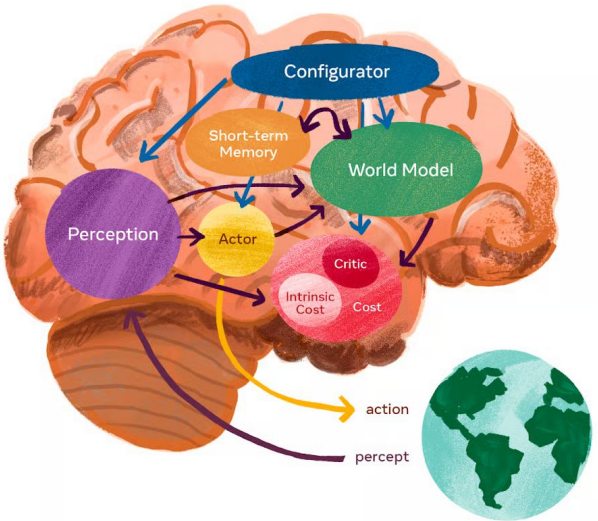
UNIVERSITY OF OXFORD

IMPERIAL

World Models

A Path Towards Machine Intelligence?

Firas Darwish, Harleen Gulati, Horia Nicolcea



Introduction: Motivation

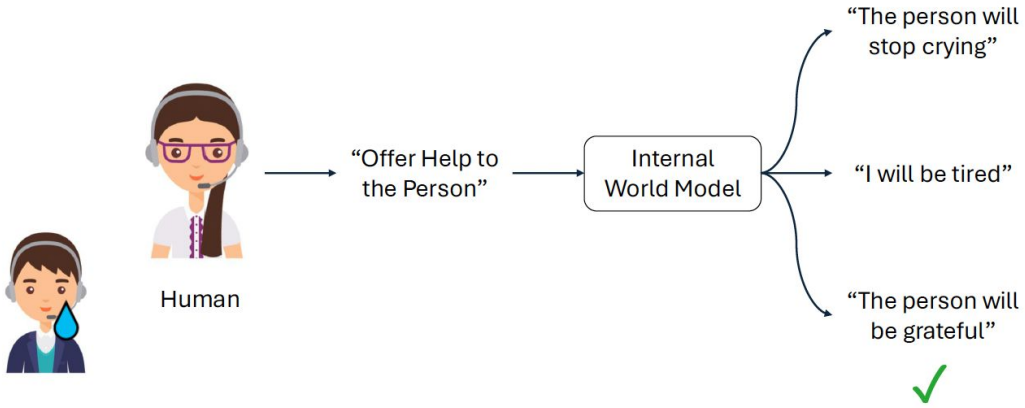


Figure 1: Familiar example of reasoning by simulation – an individual (possibly self-serving) decides to offer help to a crying person by mentally simulating multiple possible outcomes, with the best expected reward in mind.

Xing et al. [2025]

Introduction: What is a world model?

- Hypothetical thinking or “thought experiments” (Ball et al. [2025]).
- What if a machine too can perform thought experiments by simulating actions and plans in complex scenarios, extracting the best one among them (Xing et al. [2025])?
- World model is a generative model that simulates possibilities in diverse scenarios (Xing et al. [2025]).
- Takes the previous world state and action and predicts or simulates the next world state through a transformative function (e.g. a conditional probability distribution) (Xing et al. [2025]).

$$s' \sim p(s'|s, a)$$

Introduction: Why do we need them?

- Fundamental for physical reasoning (Zhou et al [2025]).
- Example: autonomous vehicles driving safely by forecasting future street scenarios (Zhou et al [2025]).
- Imitation learning and reinforcement learning (RL) enabled agents to learn complex behaviours across diverse tasks (Zhou et al [2025]).
- With such methods, generalisation is a major challenge (Zhou et al [2025]).
- WM trained on offline, pre-collected trajectories, with test-time optimisation and task-agnostic reasoning can generalise to diverse task families (Zhou et al [2025]).

What makes a ‘good’ WM?

- (Xing et al. [2025]) argues the following 5 aspects are desirable for the building and training of a general WM

(a) Identifying and preparing training data with the desired world information.

(b) Adopting a general representation space for the latent world state (possibly richer meaning than observation data in plain sight).

(c) Architecture that allows effective reasoning over representations.

(d) Objective properly guides the model training.

(e) Determining how to use the world model in decision making systems.

Video Joint Embedding Predictive Architectures (V-JEPA) Assran et al. [2025]

- Learn a predictive model for world dynamics in this **learned representation space**.
  - *Why?* Ignoring pixel-level reconstruction objective means we learn **semantic features**

- Encoder and predictor are parameterized as **vision transformers**

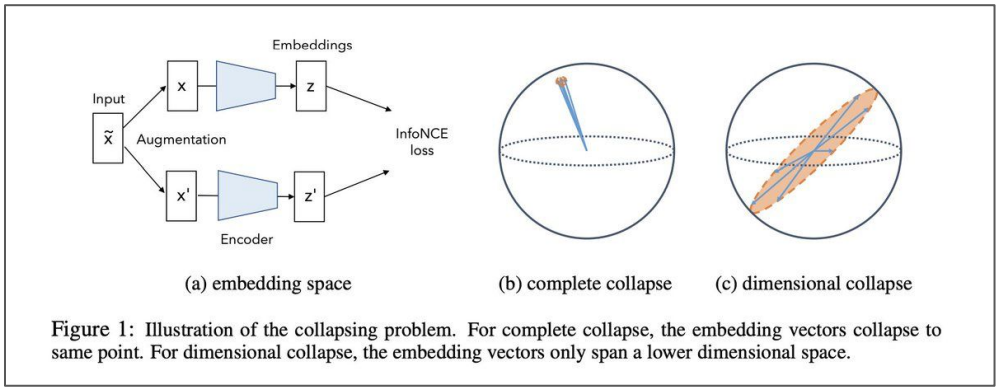
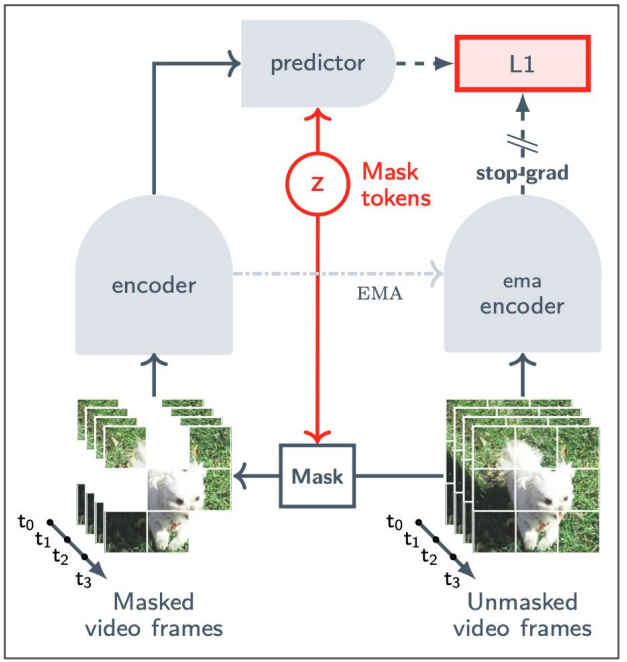


