

Critiques of World Models

Eric Xing^{◊,†*}, Mingkai Deng^{◊,†*}, Jinyu Hou^{◊,†}, Zhiting Hu^{◊,‡}

[◊]Institute of Foundation Models, Mohamed bin Zayed University
of Artificial Intelligence

[†] School of Computer Science, Carnegie Mellon University

[‡]Halicioğlu Data Science Institute, UC San Diego

eric.xing@mbzuai.ac.ae, {mingkaid, jinyuhou}@cs.cmu.edu,
zh019@ucsd.edu

July 2025

Abstract

World Model, the supposed algorithmic surrogate of the real-world environment which biological agents experience with and act upon, has been an emerging topic in recent years because of the rising needs to develop virtual agents with artificial (general) intelligence. There has been much debate on what a world model really is, how to build it, how to use it, and how to evaluate it. In this essay, starting from the imagination in the famed Sci-Fi classic *Dune*, and drawing inspiration from the concept of “hypothetical thinking” in psychology literature, we offer critiques of several schools of thoughts on world modeling, and argue the primary goal of a world model to be *simulating all actionable possibilities of the real world for purposeful reasoning and acting*. Building on the critiques, we propose a new architecture for a general-purpose world model, based on hierarchical, multi-level, and mixed continuous/discrete representations, and a generative and self-supervised learning framework, with an outlook of a Physical, Agentic, and Nested (PAN) AGI system enabled by such a model.

1 Introduction

A Large Language Model (LLM) simulates the **next word** in human languages, which has led to systems like ChatGPT that allow people to perform a wide range of tasks facilitated through language, such as common conversation, standardized tests, professional writing, and advanced math reasoning, at a level rivaling human intelligence.

What would you do if you could perfectly simulate the **next world** – every possible future in the environment that we reside? *Dune*, a classic of science fiction that inspired the likes of George Lucas’ *Star Wars* and Miyazaki’s *Valley of the Wind*, boldly imagines such a possibility. The series is centered around the *Kwisatz Haderach*, a prophesized human being who inherits their ancestors’ memories and simulates the outcomes of all possible plans in order to chart the best path to achieve their goals [25]. Such superhuman ability allows them to command armies to win galactic wars, or to oversee global-scale projects that turn a desert planet into a green paradise. Is it possible to build towards computer systems with similar functionalities using a similar approach?

* Co-first author

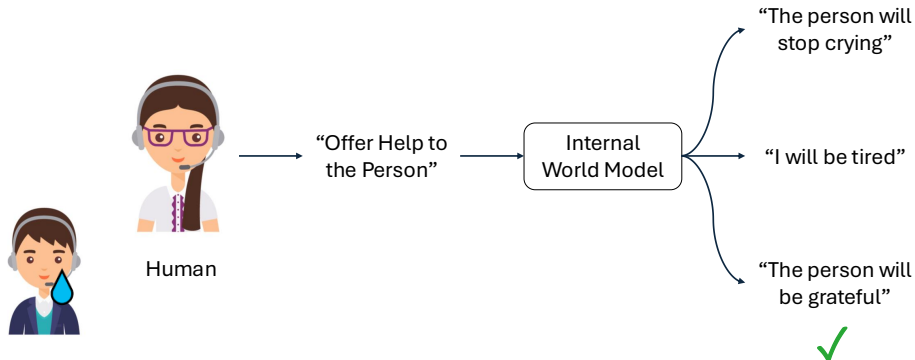


Figure 1: Familiar example of reasoning by simulation – an individual (possibly self-serving) decides to offer help to a crying person by mentally simulating multiple possible outcomes, with the best expected reward in mind.

