Video World Models

Learning the Physical World from Videos

Zeyu Zhang

Talk @ Zhejiang University, Oct 17, 2025

Some Quotes

"I think world modeling is a really promising research direction to explore, **but** *one thing I've found in some of my recent research is that it's very easy to generate great-looking videos that make critical modeling errors... Better downstream performance is really the gold standard for evaluation."*

— *Chelsea Finn*

What Are Video World Models?

"One approach to tackle the high visual complexity of the world is to learn an action conditioned video prediction model... Learning a world model to predict the outcomes of potential actions enables planning in imagination, reducing the amount of trial and error needed in the real environment."

— *Pieter Abbeel*

"Video world models try to develop predictive models of what future perceptual input will be, based on a current observation we have."

— *Chelsea Finn*

Learn from Videos

The core idea of a video world model is to learn from videos (usually large-scale, unlabelled, sequential visual data) in order to build an internal predictive model of the world — one that captures spatial, temporal, and causal structure. There are following perspectives:

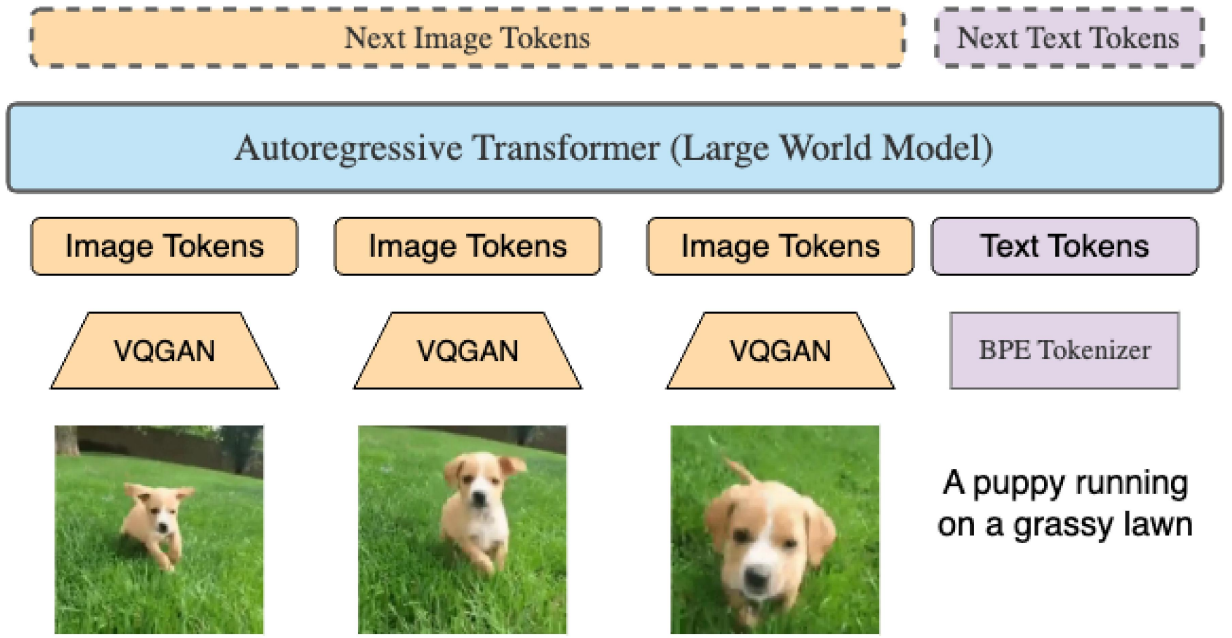
- Understanding complex, long-form videos.
- Predicting future states in video frames.
- Physical interaction through video prediction.

Challenges in Learning from Long Videos

There are two significant challenges in learning from long videos:

- **Context size constraints**: Most prior vision, video, and multimodal models are limited in how many tokens or frames they can attend to (short context windows). This limits their ability to reason over long videos (e.g. hours) or long documents. And attention mechanisms have quadratic complexity, so naively scaling to very long sequences is infeasible.
- **Efficiency**: Training on extremely long sequences results in high memory usage, instability, and computational cost, which become major bottlenecks when scaling to millions of tokens or frames.

Learn from Million-Length Videos: Large World Model (LWM)



LWM argues that integrating video and language into a unified, autoregressive world model provides more flexibility (e.g. image/video/text generation, question answering over long videos). Blockwise RingAttention (ICLR 2024) divides a long sequence into blocks and connects them in a circular ("ring") pattern so that each block attends to its neighbors, enabling efficient exact attention over million-token contexts with **linear** memory growth.

Pieter Abbeel et al. *World Model on Million-Length Video and Language with Blockwise RingAttention* (2025)

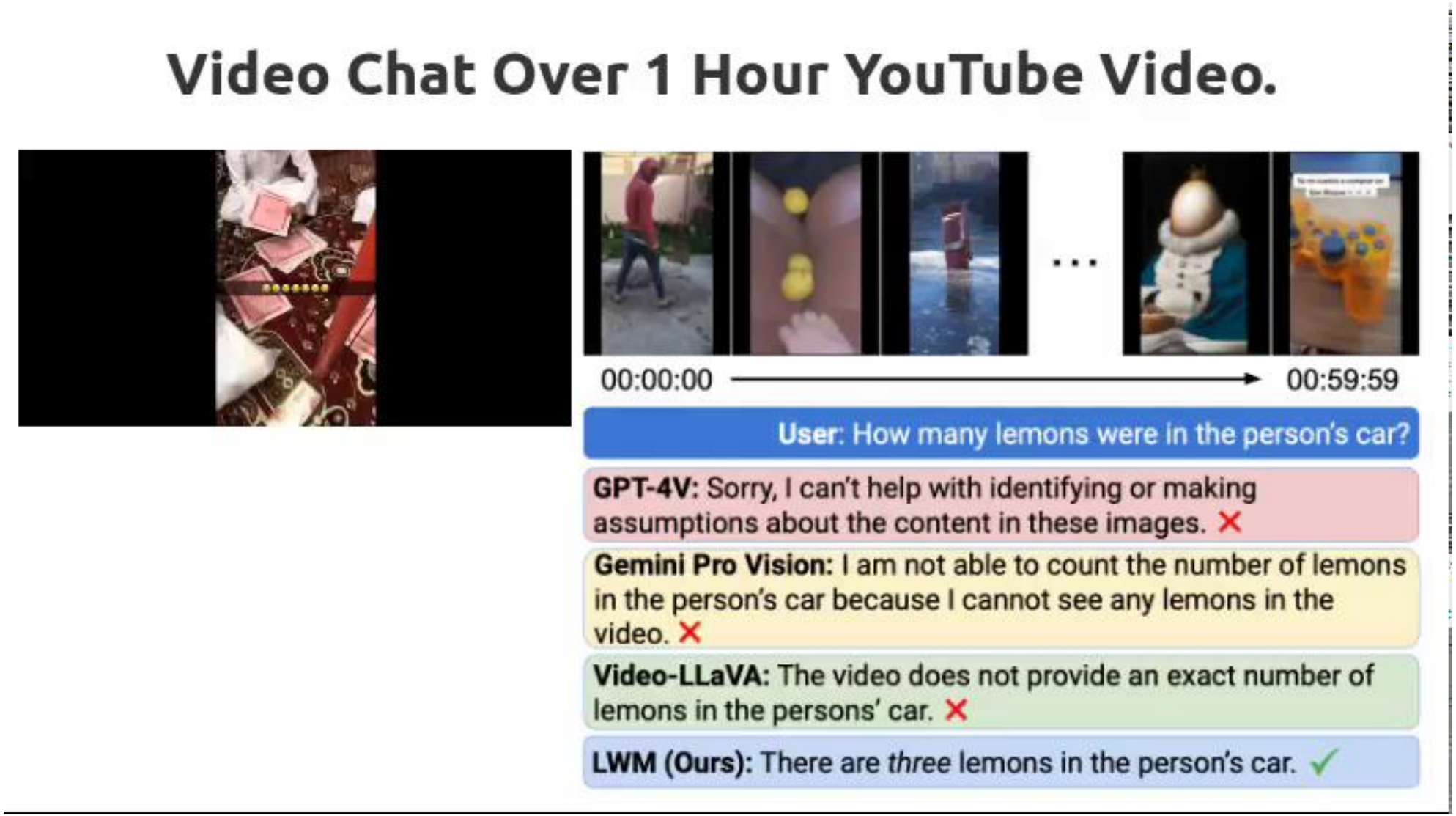
T2V Results (V2V as well and more significant)

Fireworks exploding in the sky | Waves crashing against the shore | A bustling street in London with red telephone booths and Big Ben in the background | Camera pans left to right on mago slices sitting on a table

A ball thown in the air | Slow motion flower petals falling on the ground | A burning campire in a forest | A boat sailing on a stormy ocean

Figure 5. Text to Video. LWM generates videos based on text prompts, autoregressively.