Figure 1: Illustration of the collapsing problem. For complete collapse, the embedding vectors collapse to same point. For dimensional collapse, the embedding vectors only span a lower dimensional space.

Jing et al. [2021]

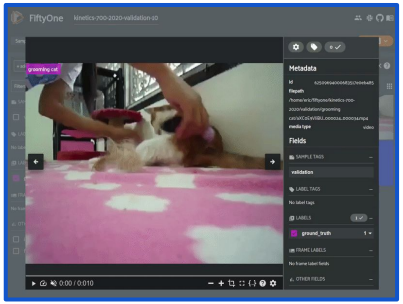


Assran et al. [2025]

UNIVERSITY OF OXFORD IMPERIAL

V-JEPA Data

Exo-centric videos



Pu et al. [2024]

Goyal et al. [2017]

Ego-centric videos

Miech et al. [2019]

YouTube videos

ImageNet (duplicate image temporally)

Russakovsky et al. [2015]

V-JEPA Evaluation

- Freeze encoder
- Train attentive probe
- Action recognition, motion classification, action localization

(a) headbanging

(b) stretching leg

(c) shaking hands

(d) tickling

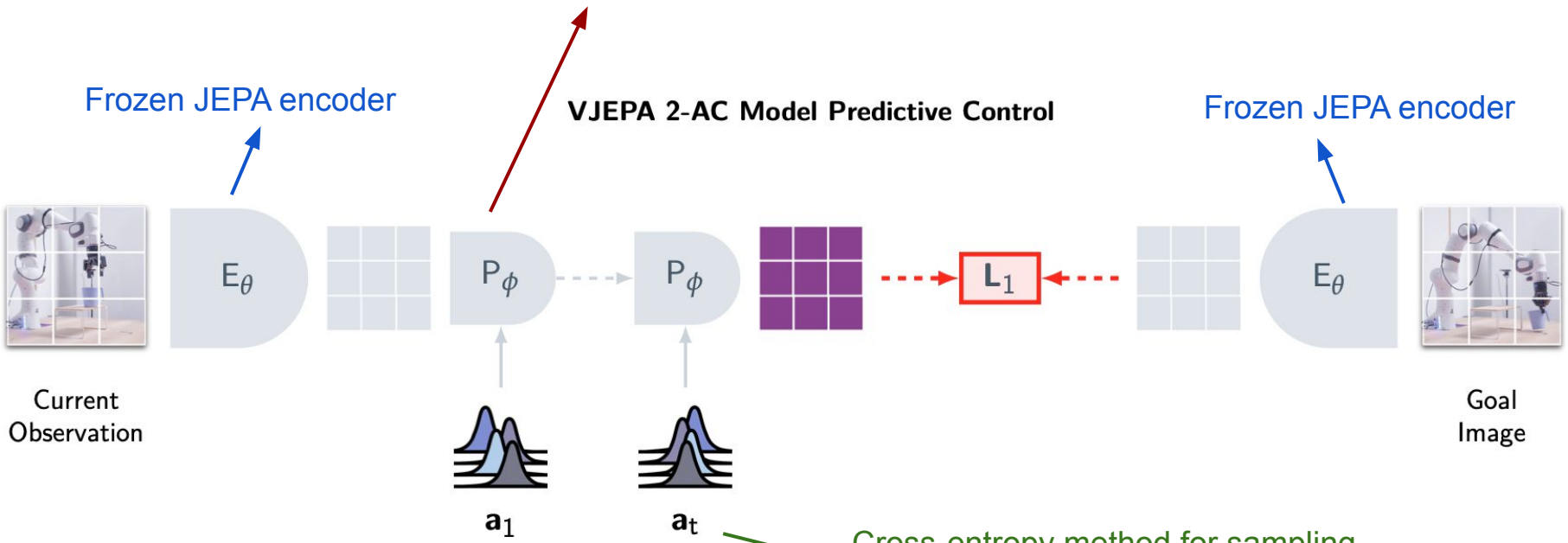
(e) robot dancing

(f) salsa dancing

Assran et al. [2025]

V-JEPA Evaluation

New model (our world model!) trained on top of frozen JEPA encoder to predict representation of future frames, conditioned on action and end-effector state



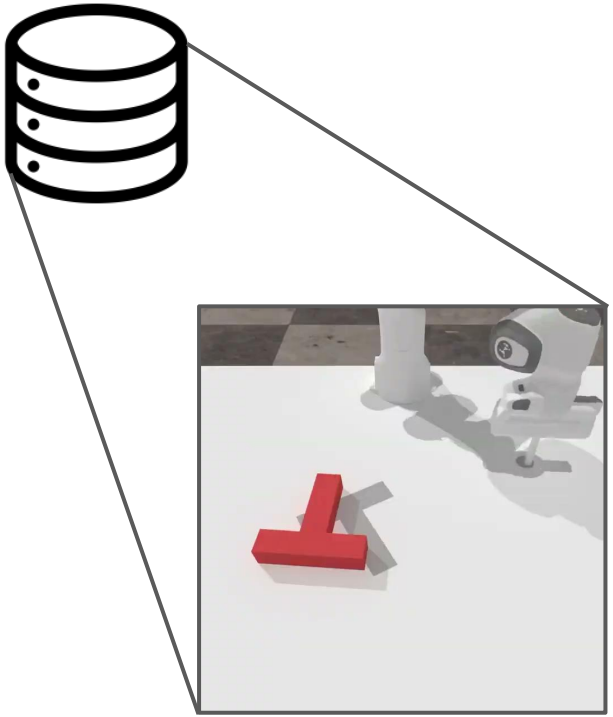
Assran et al. [2025]