Unlike a chat-bot, human consists of a hierarchy of abilities that go from immediate, concrete ones (e.g., body control/movement/action, reading/listening, and speaking/drawing) to far-reaching and abstract ones (e.g., planning, collaboration, and strategizing). Furthermore, the same human, while not necessarily perfectly, can perform a broad range of tasks (e.g., do household chores, complete risky expeditions, conduct research investigations, navigate social situations, and manage complex enterprises) all with the same cognitive architecture of the human brain. Can a single artificial intelligence (AI) system perform all these tasks? Each of these problems can be seen as a goal-oriented agent acting in a multimodal environment, requiring purposeful reasoning of massive temporal-spacial, social-physical, and emotional-cognitive complexity and depth, to the point that traditional approaches built on logical inference (e.g., induction, deduction, abduction) are often easily overwhelmed. It emerges that the key to generalized decision-making toward such complexity lies in reasoning by “seeing the future” as seen in *Dune* – an ability formally known as *Hypothetical Thinking* [4] in the psychology literature, or “thought experiments” in common practice – the ability to simulate the next worlds using a mental model of the world. We call such a mental model a **World Model**.

Specifically, a World Model (WM) is a generative model that simulates the possibilities in diverse scenarios (e.g., physical world, mental world, social world, and evolutionary world). Operationally, a WM takes previous world state s and action a , and predicts or simulates the next world state s' through a transformation function, such as a conditional probability distribution:

$$s' \sim p(s'|s, a) \quad (1)$$

With a WM, machines can perform thought experiments by simulating actions and plans in complex scenarios, including counterfactual ones, and extracting the best ones among them. This is consistent with the hypothesis that humans reason not just linearly with formal rules toward goals (e.g., imagine a self-serving individual immediately offering help to another person upon seeing her cry in the hopes of stopping her from crying, Figure 1) via a deterministic optimization algorithm, but also based on simulations using an internal mental model (e.g., imagine the same individual decides what to do by mentally *simulating* multiple possible outcomes including self-exhaustion, the other person stops crying, she continues to cry but is grateful, etc., with the best expected reward in mind) [21, 40].

Such a WM also enables transfer of knowledge to solving novel tasks, thanks to the fact that real world dynamics, even in different scenarios, share many mechanistic commonalities. For instance, a scuba diver experiences how their body moves in response to low gravity, which is likely helpful to

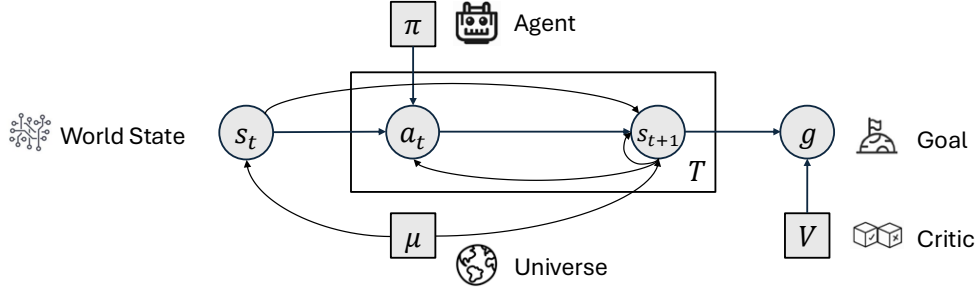


Figure 2: A possible definition of an optimal agent

walking on the moon. A mountaineer would be good at predicting how terrain conceals individual movements, which is useful when it’s their turn to lead a mountain ambush. On the other hand, a skilled gamer has deep knowledge of how digital characters respond to control signals, which will come in handy when they become drone operators. Therefore, like humans often employ their mental model to help extrapolate from past experiences to act novelly in new environments; machines can leverage a WM similarly to better achieve zero-shot capabilities in unfamiliar environments.

How should we create such a general WM? The key desiderata for building and training a WM include the following 5 aspects: identifying and preparing training **data** with the desired world information; adopting a general **representation** space for the latent world state with possibly richer meaning than the observation data in plain sight; designing an **architecture** that allows effective reasoning over the representations; choosing an **objective** that properly guides the model training; determining how to **use** the world model in a decision-making system. Recent years have seen a surge in efforts toward a world model. In this paper, we provide both empirical and technical critiques of several such efforts, including some very vocal schools of thoughts on WM where systematic proposals on the five aforementioned aspects of WM were offered.