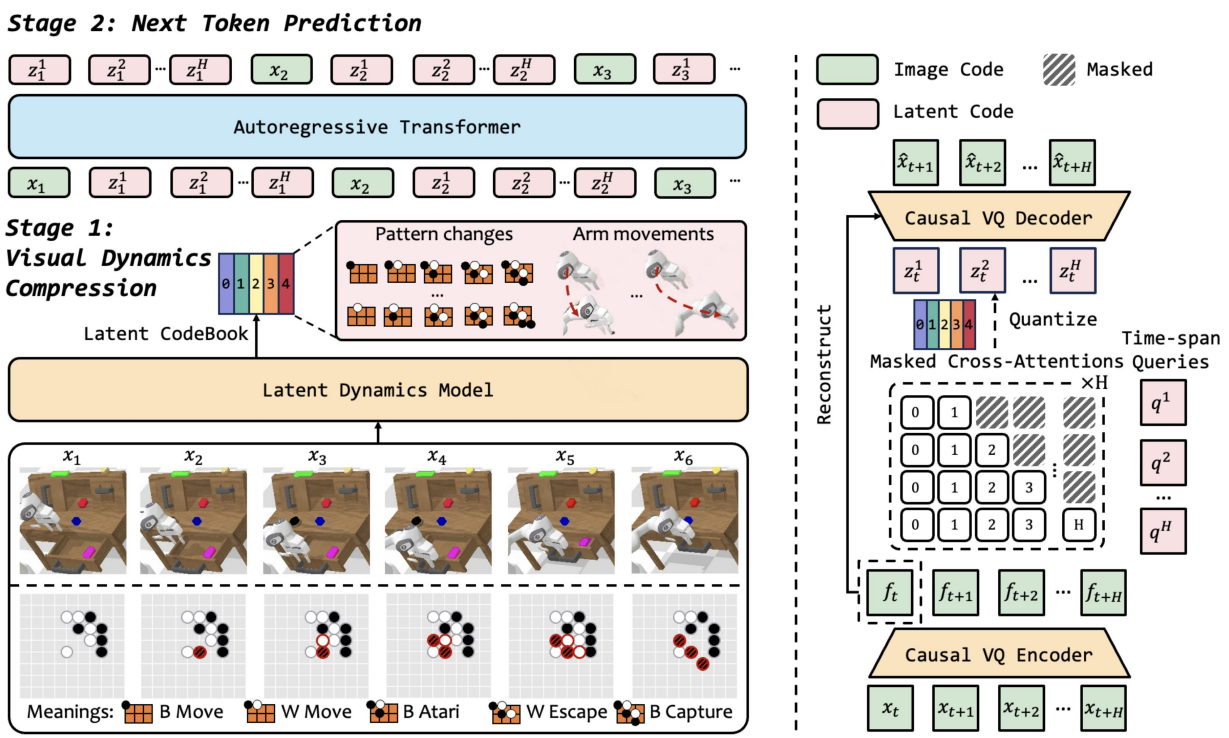
Video QA Results



Learn Solely from Videos

- Videos contain **richer** spatial and temporal information than text (LLMs), learning solely from them allows a model to infer how the world changes over time—capturing motion, causality, and physical structure—without relying on external supervision such as text or rewards.
- Hence, there are methods seek to explore whether complex knowledge (rules, reasoning, planning) can be learned **solely** from visual input (videos), without relying on textual data, labels, or reinforcement learning signals.
- **Challenge: Inefficiency in raw video representations.** Video frames contain a lot of redundant or irrelevant visual details; encoding all pixel changes can be inefficient and hinder the learning of high-level task knowledge.

Learn Solely from Videos: VideoWorlds



Latent Dynamics Model: VQ/FSQ + AR transformer. To address inefficiency and redundancy in raw video change representation, Latent Dynamics Model compresses multi-step visual changes into discrete latent codes (via VQ/FSQ) to act as a bottleneck, avoiding trivial copying. So the output is a discrete latent token representing the state change or action-like dynamics between frames.

ByteDance, *VideoWorld: Exploring Knowledge Learning from Unlabeled Videos* (CVPR 2025)

Results: Go and Robot Manipulation

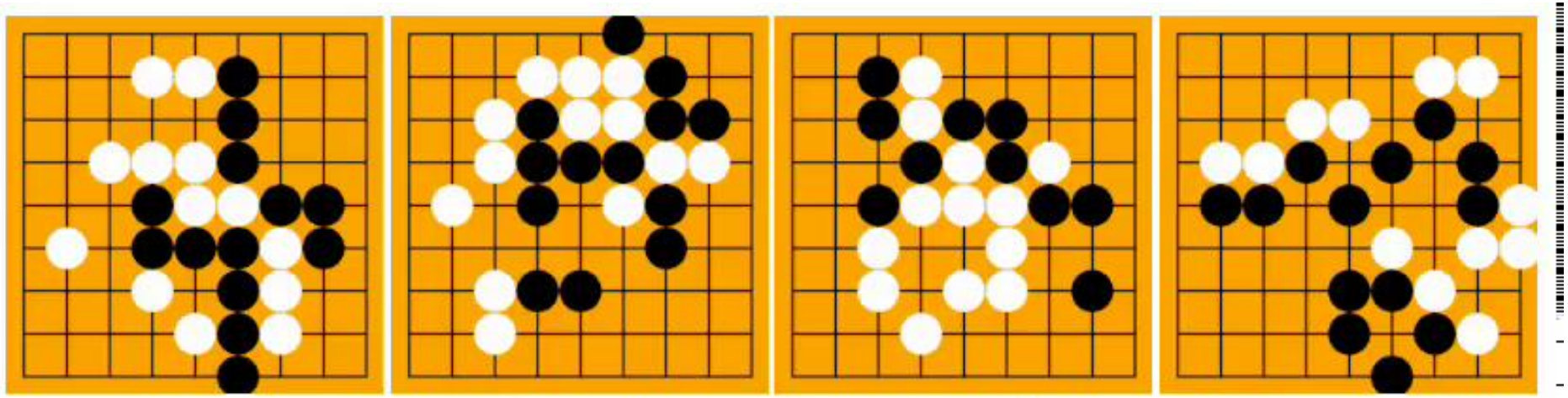


Figure 2: VideoWorld plays Go by generating next board state.