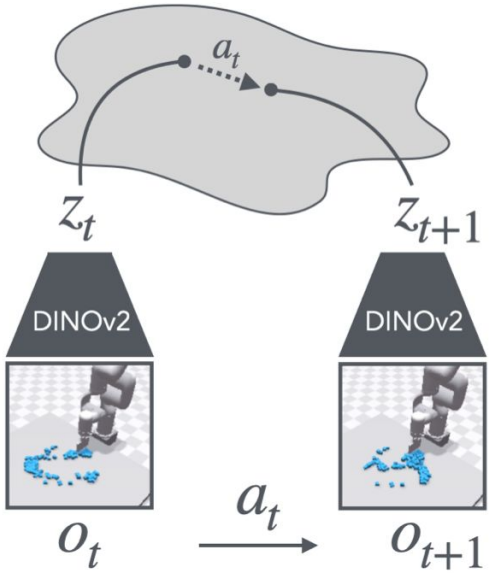
DINO-WM: World Models on Pre-trained Visual Features enable Zero-shot Planning

Core Ideas:

1. Build task-agnostic dynamic model (world model) from an **offline dataset of trajectories** *(no need for reward modelling or expert policy!)*

2. Model the world dynamics on **compact embeddings of the world**, rather than the raw observations themselves

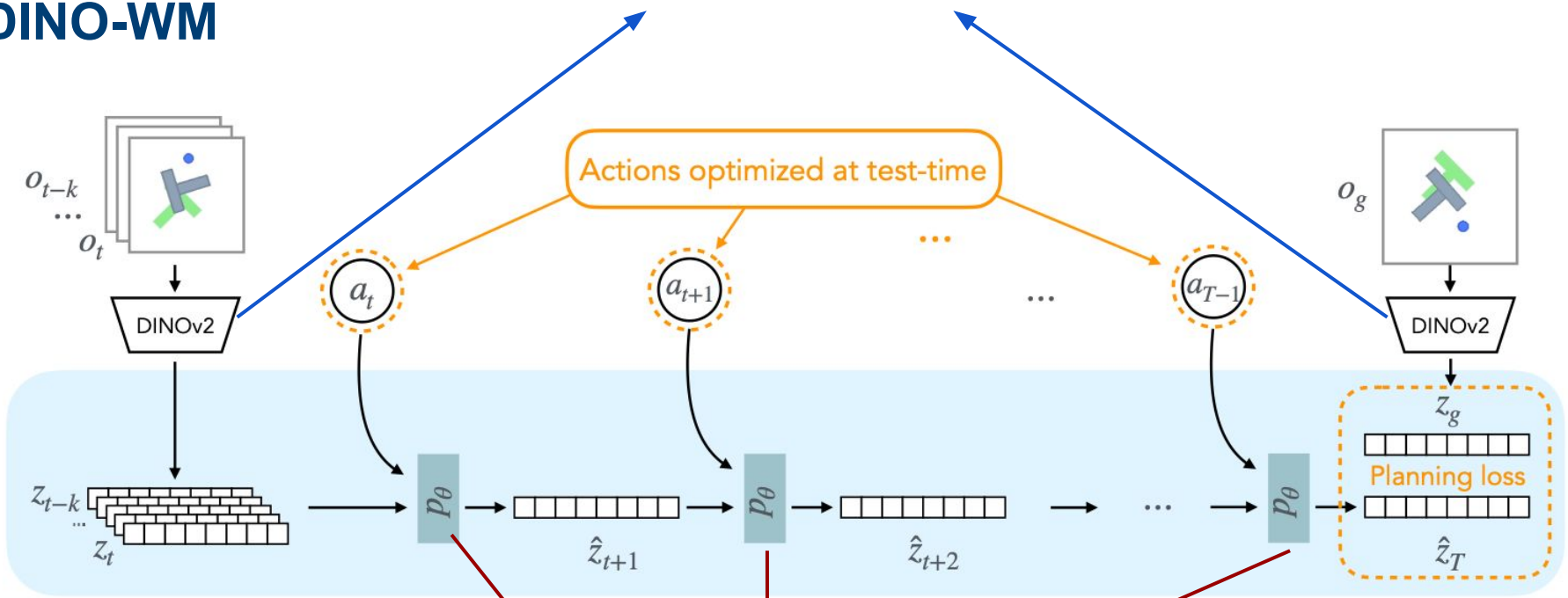




Zhou et al. [2025]

DINO-WM

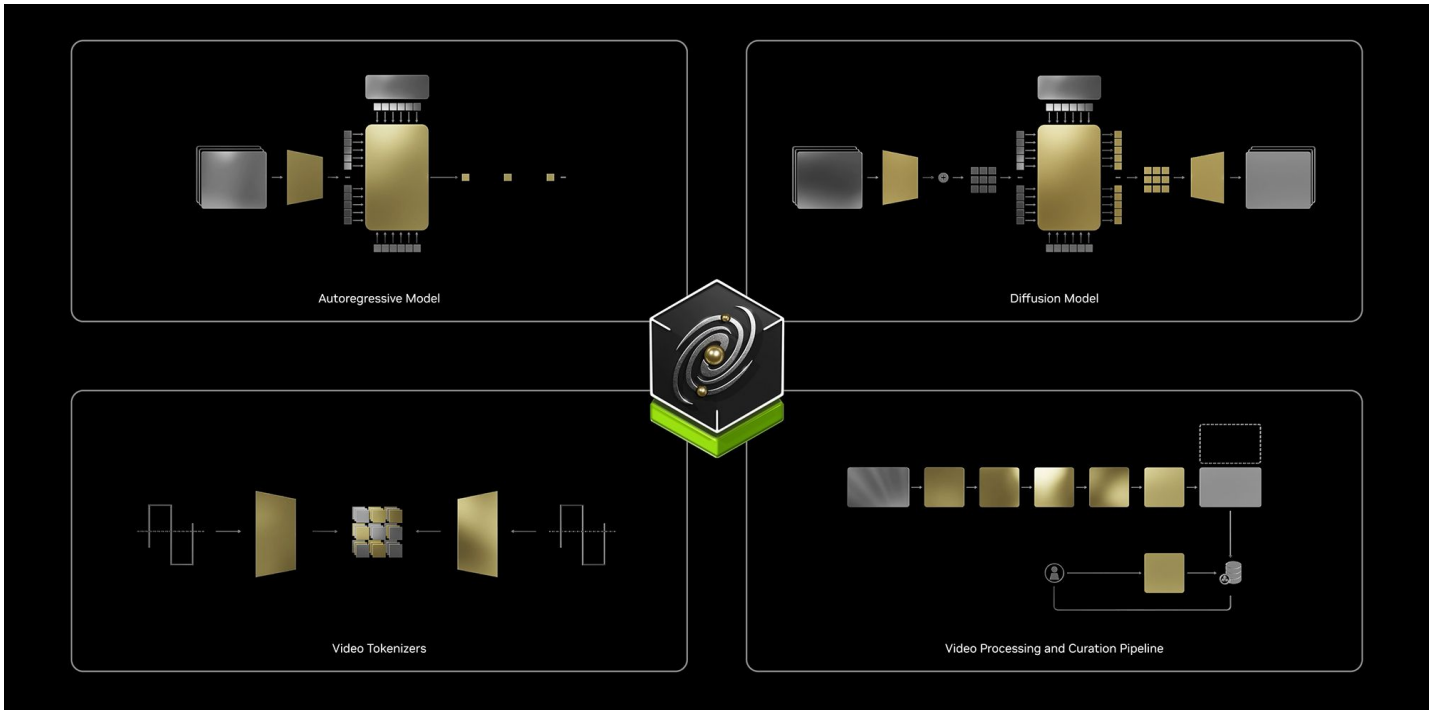
Pre-trained Observation Model (task and environment independent, captures rich spatial information)