We conclude our critiques with a brief preview of an alternative architecture – PAN – a Physical, Agentic, and Nested world model, which we argue offers the potential for a truly general-purpose and actionable world model, based on the following design principles: 1) data from all modalities of experience; 2) mixed continuous and discrete representation; 3) hierarchical generative modeling with an enhanced LLM backbone, alone with and a generative latent predictive architecture; 4) generative loss grounded in observation data; and 5) using world model to simulate experience for training agents with reinforcement learning (RL). Full details of the PAN world models and results will be provided in a separate dedicated manuscript [19].

2 World Model and Agent Decision-Making

World model arises in the context of agent decision-making. An agent is an autonomous system that acts in an environment – the *universe*, with both physical and social worlds – to achieve a goal (e.g., climbing a mountain, winning a military campaign). We consider an environment with discrete timesteps indexed by t (continuous timesteps can be approximated by infinitesimally small discrete time-steps). Formally, the agent takes the current world state s_t and outputs the next action a_t based on a distribution $p_\pi(a_t|s_t)$, known as the “policy” in reinforcement learning literature. The optimal agent is thus one that best achieves its goals across all environments. The concept of a World Model arises as a surrogate of the environment in general agentic reasoning.

2.1 Agent-Environment Model and Optimal Agent

Consider a sequential interaction between an agent and an environment (Figure 2). At time step t , the agent outputs action a_t , and the universe μ takes the current state s_t and the action a_t , and outputs the next state s_{t+1} based on distribution $p_\mu(s_{t+1}|s_t, a_t)$. The distribution of the interaction trajectory until timestep T , or $(a_t, s_{t+1}, \dots, a_{T-1}, s_T)$, given the current state s_t is thus denoted by:

$$p_\mu^\pi(a_t, s_{t+1}, \dots, s_T | s_t) = \prod_{k=t}^{T-1} \underbrace{p_\pi(a_k | s_k)}_{\text{agent}} \underbrace{p_\mu(s_{k+1} | s_k, a_k)}_{\text{universe}} \quad (2)$$

In each state s_t , the agent also receives a reward $r(g, s_t)$ based on its goal g . We evaluate the agent by its discounted cumulative reward, denoted as $\sum_{k=t}^{\infty} \gamma_k r(g, s_k)$ (with the discount parameter γ_t decaying to zero with time, i.e., $\lim_{t \rightarrow \infty} \gamma_t = 0$). Note that this reward function can be dense (e.g., gaming scores), but perhaps frequently sparse (e.g., curing a disease). The agent’s long-term success can thus be measured by its expected future discounted reward, also known as **value function** [38]:

$$\begin{aligned} V_{\pi, \mu}^g(s_t) &:= \mathbb{E}_{\pi, \mu} \left[\sum_{k=t}^{\infty} \gamma_k r(g, s_k) \mid s_t \right] \\ &= \lim_{T \rightarrow \infty} \sum_{(a_t, s_{t+1}, \dots, s_T)} \underbrace{\sum_{k=t}^T \gamma_k r(g, s_k)}_{\text{goal}} \underbrace{p_\mu^\pi(a_t, s_{t+1}, \dots, s_T | s_t)}_{\text{trajectory}} \end{aligned} \quad (3)$$

Based on Equations 2 and 3, we can define the optimal agent in this universe μ as one that maximizes the value function, written formally as below:

$$\pi_\mu^* := \arg \max_{\pi} V_{\pi, \mu}^g \quad (4)$$

Some simple derivation will show that the optimal agent in state s_t will select actions based on the following decision rule $\pi_\mu^*(s_t)$ when planning for actions $a_{t:T-1}$:

$$\pi_\mu^*(s_t) = \underbrace{\arg \max_{a_{t:T-1}}}_{\text{possible actions}} \sum_{s_{t+1:T}} \underbrace{\left(\sum_{k=t}^{T-1} \gamma_k r(g, s_k) + \gamma_T V_{\pi, \mu}^g(s_T) \right)}_{\text{goal progress}} \prod_{i=t}^{T-1} \underbrace{p_\mu(s_{i+1} | s_i, a_i)}_{\text{universe response}} \quad (5)$$

