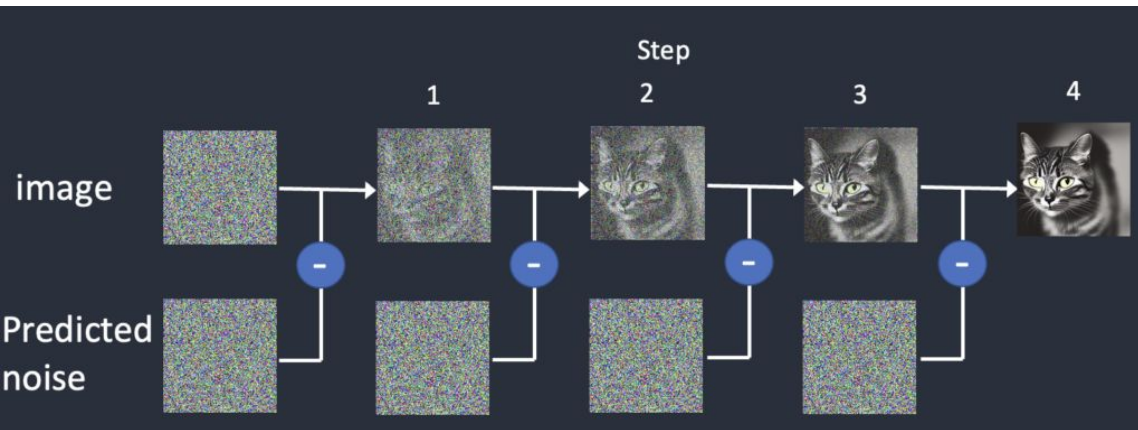


[2]



[2]

# Learning via Diffusion



Repeated over time with big video-text dataset

[2,7]

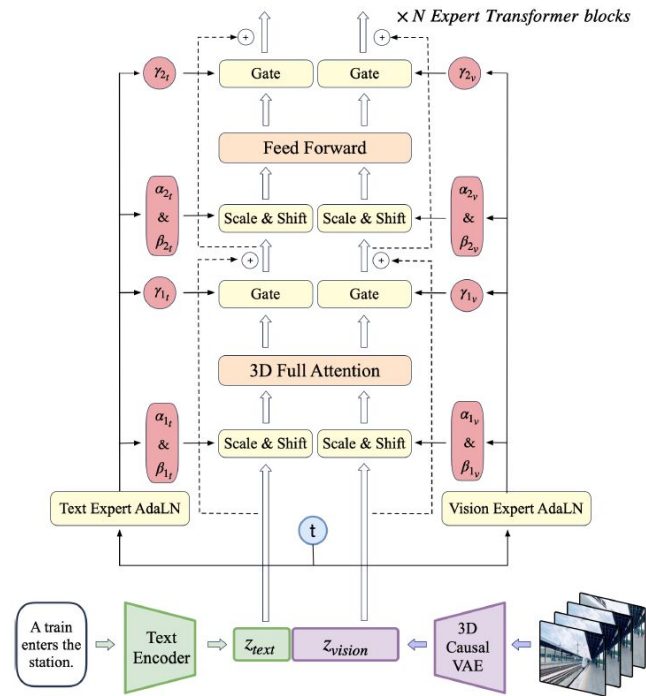


Figure 3: The overall architecture of CogVideoX.

---

# VideoPHY

# Caption Generation

Goal: Create a large set of physics-rich video prompts.

**Solid-Solid** (e.g., "Dominoes toppling over each other")

**Solid-Fluid** (e.g., "Water splashing on a rock")

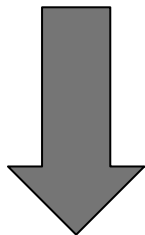
**Fluid-Fluid** (e.g., "Ink spreading in still water")

Output: 1000 raw captions representing a broad spectrum of physical phenomena



# Human Verification

Goal: Filter out unrealistic or physically inaccurate captions



Output: 688 high-quality, human-verified captions that capture realistic physical dynamics

# Difficulty Annotation

Goal: Capture complexity of simulating each interaction.

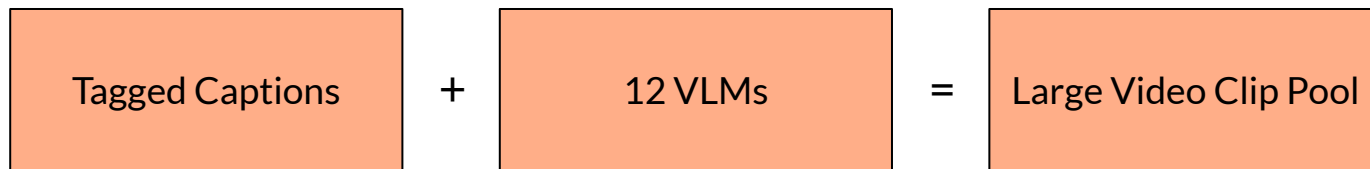
Rate by:

Easy: "A bottle topples off the table"