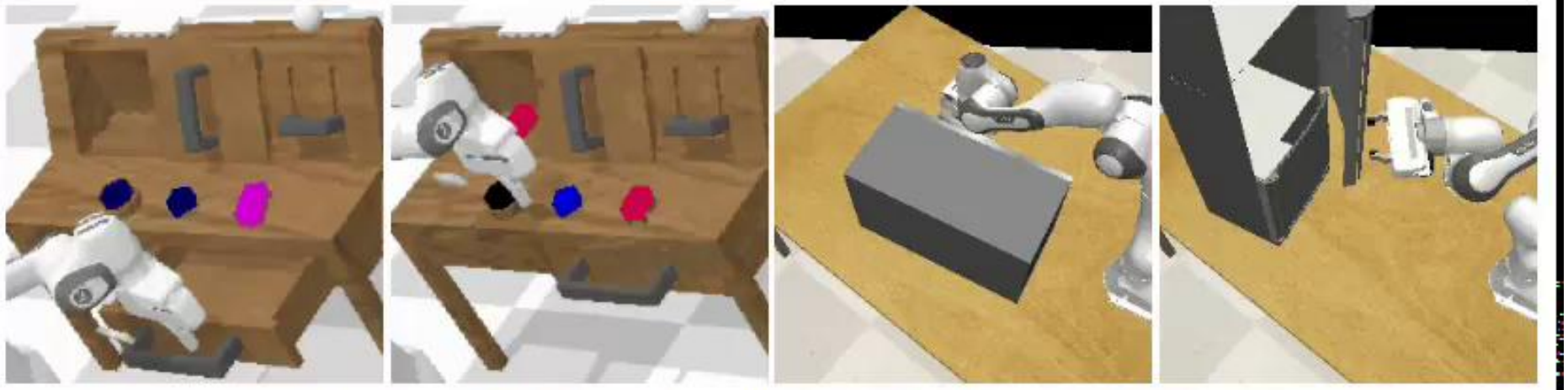
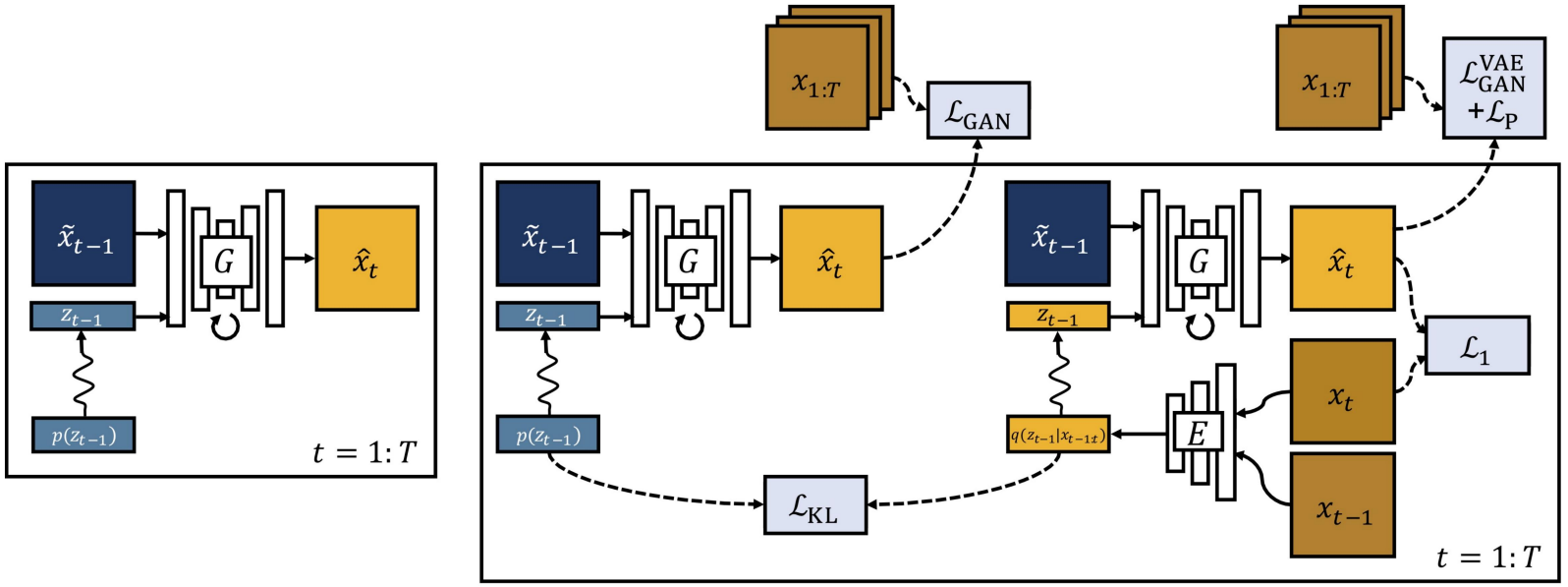


Figure 3: VideoWorld controls robotic arms across different environments.

Video Generation for Future State Prediction

- The idea of video prediction in world modeling emerged **prior** to the modern era of video generation.
- Predicting only second-level videos is limited in practical applications. But what is the proper paradigm to generate future video frames given a current visual perception in a **long** and **streaming** fashion? (AR and Semi-AR)
- How can we generate a **geometrically consistent** and **physically plausible** video that can be grounded into the real world?

Starting Point: Stochastic Adversarial Video Prediction (SAVP)

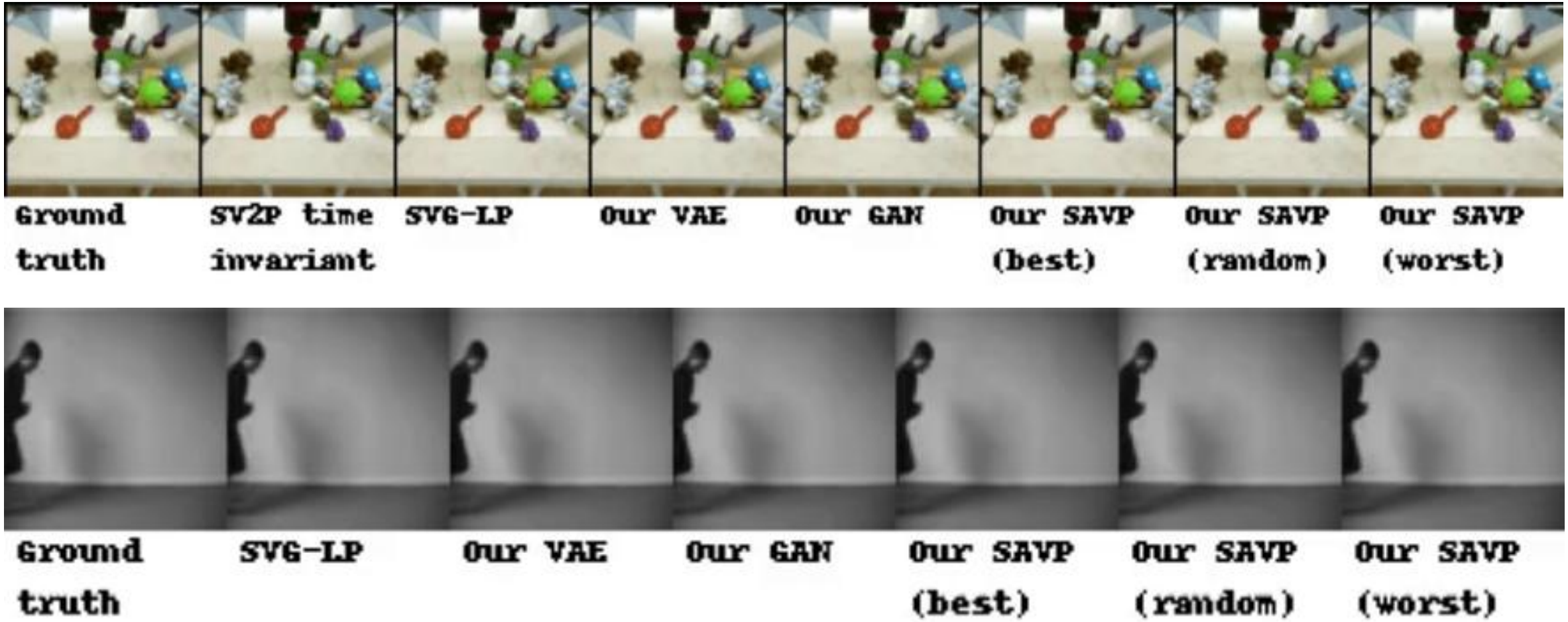


(a) Testing Usage (b) Our SAVP Model (Training Usage)