Dynamic model takes in history of past latent states and actions and predicts the latent state at next time step.

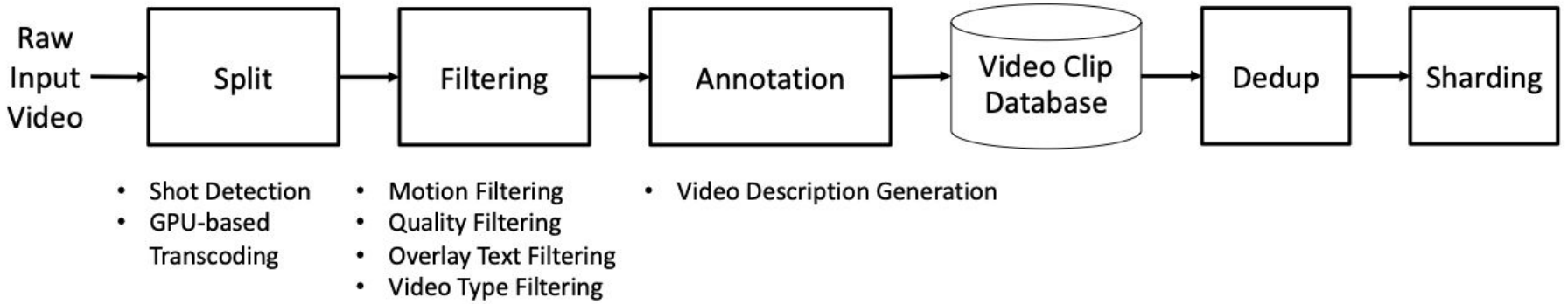
Zhou et al. [2025]

COSMOS World Foundation Model Platform for Physical AI



NVIDIA et al. [2025]

COSMOS: Data Curation

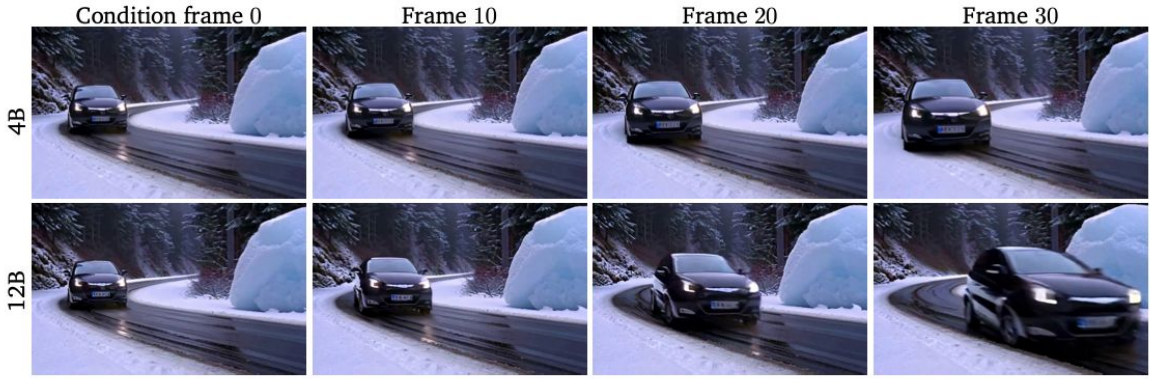


NVIDIA et al. [2025]

UNIVERSITY OF OXFORD IMPERIAL

COSMOS: Models

- **Autoregressive:**
  - With discrete tokenizer (encodes visual data into discrete latent codes)
- **Diffusion**:
  - With continuous tokenizer (encodes visual data into continuous latent embeddings)



Prompt: None.

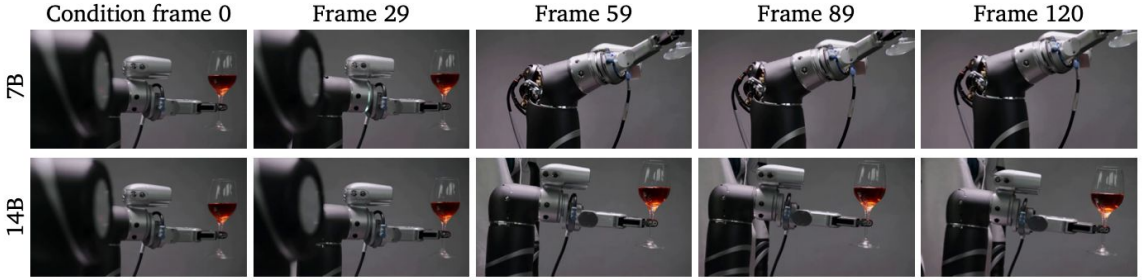
Prompt: The video of a car moving forward, passing under a large overpass. The road is clear, and there are a few other cars visible in the distance. The weather appears to be sunny, and the time of day is daytime. The scene is set on a busy highway with concrete structures and greenery on the sides.