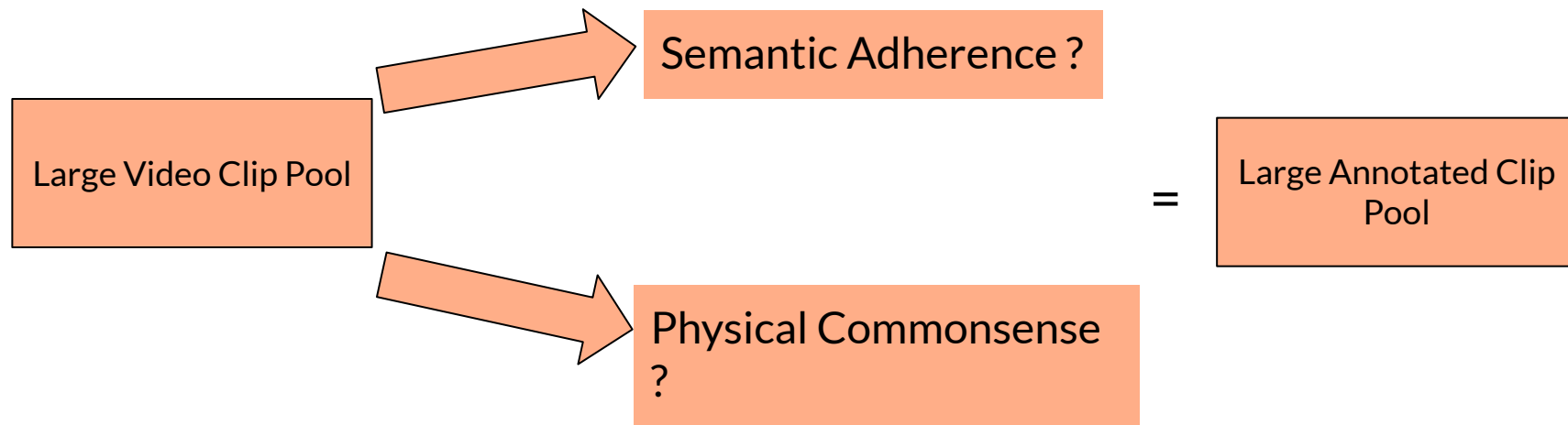
Hard: "A swimmer splashing in the sea"

Output: Final set of 688 captions, each tagged as "easy" or "hard"

# Video Generation



# Binary Labeling



Initial Frame

Middle Frame

Last Frame

LaVIE



**Prompt:** Pouring milk into a cup.

**Conversation of Mass Violation:**  
The level of the milk in the cup does not increase.

Gen-2



**Prompt:** Shoveling sand into a bucket.

**Newton's First Law Violation:** The sand in the front shifted without any force by shovel.

OpenSora

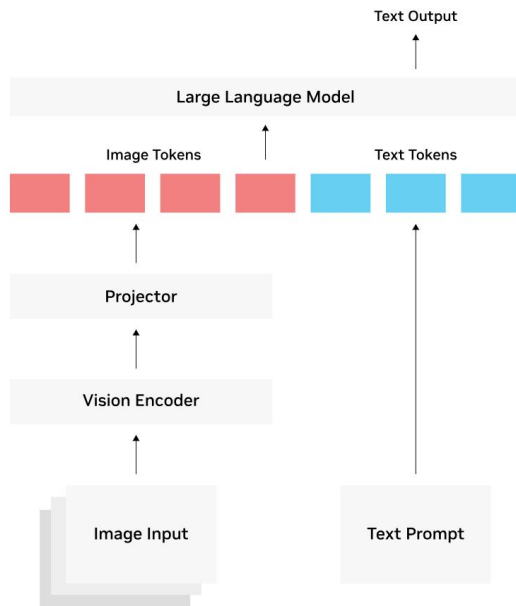


**Prompt:** Wood floats down a canal.

**Solid Constitutive Law Violation:**  
Rigid objects (wood) should not deform under small force load.

# VIDEOCON-PHYSICS

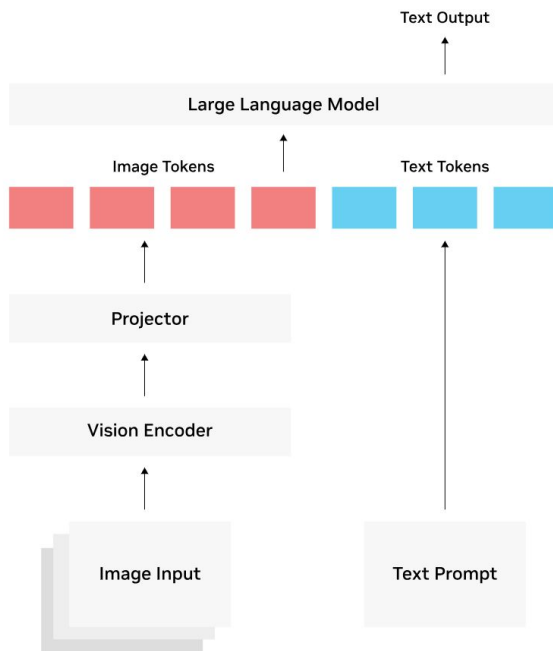
*“A baseline model capable of understanding video-text relationships but without any explicit physical reasoning”*