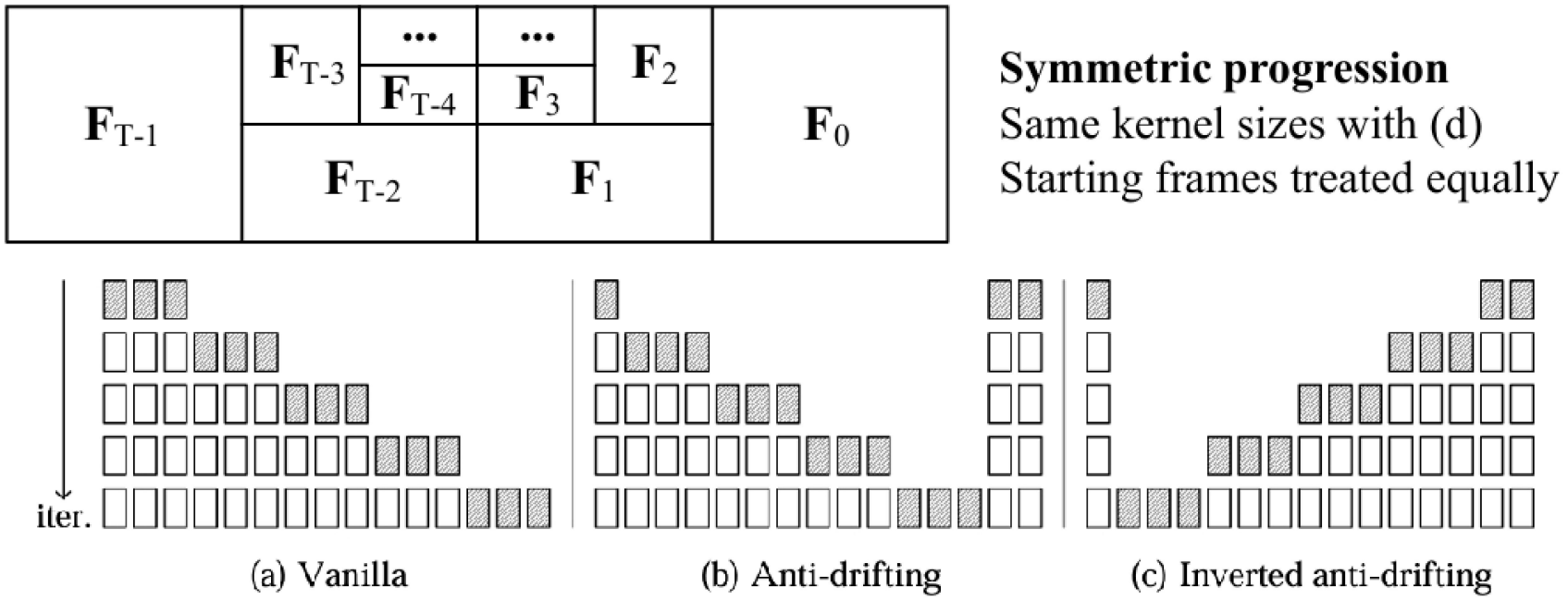
Predicting future frames from video input is inherently uncertain and multimodal: many different futures could plausibly follow the same past. SAVP addresses this uncertainty by combining VAE-based stochastic latent modeling with GAN-based adversarial training, enabling the generation of diverse yet realistic future video frames.

Pieter Abbeel, Chelsea Finn, Sergey Levine et al., *Stochastic Adversarial Video Prediction* (2018)

Results: Robot Manipulation and Human Motion



Diffusion: Bidirectional Attention

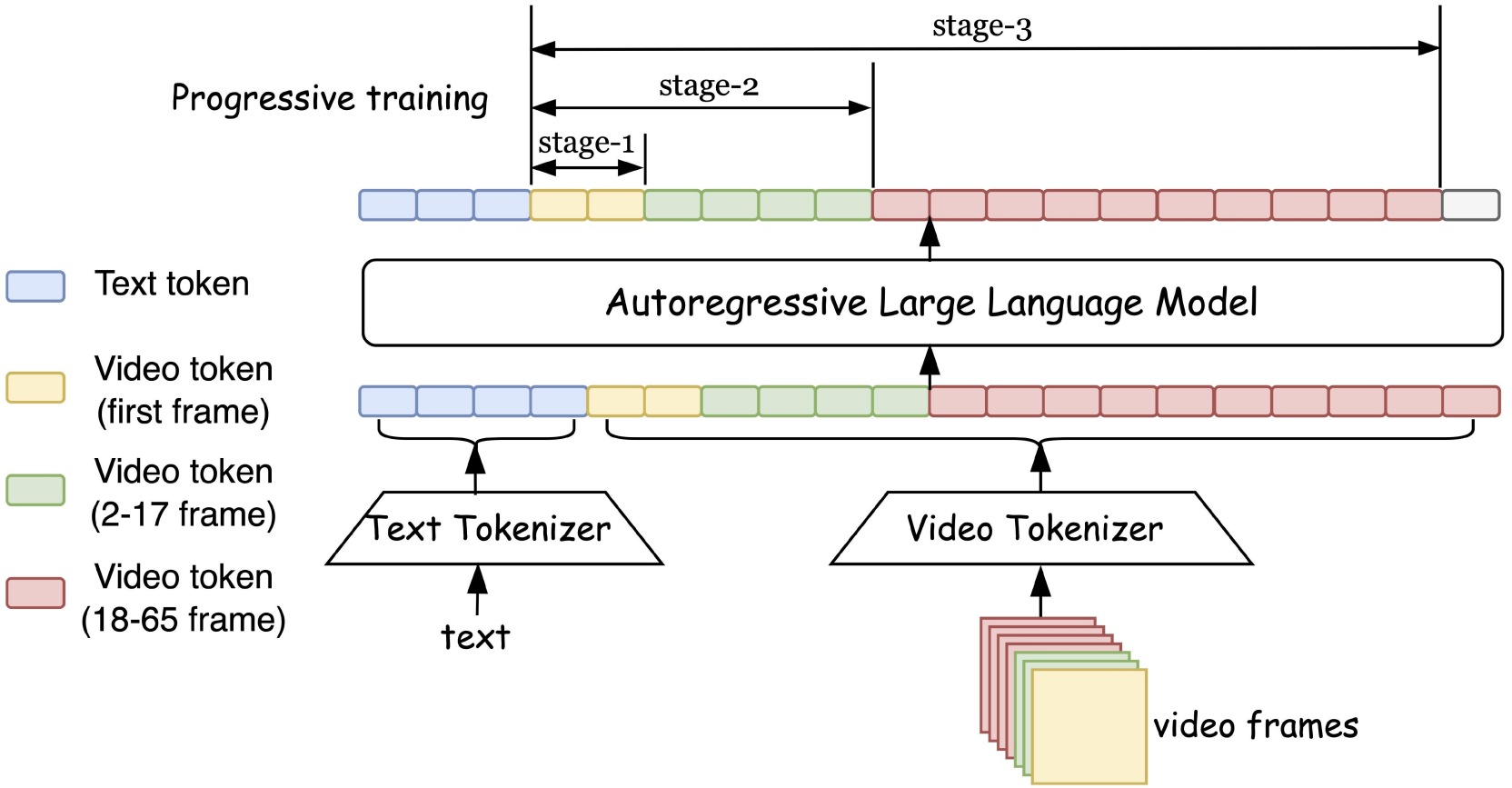


FramePack compresses input frames into a fixed-length representation by prioritizing them by importance so that the transformer's context size remains constant regardless of video length, and uses bi-directional sampling (i.e. generating frames in non-causal order with both past and future context) to mitigate error propagation ("drift").

Lvmin Zhang et al., *Frame Context Packing and Drift Prevention in Next-Frame-Prediction Video Diffusion Models* (NeurIPS 2025 Spotlight)

Results: FramePack

AR: Causal Attention



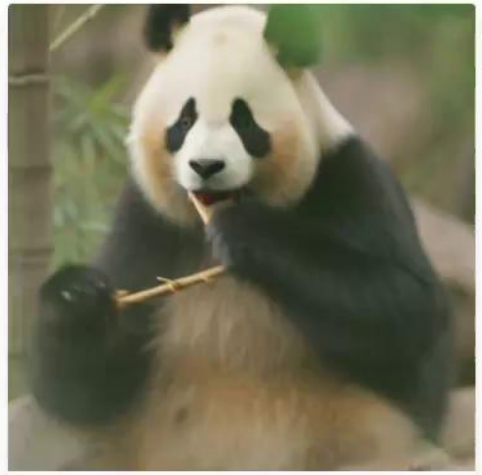
Loong uses causal attention in an autoregressive fashion to generate video frames step by step, treating each future frame as conditioned only on past frames to ensure temporal consistency.