NVIDIA et al. [2025]

UNIVERSITY OF OXFORD IMPERIAL

COSMOS: Diffusion Model

- **Continuous Tokenizer**
  - Encode visual data into continuous latent embeddings
- **Video2World**
  - predict future video based on the current observation (input video) and perturbation (text prompt)



Prompt: The video depicts a robotic arm holding a wine glass filled with red wine. The robotic arm, equipped with multiple joints and mechanical components, appears to be designed for precision tasks. The glass is held delicately, showcasing the robot's capability to handle fragile objects. The background is minimalistic, emphasizing the interaction between the robot and the wine glass.

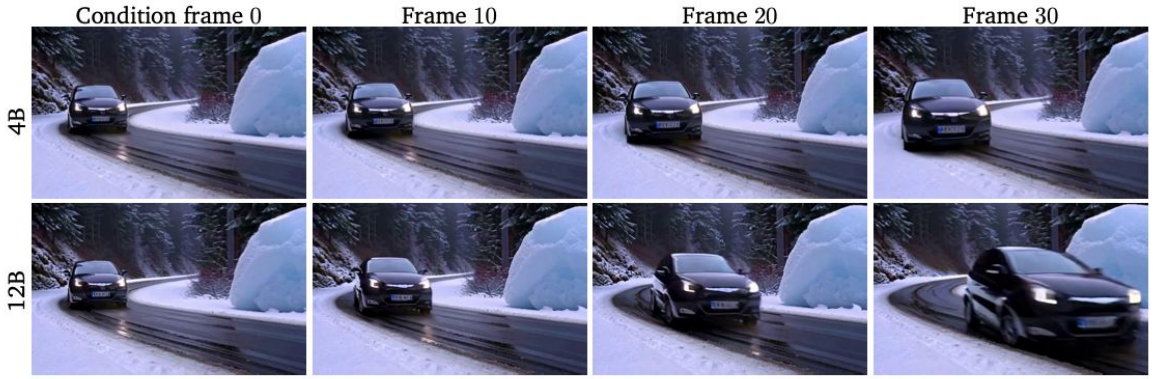
Prompt: The video depicts the interior of a large industrial facility, likely a factory or warehouse. The space is expansive with high ceilings and metal framework. Overhead cranes and various machinery are visible, indicating a setting for heavy manufacturing or assembly. The floor is mostly empty, with some scattered debris and marked lines. Safety signs and barriers are present, emphasizing the industrial environment. The lighting is natural, streaming through the high windows, illuminating the workspace.

NVIDIA et al. [2025]

UNIVERSITY OF OXFORD IMPERIAL

COSMOS: Autoregressive Model

- **Discrete Tokenizer**:
  - Encode visual data into discrete latent codes
- **Video2World**:
  - World generation as a **next-token prediction** based on input video and text prompt



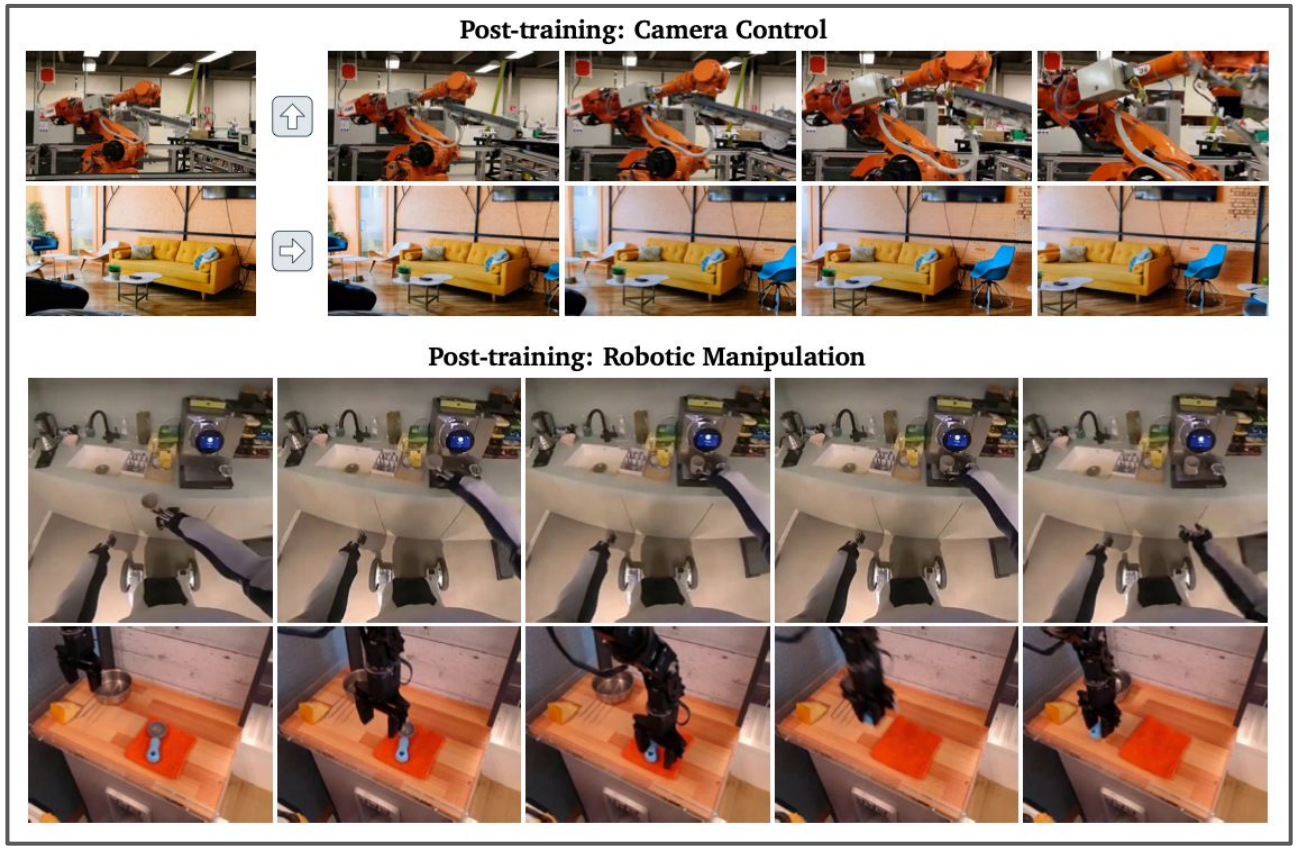
Prompt: None.