# Finetuning

Large Annotated Clip  
Pool

+



=

improved semantic  
adherence and  
physical commonsense

[4]

---

# Methodology Comparison

---



# Recap

Model	Approach	Pros	Cons
Cosmos-Reason	Ontology-based learning with structured SFT and RL	Deep physical understanding, long-term reasoning	High computational cost, rigid ontology
CogVideoX	Diffusion-based, expert transformer for video denoising	High-resolution outputs, temporal coherence	High complexity, potential for overfitting
VideoPhy	Dataset-centric, human-verified physical prompts	Scalable evaluation, broad dataset coverage	Annotation costs, limited flexibility

---

# Real World Gap

# Real-World Gap for Physical Knowledge in VLMs

1. Lack of Intuitive Physical Understanding
2. Limited Real-World Data Representation
3. Poor Generalization to Unseen Scenarios

---

# Conclusion & Outlook

# Conclusion

- Current approaches often rely on statistical patterns and predefined ontologies, missing deeper physical insights.
- Bridging this gap is critical for advancing AI systems capable of interacting with the real world.

# Outlook

Real-World Embodied Learning

Long-Horizon Training

Automated Physical Benchmarking

# References

- [1] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning - <https://arxiv.org/abs/2503.15558v2>
- [2] CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer - <https://arxiv.org/abs/2408.06072v3>
- [3] VideoPhy: Evaluating Physical Commonsense for Video - <https://arxiv.org/abs/2406.03520>
- [4] <https://www.nvidia.com/en-us/glossary/vision-language-models/>
- [5] <https://www.yahoo.com/lifestyle/does-a-glass-of-water-ever-go-bad-experts-weigh-in-151447901.html>
- [6] <https://www.bmwgroup.com/de/news/allgemein/2024/humanoide-roboter.html>
- [7] <https://stable-diffusion-art.com/how-stable-diffusion-work/>

## EMBODIED REASONING BENCHMARK



Based on the ego vehicle's actions in the video, what is the most likely immediate action it will take next?

A: turn right, B: turn left, C: change to right lane, D: change to left lane

## RESPONSE BEFORE RL

Okay, let's break this down. The video shows the ego vehicle is driving on a single-lane road with a 25 mph speed limit. The road has parked cars on the right, so the driver is maintaining a safe distance there. The left side is a two-way divider with oncoming traffic, which is coming from the left in the video.

Since the road is single-lane, changing lanes (right or left) might not be possible. But the video shows a single-lane road, which usually means one lane in each direction, so there's no shoulder to the right for parking, but parked cars are already on the right. Maybe the road is a single lane each way, but parked cars are on the right side of that lane. So the driver is in the correct lane, just keeping a safe distance from the parked cars. The oncoming traffic is in the opposite lane, so the driver is staying in their lane. Changing to the left lane would be into oncoming traffic. So that's not possible. Therefore, the answer must be D.



## RESPONSE AFTER RL

Okay, let's break this down. The user provided a driving scenario and asked for the most likely immediate action based on the ego vehicle's current actions. First, I need to recall the details given. The ego vehicle is on a narrow, two-way residential street with parked cars on both sides. The driver is maintaining a steady speed, staying in their lane, and being cautious because there's oncoming traffic in the opposite lane. The speed limit is 25 mph. The parked cars limit visibility, so they have to be extra careful. The road markings are double yellow lines, meaning no passing, so lane changes aren't allowed. The driver is keeping a consistent pace and staying vigilant.

Now, the question is about predicting the next immediate action. The options are turning right, left, changing to the right lane, or left lane. Let's consider each option.

Option A: Turn right. The video shows a two-way street. The road is narrow with parked cars on both sides, so turning right might not be possible.

Option B: Turn left. Similar reasoning as A. Unless there's a left turn ahead, but the video doesn't indicate any upcoming intersections or turns.

Option C: Change to right lane. The current lane is the right one since it's a two-way street. Changing to the right lane again doesn't make sense because there's already parked cars on both sides. So this option is probably not correct.

Option D: Change to left lane. Wait, but in a two-way street with double yellow lines, there's no left lane to change into. The driver is in their respective lane (right for their direction), and the opposite lane is for oncoming traffic. So changing to the left lane would be into oncoming traffic, which is dangerous and not allowed. So this option doesn't fit either.

Wait, maybe the question is a trick. The video shows the road markings are double yellow lines, which means no passing, so changing lanes is prohibited. Therefore, the driver cannot change lanes, so options C and D are impossible. The only remaining options are turning right or left. But since the video doesn't show any turns, and the driver is maintaining their path, perhaps the answer is that none of the options are correct, but since we have to choose, I have no choice but to select none. Therefore, the answer is none.