Yuqing Wang et al., *Loong: Generating Minute-level Long Videos with Autoregressive Language Models* (2024)

Results: Loong's T2V

Clown fish swimming through the coral reef

Aerial view of Santorini during the blue hour, showcasing the stunning architecture of white Cycladic buildings with blue domes. The caldera views are breathtaking, and the lighting creates a beautiful, serene atmosphere



A panda eating bamboo on a rock

Hulk wearing virtual reality goggles

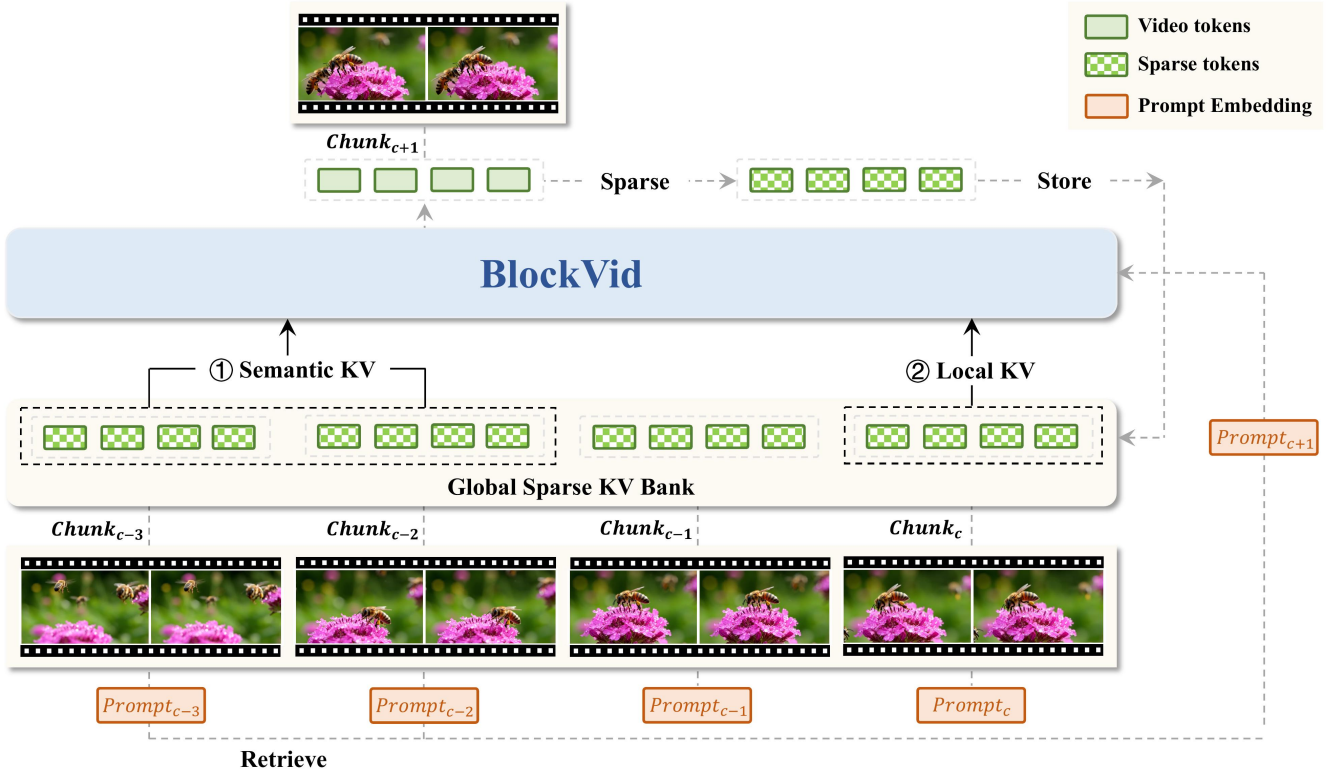
A bigfoot walking in the snowstorm

Two pandas sitting at a table playing cards

Modern Video Generation Paradigms

- Although diffusion-based and diffusion-forcing models can produce high-quality video clips, their reliance on bidirectional attention makes inference inefficient. The bidirectional attention prevents KV cache technique, leading to redundant computation and prohibitive latency for long videos.
- For AR models with causal attention, they can leverage cached KV states for faster inference, but they often exhibit degraded quality when generating long videos.
- What comes next to address the shortcomings of these two approaches? (Semi-AR)

Semi-AR: BlockVid



BlockVid introduces a semi-AR "block diffusion" framework that generates videos chunk-by-chunk — adapting diffusion within each chunk for high-quality denoising and AR causal conditioning across chunks for temporal continuity.

Zeyu Zhang et al., *BlockVid: A Scalable and Efficient Block Diffusion Framework for Minute-Long Video Generation* (2025)

Results: Chunk-by-Chunk Prediction and Minute-Long Generation

BlockVid: Block Diffusion for High-Fidelity and Coherent Minute-Long Video Generation

Zeyu Zhang[1] Shuning Chang[1] Yuanyu He[1,2] Yizeng Han[1]
Jiasheng Tang[1,3*] Fan Wang[1†] Bohan Zhuang[1,2*†]

[1]DAMO Academy, Alibaba Group [2]ZIP Lab, Zhejiang University [3]Hupan Lab

* Project leads. † Corresponding authors.

Long Video Generation Form

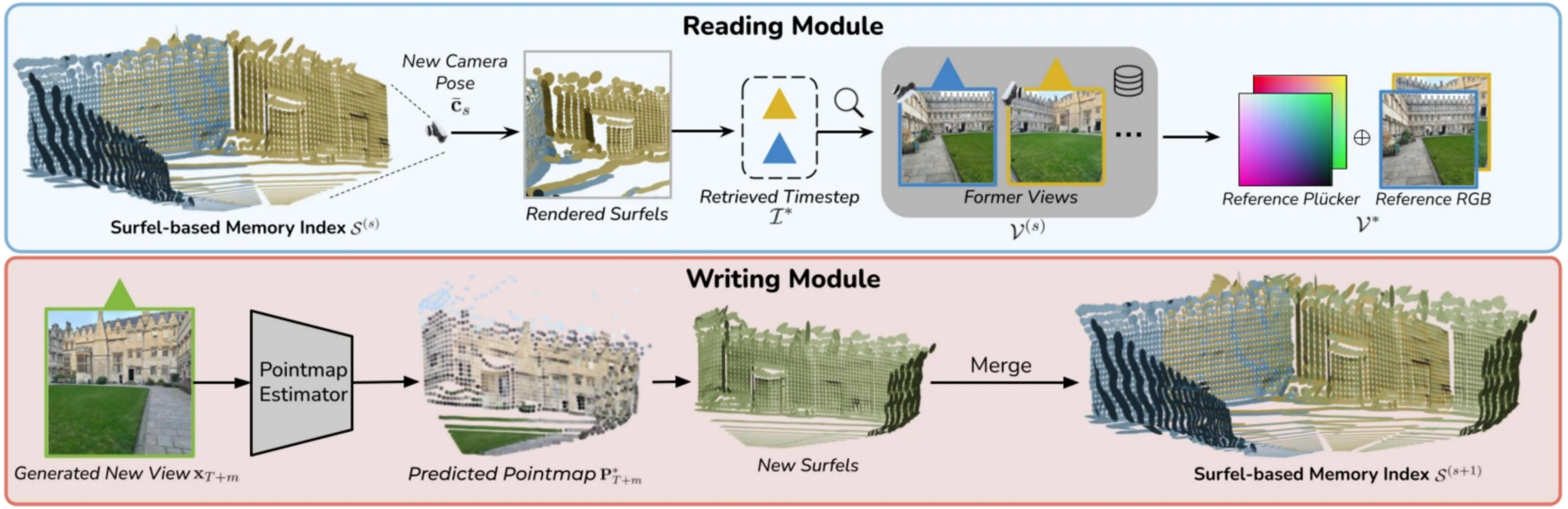
Single-shot video with static camera.

Multi-shot video concatenation.

Evolve

Single-shot video with camera motion.

Geometrically Consistent: VMem



VMem enforces **geometrically consistency** by indexing past views to surface elements (surfels, typically computed from multi-view images or depth maps) they observed; when generating a new view, it retrieves only those past views that correspond to the same 3D surfaces and uses them to condition generation, thus anchoring novel-to-past view coherence.