Prompt: The video of a car moving forward, passing under a large overpass. The road is clear, and there are a few other cars visible in the distance. The weather appears to be sunny, and the time of day is daytime. The scene is set on a busy highway with concrete structures and greenery on the sides.

NVIDIA et al. [2025]

UNIVERSITY OF OXFORD IMPERIAL

COSMOS: Post Training

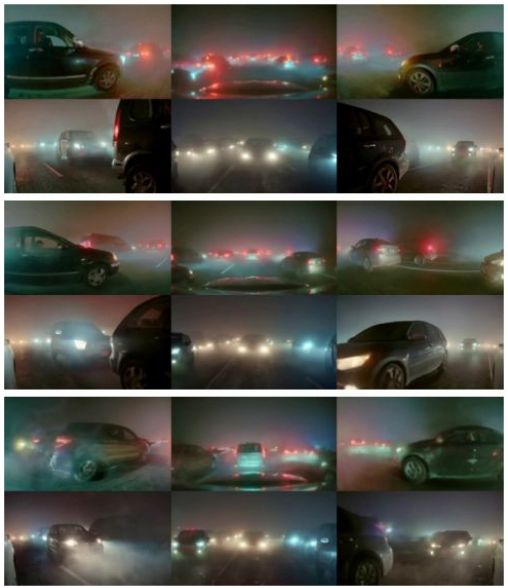


NVIDIA et al. [2025]

UNIVERSITY OF OXFORD IMPERIAL

COSMOS: Post Training

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.

NVIDIA et al. [2025]

School of common thoughts

- (Xing et al. [2025]) argues the vocal school of thought (MacKay et al. [2003]) making the following claims in building a WM warrant a closer examination:

1. Sensory inputs are superior to text (larger data volume from the physical world).
2. World state should be represented by continuous embeddings rather than discrete tokens (for gradient-based optimisation).
3. Autoregressive, generative models (e.g. LLMs) ‘doomed’.
4. Probabilistic, data-reconstruction objectives do not work.
5. World model should be used in MPC instead of RL, latter requires too many trials.

Closer examination…

Sensory inputs are superior to text (larger data volume from the physical world).

For a WM to be successful (i.e. general-purpose) it must learn from all modalities of experience to generalise across diverse tasks. Ignoring low-level perception or high-level abstraction risks omitting crucial information needed for intelligent behaviours (Xing et al. [2025]).

World state should be represented by continuous embeddings rather than discrete tokens (for gradient-based optimisation).

Continuous embeddings allow for smoother gradient flow, but this argument overlooks the noise and high variability with continuous sensory inputs. Humans counter this variability by categorising the raw perception into discrete concepts. Therefore, (Xing et al. [2025]) proposed mixed representations to leverage discrete tokens for robust, interpretable and symbolic forms of reasoning whilst continuous embeddings capture fine-grained sensory nuance (Xing et al. [2025]).

Closer examination…

Autoregressive, generative models (e.g. LLMs) 'doomed'.

Not necessarily, for many real-world systems are fundamentally chaotic with small deviations growing exponentially with time. However, autoregressive models can learn abstract properties of systems which are surprisingly stable and accurate (Ornstein et al. [1991]).

Probabilistic, data-reconstruction objectives do not work.

Latent reconstruction losses avoid a decoder. (Xing et al. [2025]) argues for learning from the pixel space.

World model should be used in MPC instead of RL, latter requires too many trials.

MPC can suffer from practical limitations (Xing et al. [2025]). RL learns a policy function reflecting long-term cumulative rewards, enabling more strategic reasoning over extended time horizons (Xing et al. [2025]).

Proposed WM framework

- Drawing on this analysis, (Xing et al. [2025]) argues that a general purpose WM should follow the following design principles:

1. Use data from all modalities of experience, including text.
2. Employ a mixed continuous and discrete latent representation.
3. Adopt a hierarchical generative architecture (specialised generative layers for different levels of abstraction / latent prediction).
4. Be trained with a generative, observation-grounded loss. (encoder - WM - decoder)
5. Be used to simulate future trajectories for offline agent learning via RL.

The PAN model

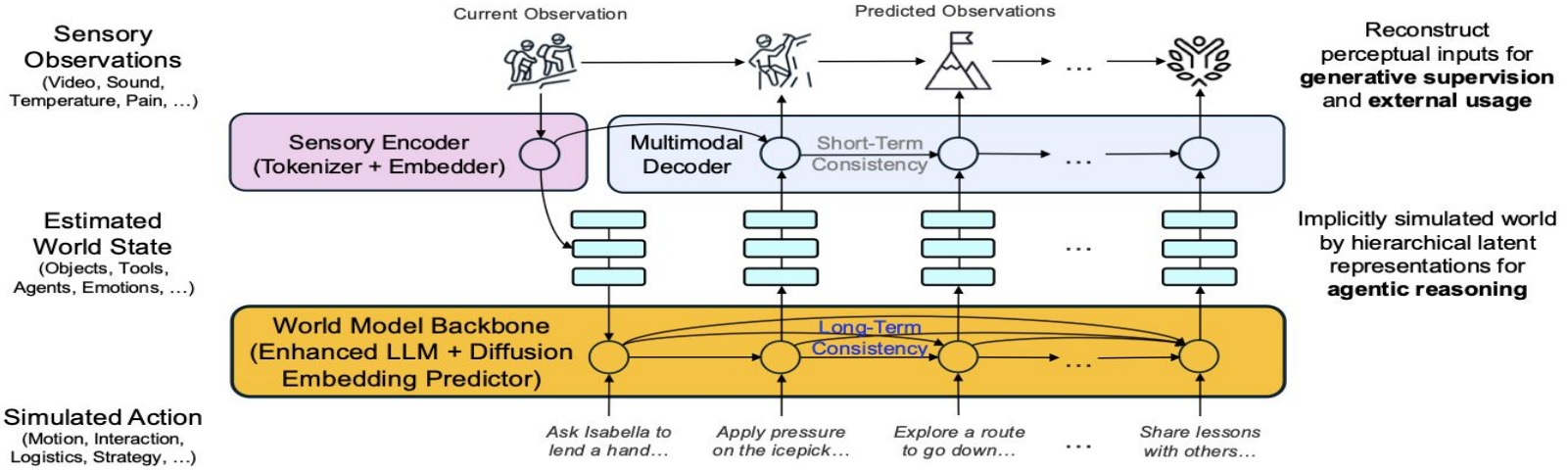


Figure 10: Illustration of the architecture of PAN-World, our proposed framework for general-purpose world modeling. At each time point, PAN-World estimates the world state $\hat{s}$ based on the current sensory observation o, predicts the future world states $\hat{s}'$ based on proposed actions a for agentic reasoning, and reconstructs the future observations $\hat{o}'$ for generative supervision and external usage.