2.2 World Model and Simulative Reasoning

Note that optimal decision-making defined in Equation 5 requires the agent to have access to the ground-truth world state s from the universe μ to experience and optimize. However, these are often not the case aside from simple scenarios like the Go and Chess games [33, 34] – imagine building an agent to land on Mars, or even a real-world robot relying on noisy sensors in daily environments. World Model thus arises as a crucial component for predicting the universe’s response to a general agent. Specifically, as illustrated in Figure 3, a WM f operates on an internal (continuous or discrete) representation of the world state, denoted as a *belief state* \hat{s}_t , which is derived from sensory inputs o_t via an *Encoder* h (unlike the optimal agent described in §2.1 which has direct access to the true world state s_t). Given a proposed action a'_t (as opposed to the true action a_t used by the optimal agent), the WM predicts the next belief state \hat{s}_{t+1} according to the distribution $p_f(\hat{s}_{t+1} | \hat{s}_t, a'_t)$. This predicted belief state then allows the agent to propose the next action, continuing the cycle of prediction and action up to a desired time horizon T' .

The agent can simulate multiple such sequences of proposed actions and belief states, and select the actual action a_t (upon observing o_t) based on some external function such as a *Critic* that evaluates the outcome against a given *Goal*. Thus, a WM essentially functions as a generative

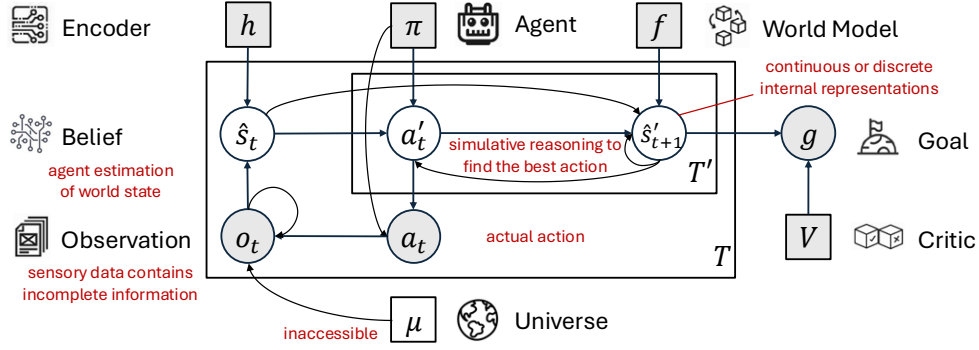


Figure 3: An agent in real world where groundtruth world state and universe are unavailable to experience or experiment, so world model is crucial for simulation.

model of possible future world states, which enables simulative reasoning, or ”thought experiments”. Formally, for the optimal agent π_f^* equipped with WM f in belief state \hat{s}_t , we define the simulation-based decision rule in Equation 6 as follows:

$$\pi_f^*(\hat{s}_t) = \underbrace{\arg \max_{a'_{t:T'-1}}}_{\text{possible actions}} \sum_{\hat{s}_{t+1:T'}} \underbrace{\left(\sum_{k=t}^{T'-1} \gamma_k r(g, \hat{s}_k) + \gamma_{T'} V_{\pi, f}^g(\hat{s}_{T'}) \right)}_{\text{goal progress}} \prod_{i=t}^{T'-1} \underbrace{p_f(\hat{s}_{i+1} | \hat{s}_i, a'_i)}_{\text{simulation with world model}} \quad (6)$$

A general-purpose WM enables simulation of diverse possibilities across a wide range of domains, enabling agents to reason about outcomes without direct interaction with the environment. This includes, but is not limited to the following examples:

- **Physical dynamics:** Mechanics of the real world, such as how water pours, how an object moves when thrown, or how a machine operates under varying conditions.
- **Embodied experiences:** Internal bodily states (e.g., balance, posture), sensations (e.g., heat, pain, dizziness), and complex motor activities like getting dressed or tying shoes.
- **Emotional states:** Affective responses such as happiness, sadness, or fear, which can facilitate planning in emotionally charged contexts (e.g., therapy or social interactions).
- **Social situations:** The actions and internal states of other individuals, including their embodied or emotional experiences, needs, intentions, and expectations.
- **Mental world:** Abstract ”thought processes” such as logistics, tactics, and strategies, potentially in multi-agent or adversarial settings.
- **Counterfactual world:** Alternative realities or ”what if” scenarios to guide better decision-making under uncertainty or incomplete information.
- **Evolutionary world:** Generational dynamics such as genetic inheritance, adaptation, and survival of organisms.

As stated earlier, a major function of the WM is to enable *simulative reasoning* – where an agent performs a series of thought experiments by simulating the outcomes of plans with the WM, based on which it can choose the best plan. This reasoning approach contrasts with alternative approaches also explored in AI systems, such as *logical reasoning* inspired by a type of human mental activity that aims to arrive at a conclusion through rigorous inferences or arguments starting from a set

of premises and then using stepwise relational formal reasoning (e.g., Lambda calculus) to reach a conclusion supported by these premises; or *model predictive control* based reasoning used in the process industries in chemical plants and oil refineries to infer optimal action sequence through mathematical programming (e.g., convex optimization) to satisfy a set of constraints. A hallmark of simulative reasoning enabled by a WM is its flexibility, generalizability, and scalability with respect to changing computing resource, memory, environment, and problem complexity, thanks to the WM’s intrinsic ability to simulate all possibilities across domains. In this regard, a WM bears important empirical (e.g., end-to-end experience) similarity to a modern LLM such as ChatGPT that can operate in a subject-agonistic way in the lingual intelligence space. Indeed, as we discuss later, a general-purpose WM can make use of LLMs as its key building blocks.

Combined with an encoder that estimates beliefs of the world states from arbitrary sensory observations, WM supports machines to perform thought experiments computationally with controlled depth (i.e., number of steps) and width (number of trajectories). AlphaGo [33], for example, can be seen as a special case of simulative reasoning with Monte-Carlo Tree Search (MCTS) using a known (trivial) WM. In the physical world, simulative reasoning enables autonomous vehicles to drive safely by forecasting future street scenarios [e.g., 41, 37] or military commanders to develop battle-winning tactics by anticipating the outcomes of troop movements [36]. WM also enables simulations on different time scales, allowing one to answer questions about billions of years on Earth’s evolutionary path, or just a few moments on a hypothetical Martian civilization.

3 The World Model Landscape

Recent work on world modeling have led to a variety of systems, many of which optimized for a specific domain or type of simulation. Interestingly, a commonality nevertheless can be found in these diverse systems in that they all put a significant emphasis on video/image generation, and on visual quality of the generated contents:

Gaming World Models Systems such as Genie 2 [28] (Google DeepMind), Muse [22] (Microsoft), and Oasis [11] (Decart and Etched) simulate video game environments using generative models. These models can render plausible trajectories from visual and action input, producing up to 1-2 minutes of continuous gameplay content. Despite their advances, these systems remain domain-specific – for instance, Genie 2 and Muse rely on restrictive console-style inputs, while Oasis is limited to Minecraft-like settings. Furthermore, their temporal coherence remains shallow, as their current generation horizons (1-2 minutes) fall short of representing full gameplay sessions, which often span several hours. As such, current gaming world models lack the flexibility, generality, and long-term reasoning capabilities required for more open-ended, agent-driven tasks.