Runjia Li, Philip Torr, Andrea Vedaldi, et al., *VMem: Consistent Interactive Video Scene Generation with Surfel-Indexed View Memory* (2025)

Results: VMem

Evaluation of Geometrically Consistent: WorldScore

Score: 94.41

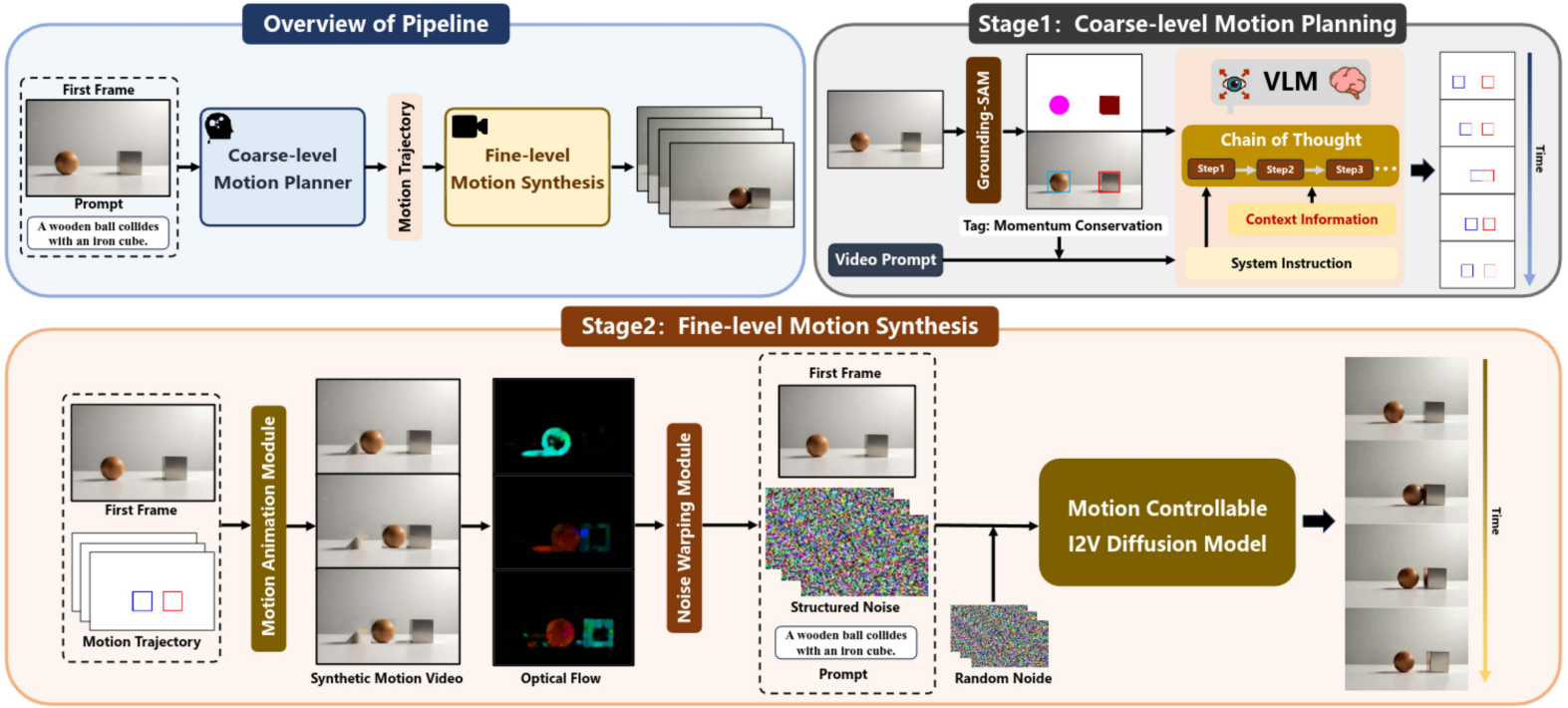
Score: 0

Score: 92.88

Score: 0

Haoyi Duan et al., *WorldScore: A Unified Evaluation Benchmark for World Generation* (2025)

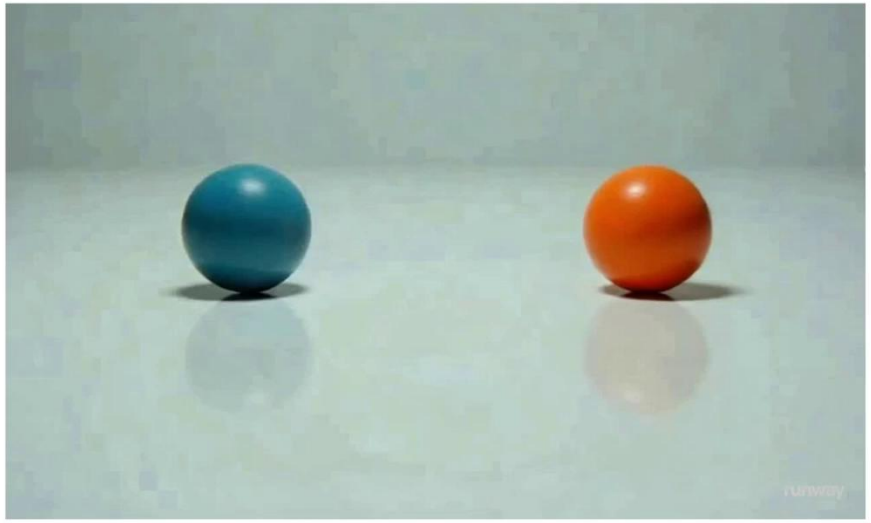
Physically Plausible: VLIPP



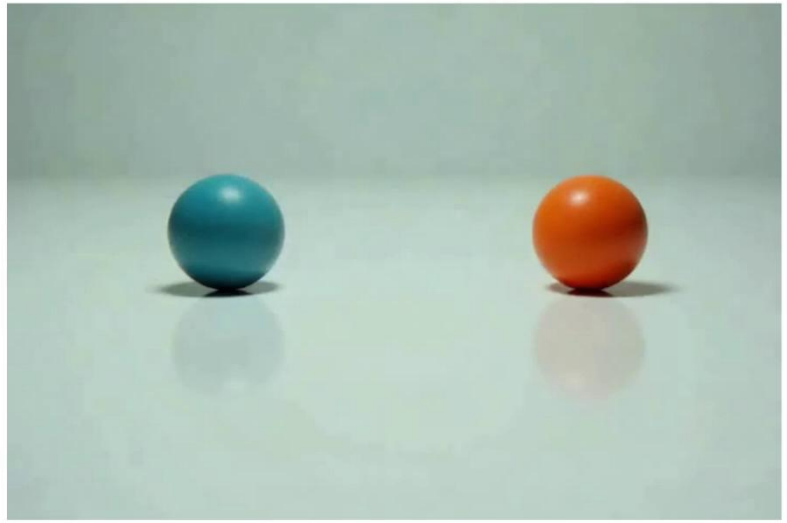
VLIPP improves physical plausibility in video generation by using a two-stage framework: first a VLM plans coarse, physically consistent trajectories (e.g. object bounding box changes) via chain-of-thought and physics reasoning, then a video diffusion model is conditioned on those trajectories (via optical flow and structured noise) to generate fine motion that aligns with physical laws.

Xindi Yang et al., *VLIPP: Towards Physically Plausible Video Generation with Vision and Language Informed Physical Prior* (2025)

Results: VLIPP



Gen-3

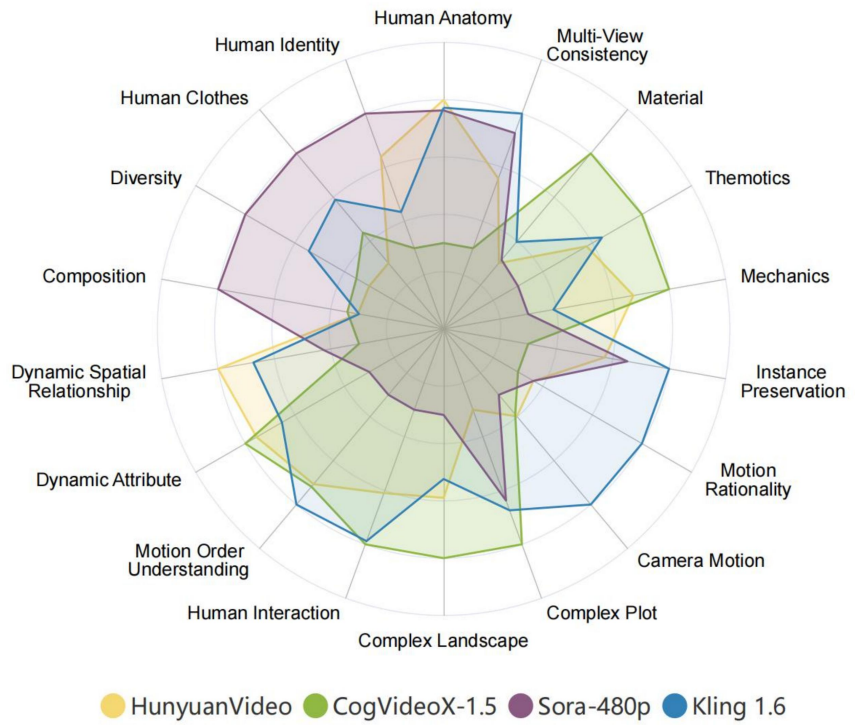


Ours

Kling 1.6

Ours

Evaluation of Physically Plausible: VBench-2.0



VBench-2.0 evaluates physical plausibility (under its "Physics" dimension) by applying specialized tests (via video-language models, anomaly detectors, etc.) that assess whether generated motion adheres to physical rules (e.g. gravity, collision consistency, motion rationality) and comparing the results to human judgments.