Xing et al. [2025]

The PAN model

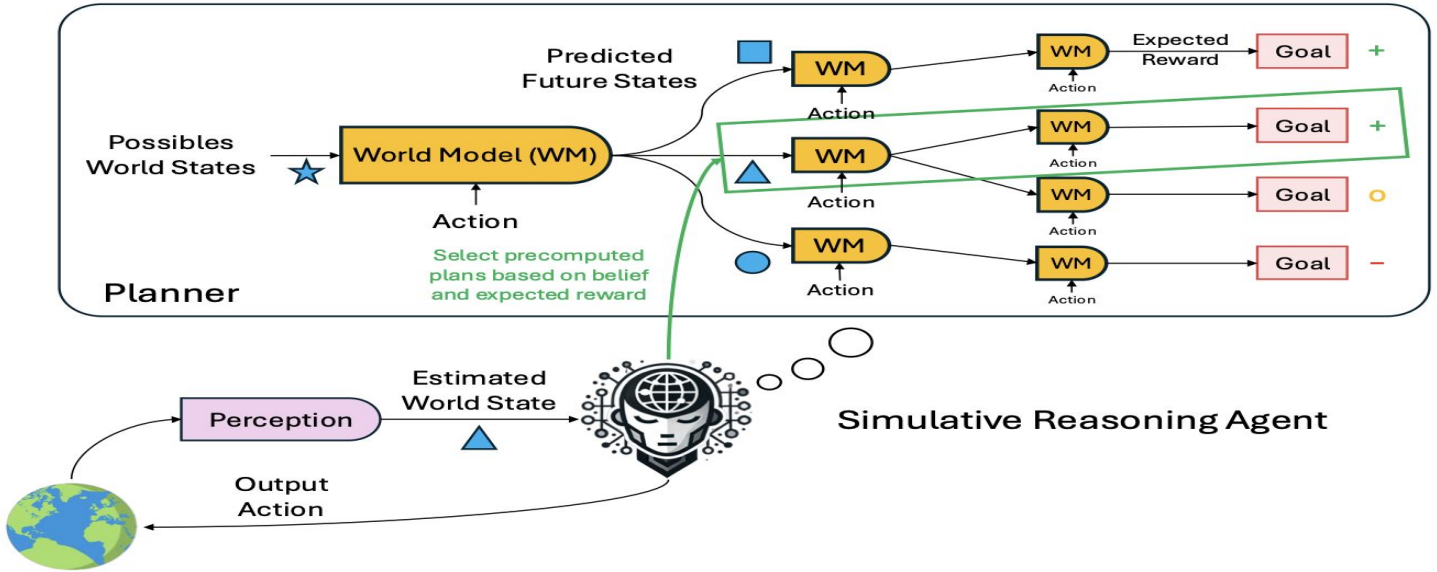


Figure 11: Illustration of our proposed simulative reasoning agent powered by the PAN World Model. Unlike traditional RL agents that rely on reactive policies, or model-precitive control (MPC) agents that expensively simulates futures at decision time, this agent leverages a cache of precomputed simulations generated by the PAN WM. During decision-making, the agent selects actions based on its current beliefs and expected outcomes, enabling a more efficient, flexible, and intentional form of planning, which, as we argue, is closer to the flexibility of human reasoning.

Xing et al. [2025]

UNIVERSITY OF OXFORD IMPERIAL

References (1/2)

- Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985.

- Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In Proceedings of the IEEE international conference on computer vision, pages 5842–5850, 2017.

- Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. CoRR, abs/2110.09348, 2021. URL https://arxiv.org/ abs/2110.09348.

- Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. CoRR, abs/1906.03327, 2019. URL http://arxiv.org/abs/1906.03327.

- David JC MacKay. Information theory, inference and learning algorithms. Cambridge university press, 2003.

- Donald S Ornstein and Benjamin Weiss. Statistical properties of chaotic systems. Bulletin of the American Mathematical Society, 24(1):11–116, 1991.

References (2/2)

- NVIDIA, :, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei- Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai, 2025. URL https://arxiv.org/abs/2501.03575.

- Shi Pu, Kaili Zhao, and Mao Zheng. Kinetics-700 dataset, dec 2024. URL https://service.tib. eu/ldmservice/dataset/kinetics-700-dataset.

- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

- Eric Xing, Mingkai Deng, Jinyu Hou, and Zhiting Hu. Critiques of world models, 2025. URL https://arxiv.org/abs/2507.05169.

- Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning, 2025. URL https://arxiv.org/abs/2411.04983.

- Linden J. Ball. Hypothetical Thinking, page 514–528. Cambridge Handbooks in Psychology. Cambridge University Press, 2020.

UNIVERSITY OF OXFORD IMPERIAL

One-liners for why (a)-(e) common school of thoughts is challenged

(a) ***World state should be represented by continuous embeddings rather than discrete tokens (for gradient-based optimisation).***

(b) Indeed continuous embeddings allow for smoother gradient flow, but the argument overlooks inherent noise and high variability with continuous sensory inputs. Human cognition has evolved to counter this variability by categorising raw perception into discrete concepts typically encoded in language, symbols, structured thoughts. Space of language is man-made latent space of percievable and describable universe which humans live in, substantial subspace of whole universe. Can discete rep faithfully capture richness of high-dim, cts sensory data due to risk of info loss thro distillation. Theory shows yes in principle can preserve arbitrary fine distinctions between real-valued inputs, provided scaled appropriately. Given discrete and cts latent representations offer complementary levels of abstraction, rep power and operational viability, mixed representations. Leverage discrete tokens for more robust, interpretable and symbolic forms of reasoning. Continuous embedding role: capturing fine-grained sensory nuance

‘Arguement overflooks inherent noise and high variability with continuous sensory inputs, we counter this by categorising raw perception into discrete concepts typically encoded in language, symbols and structured thoughts/. In principle discrete reps can capture richness of high dim, cts sensory data given scale appropriately. Propose mixed rep to leverage discrete tokens for robust, intepretable symbolic forms of reasoning and cts embeddings for fine grained sensory nuance.