3D Scene World Models World Labs [47] and related efforts focus on stylized 3D scene generation and egocentric navigation. While visually appealing, technical details remain scarce. From available demonstrations, these models appear to simulate static environments without dynamic agents, physics, or rich interactivity. This results in incomplete simulations that are insufficient for tasks involving physical causality, multi-agent behavior, or goal-driven planning. Although these systems push the boundary of spatial realism, they do not support full-fledged world modeling for decision-making or agent learning.

Physical World Models Generative models such as Wayve GAIA-2 [32] and NVIDIA Cosmos [1] are trained specifically on physical control tasks, including autonomous driving, robotic manipulation, and embodied navigation. These systems demonstrate impressive fidelity in modeling low-level physics and sensory-motor control under diverse conditions (e.g., varying weather, lighting, and

geography). However, they are tightly coupled to their respective domains, often relying on task-specific sensors, data, or control architectures. These models excel in their respective constrained settings, but still yet to address the broader challenge of simulating complex, multi-agent, or socially grounded worlds.

Video Generation Models Another popular class of models focus on general-purpose video generation, with recent examples including OpenAI’s Sora [7] and Google DeepMind’s Veo [15]. These models aim to generate high-quality video sequences from textual prompts and/or prior frames. While visually stunning, these models only generate fixed trajectories and do not support interactions based on alternative actions. Specifically, they lack explicit notions of state, action, and possibly even object-level or conceptual representation within the video frames. They also provide no simulation control that would allow for reasoning about counterfactual outcomes or evaluate different decisions. Consequently, these systems fall outside the definition of world models for reasoning and planning, and are better understood as strict video-generation tools (focused on pixel-level synthesis) rather than parts of decision-making systems.

Joint Embedding Predictive Models Last but not least, the joint embedding predictive architecture (JEPA) family, including the V-JEPA series [5, 3], DINO-WM [50], and PLDM [35] from Meta FAIR, has attracted significant attention for its conceptually elegant approach to world modeling. These models forego pixel-level generation and instead predict future latent embeddings, often using an encoder-encoder architecture and supervising the outputs in the latent space with energy-based losses. While this design promises to improve tractability, the evidence for practical usability remains scarce as these models have mainly been demonstrated in toy environments with simple heuristics and action spaces. The very recent V-JEPA 2 [3] marks a step forward by applying joint embedding prediction to robotic arm manipulation tasks, but it remains unclear whether such models can generalize across more diverse tasks (e.g., making breakfast) or scale to higher-complexity environments with long-term dependencies (e.g., mountaineering).

In summary, although the systems surveyed above have demonstrated significant progress in modeling aspects of the world, most fall short of enabling purposeful reasoning and planning in real world applications due to limitations in scope, abstraction, controllability, interactability, and generalizability. In particular, except for JEPA, contemporary WM systems almost universally emphasize video generation as a core function, yet this emphasis remains largely unexamined and lacks a compelling justification. This focus may reflect the underlying conceptual ambiguities, or even misunderstandings, regarding a fundamental question – what *is* a world model? We argue that **a world model is *not* fundamentally about generating videos, but about serving as a sandbox for reasoning and thought-experiment**. Our discussion below is organized around this definition to examine the plausibility and feasibility of current technical approaches to world modeling.

4 Critiques of World Modeling

A vocal school of thought [23] has made the following claims along the five dimensions in building a world model: data, representation, architecture, objective, and usage, which we found to warrant closer examination.

1. **Sensory inputs are superior to texts**, because of the larger data volume from physical world (i.e., “even a 4-year-old would have processed 1.1^{14} bytes of visual data”, whereas all textual data used to train modern LLMs amount to merely 0.9^{14} bytes).
2. **World state should be represented by continuous embeddings** instead of discrete tokens to enable gradient-based optimization.