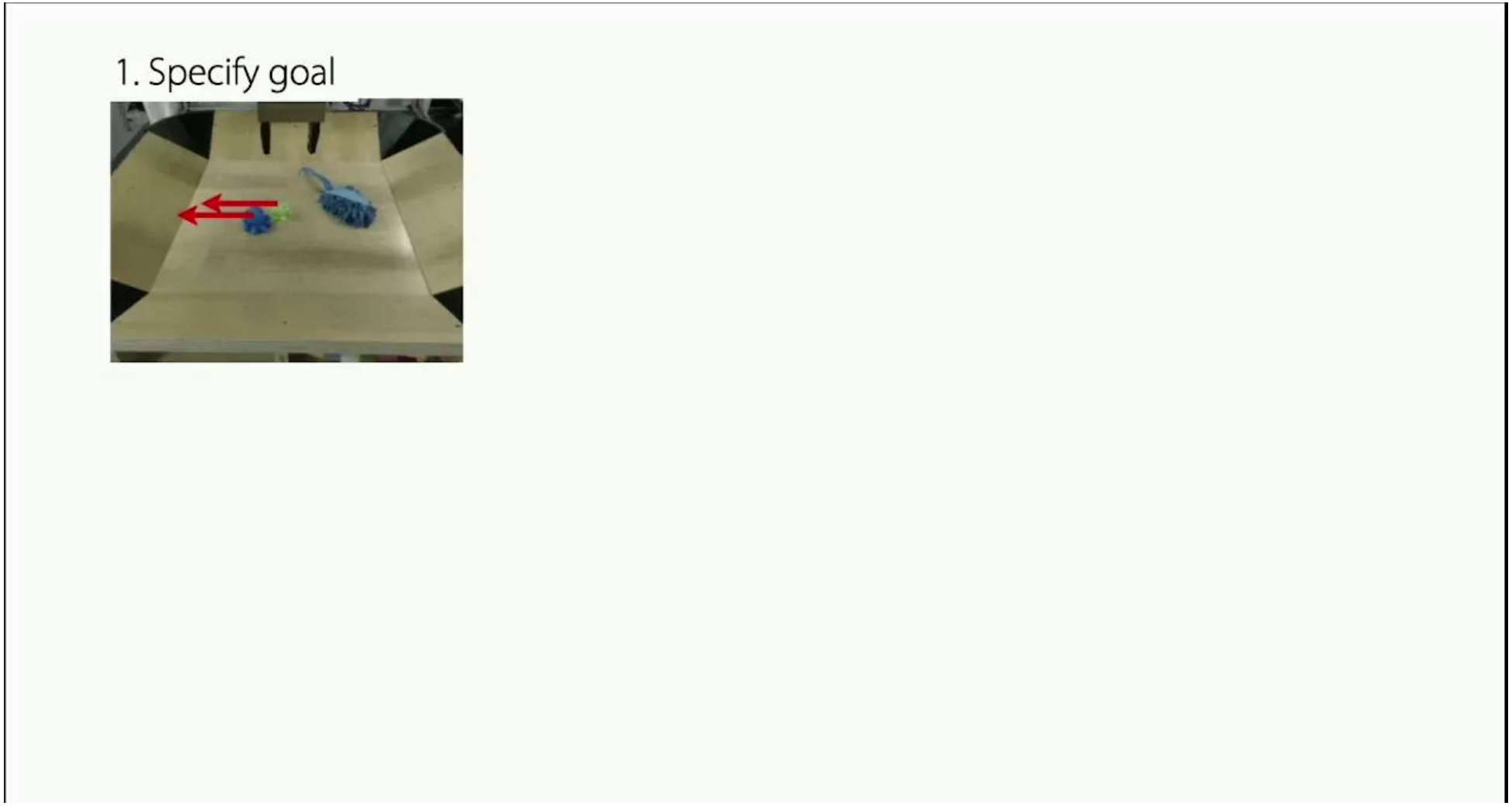
Dian Zheng et al., *VBench-2.0 : Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness* (2025)

Physical Interaction through Video Prediction

"One of the things that is really promising about leveraging video prediction models for robot interaction is that through planning, we actually accomplished a wide range of tasks, including those involving objects that the robot has never seen before."

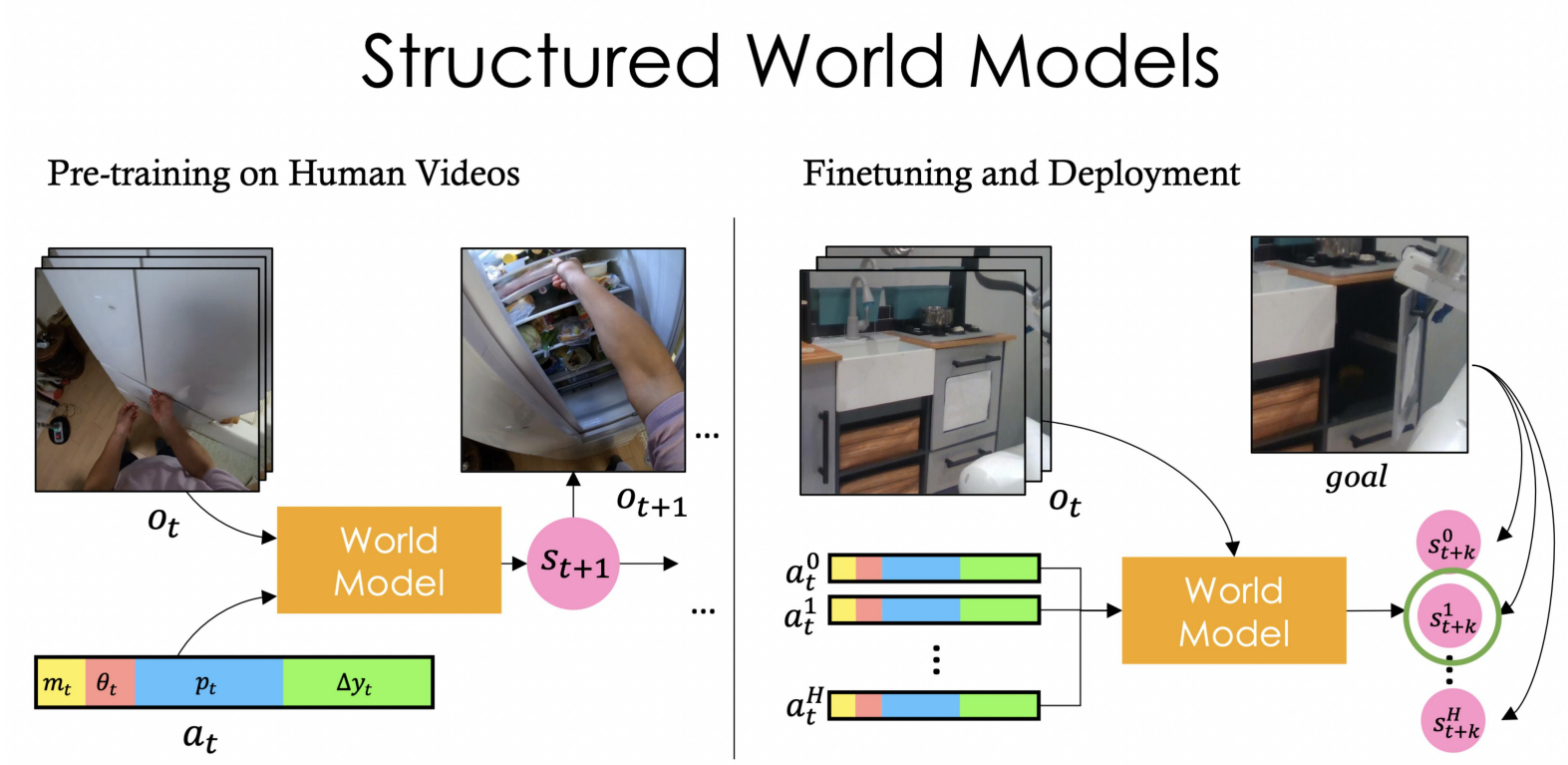
— Chelsea Finn

Early Attempts: Visual Foresight



Sergey Levine, Chelsea Finn et al., *Improvisation through Physical Understanding: Using Novel Objects as Tools with Visual Foresight* (RSS 2019)