One-liners for why (a)-(e) common school of thoughts is challenged

(a) Text captures mental, social and counterfactual phenomena and provides an interface to collective human memory which is difficult to derive from raw perceptual input alone [42c]. Path towards general-purpose world modeling must leverage all modalities of experience as each modality captures a different level of experience, overemphasising certain modalities over others reflects limitation or bias over fundamental perception of what world model is and what it can do successful WM must learn from these structures of experiences to generalise across diverse tasks, ignoring any level, be it low-level perception or high-level abstraction, risks omitting crucial information needed for intelligent behaviours.

‘Text captures mental, social phenomena and high-level abstraction. If want GP WM must leverage all modalities as each capture a different level of experience. Overemphasising certain modalities over others reflects limitation or bias.

UNIVERSITY OF OXFORD IMPERIAL

One-liners for why (a)-(e) common school of thoughts is challenged

(a) Autoregressive, generative models (e.g. LLMs) ‘doomed’.

Autoregressive generative model: autoregressive here means automatically predict the next component in a sequence by taking measurements from previous inputs in the system, generative is capable of creating new and unique content (for example, GPT in LLMs, deep NN consisting on encoder-decoder, enables natural language understanding AND natural language generation resp, GPT uses only decoder for autoregressive language modeling. Allows GPT to understand natural languages and respond in ways humans comprehend A GPT-powered large language model predicts the next word by considering the probability distribution of the text corpus it is trained on.

Compounding errors issue

Many real world systems fundamentally chaotic, small deviations which grow exponentially with time BUT autoregressive models can still learn abstract properties of system surprisingly stable and predictable

Systems using latent reconstruction losses justify this by saying generative reconstruction losses involving decoder reconstruct aspects of environment either inherently unpredictable or irrelevant to task performance (fine-grained visual details, out-of-scence content) which may mislead the model into learning unstable or spurious correlations

Encoder-only architectures suggest by avoiding this reconstruction step, resulting WM can focus more selectively on predictable and task-relevant elements

But remains unclear if encoder-only architecture effective remedy (could be degradation of next-state predictive performance in face of data signal variability stems less from presence of generative decoders than from use of continuous embeddings itself which squeeze massive amounts of info into limited subspace with fixed dimension, instability also from energy-based loss functions used for next latent prediction, require heuristic-based regualizers, behaviour difficult to understand and control

Theory shows that latent loss <= generative loss + small reconstruction error (only equal when encoder and decoder perfectly invert each other), unrealistic in practice
Minimising latent loss does not guarantee consistency with observed data, usually required for minimising generative loss
Error term small, so latent <= gen implies former can miss semantically important mistakes that latter will penalise and less understood regualisers use in conjunction with latent as objective makes its outcome even more difficult to assess

Therefore argue for learning from pixel space, predicting new observation to ensure predicted latent representations meaningfully grounded in real world, reliable prediction in latent space depends on continual validation through observable data, generative reconstruction objectives tether learned representations to observable world, richer stable learning supporting meaningful distinctions, general usabillity and human interpretability

Critiques says instead of modelling full world at a single level of detail, GLP decomposes problem across multiple layers of latent prediction, each specialised for different representational granularites (e.g. continuous perceptual features or discrete conceptual tokens)
Each layer operates at appropriate level of abstraction whilst remaining generative and predictive

Probabilistic, data-reconstruction objectives do not work.

Latent reconstruction losses avoid a decoder as there is a concern that they compel model to reconstruct aspects of the environment that are unpredictable or irrelevant to task performance. By avoiding a decoder, latent reconstruction losses suggest the resulting WM can focus more selectively on predictable and task-relevant elements. However, it is unclear if removing the decoder is the effective remedy. Theory also shows the result below, where the inequality is equal only when the encoder and decoder perfectly invert each other. Hence, minimising the latent loss does not guarantee consistency with the observed data. A small error term usually implies the former can miss semantically important mistakes the latter would penalise.

Hence, [c] argues for learning from the pixel space. They use a Generative Latent Prediction (GLP) architecture which instead of modelling the full world at a single level of detail, decomposes the problem across multiple layers of latent prediction.

$$\mathcal{L}_{\text{latent}}(h, f) = \mathbb{E}_{(o,a,o')\sim\mathcal{D}}\left[\|f(h(o), a) - h(o')\|\right],$$

$$\mathcal{L}_{\text{gen}}(h, f, g) = \mathbb{E}_{(o,a,o')\sim\mathcal{D}}\left[\|g \circ f(h(o), a) - o'\|\right].$$

UNIVERSITY OF OXFORD IMPERIAL

World model should be used in MPC instead of RL, latter requires too many trials.

- What is MPC *talk about*
- Include the loss

- Its appeal

Appeal lies in learning from offline trajectories without potentially unsafe exploration in the real world, higher quality decision-making from WM based simulation.

MPC can suffer from practical limitations, such as high computational overload and difficulty to respond effectively in fast-changing environments. *Talk about why (the reconstruction of trajectories AND small planning times in terms of horizon searching).*

MPC has shown promise primarily in simplified settings, where environment dynamics is simpler and slower decision making is rewarded but does struggle to extend to real-world tasks which typically involve complex dynamics and require a mixture of short- and long- term decision-making.

RL is a general, flexible and scalable approach to training agents without restrictions on decision-making or search horizon. RL learns a policy function reflecting long-term cumulative rewards, enabling more strategic reasoning over extended time horizons.

such as high computational overload and difficulty to respond effectively in fast-changing environments. RL is a general, flexible and scalable approach to training agents without restrictions on decision-making or search horizon.

Sample Body

- How do you tell when coffee is sick? It keeps coffee-ng! (Harleen et. al)
- What does the tissue say when it has a cold? A-tissue! (Harleen et. al)
- Copy and paste this slide to make your body text!

References

- Harleen, James, Chris. *1001 Jokes*. 2025

UNIVERSITY OF OXFORD IMPERIAL

How do the discussed models compare

	All data modalities	Mixed latent representations	Hierarchical generative architecture	Reconstruction loss	RL policy training
V-JEPA					
DINO-WM					
COSMOS	✅		✅		