Structured World Models for Intentionality (SWIM)



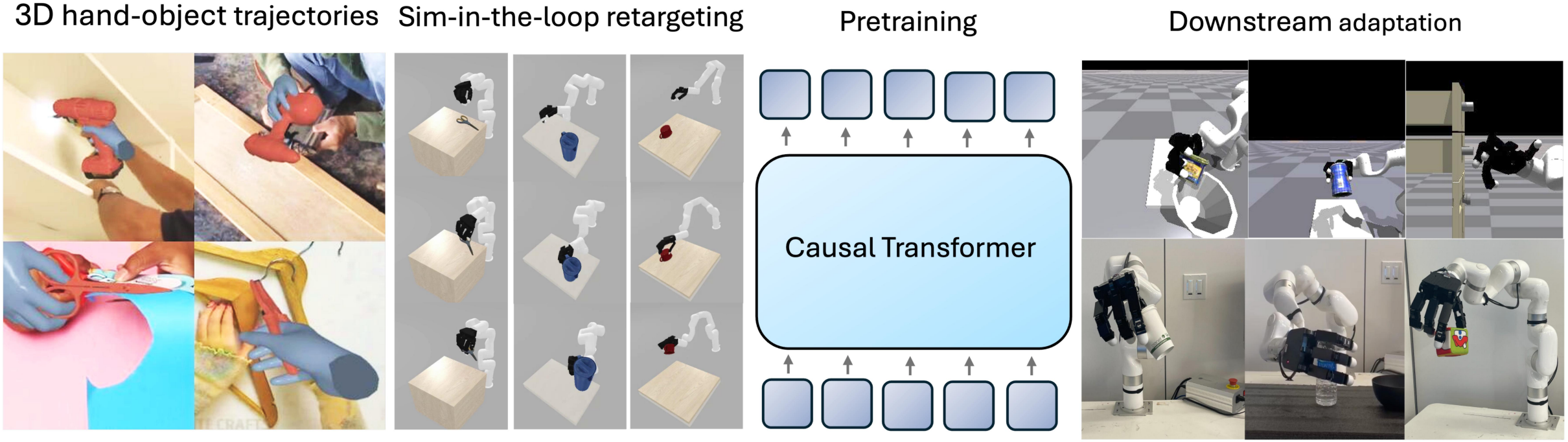
Structured World Model involves 3 steps: (1) pretraining a world model on human videos, (2)finetuning the world model on unsupervised robot data, and (3) using the finetuned model to plan to achieve goals.

Russell Mendonca et al., *Structured World Models from Human Videos* (RSS 2023)

Pretraining from Human Videos: SWIM

Results: SWIM

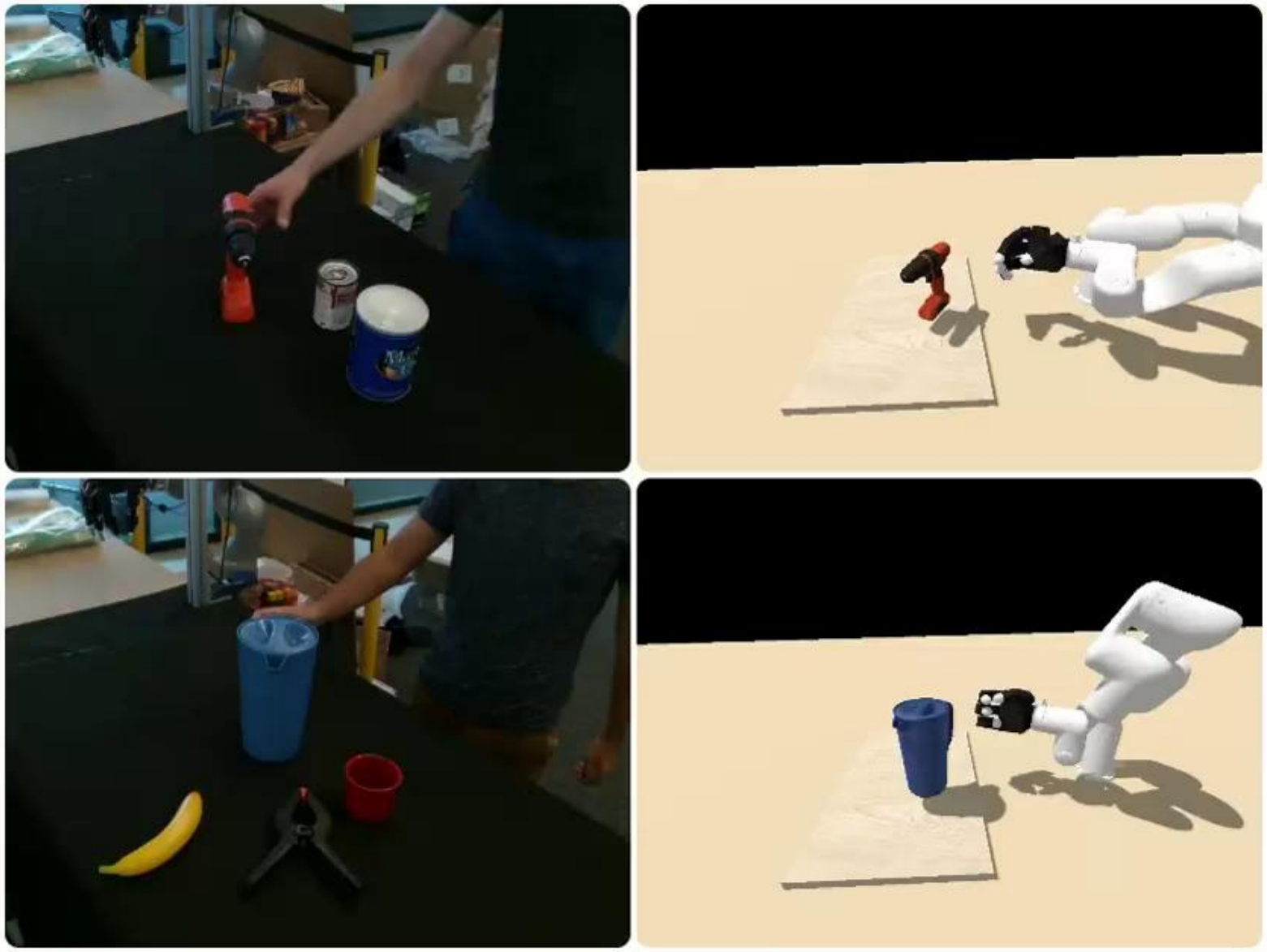
HOI Pretraining from Videos (HOP)



HOP extracts 3D hand-object trajectories from in-the-wild videos, retarget them to a robot embodiment via simulation, and train a task-agnostic manipulation prior.

Pieter Abbeel, Jitendra Malik et al., *Hand-Object Interaction Pretraining from Videos* (ICRA 2025)

Results: HOP



Takeaways

- Work on unimodal LLMs that perform next-token prediction will not achieve advanced machine intelligence. If you are interested in human-level intelligence, do not rely solely on LLMs; instead, enhance spatial awareness in visual foundation models and learn from videos.
- Downstream task is the gold standard for evaluation of video world models.

End

Thank you.