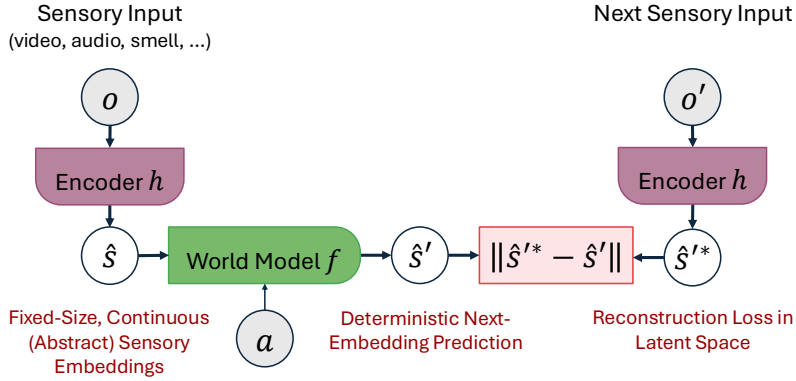


Figure 4: Framework for world model proposed by a vocal school of thought.

3. **Autoregressive, generative models (e.g., LLMs) are doomed**, as they are guaranteed to make mistakes eventually and can't model outcome uncertainties.
4. **Probabilistic, data-reconstruction objectives (e.g., an encoder-decoder scheme) do not work**, as they are intractable and force the model to predict irrelevant details.
5. **World models should be used in model-predictive control (MPC)**, instead of a reinforcement learning (RL) framework, as the latter requires too many trials.

This school of thought also proposes the following alternative framework for world models, as illustrated by Figure 4, centered around an idea that can be summarized as “next representation prediction” rather than next data prediction:

- **Text-Free Pretraining:** The proposed framework completely eschews text data in favor of continuous sensory data like video, audio, smell, etc.
- **Fixed-Size, Continuous State Embeddings:** Given sensory input o , an encoder h estimates the world state $\hat{s} = h(o)$ as an abstract continuous embedding with fixed dimensions (e.g., $\hat{s} \in \mathbb{R}^d$).
- **Encoder-Encoder Architecture:** Based on action input a , the world model f predicts the next state embedding $\hat{s}' = f(\hat{s}, a)$ in a deterministic manner. In particular, the architecture does not use a decoder g to reconstruct the next observation o' , but instead applies the encoder again to bootstrap $\hat{s}'^* = h(o')$ as a *ground-truth next state* for supervision.
- **Reconstruction Loss in Latent Space:** Instead of supervising the world model by the difference between the reconstructed next sensory input \hat{o}' and the actual data o' , the proposed framework bases learning on the deviation between the predicted next state \hat{s}' and the bootstrapped groundtruth \hat{s}'^* (e.g., L2 loss $\|\hat{s}' - \hat{s}'^*\|$).
- **Action Selection via MPC:** Given current observation o_t , the framework favors proposing an initial action sequence $(a_t, a_{t+1}, \dots, a_{T-1})$, using the world model f to simulate the next states $(s_{t+1}, s_{t+2}, \dots, s_T)$, and optimizing the actions based on goal progress $V_g(s_T)$.

While these thoughts raise valid questions on some current practices in world modeling and promises appealing solutions, we argue that each of their underlying assumptions introduces critical limitations when applied to general, scalable, and robust world modeling for the purpose of agentic reasoning and decision making. In the following discussion, we offer analytical critiques of the claims and proposals presented above along the dimensions of **data**, **representation**, **architecture**, **objective**, and **usage** of building a world model.

4.1 Data: Information Density, Not Just Volume

World model needs to be trained from sensory inputs due to larger data volume: LLM with 0.9^{14} bytes textual data vs. 4-year-old with 1.1^{14} bytes vision data.

Although sensory data streams such as video appear massive in raw volume, much of that data is low in semantic content and highly redundant [14, 44]. In contrast, natural language is an evolved compression of human experiences, optimized over generations of abstract communication and conceptual reasoning [24, 30].

Text captures not only physical realities but also mental, social, and counterfactual phenomena, which are otherwise difficult or impossible to observe directly [10]. For example, concepts such as justice, motivation, or regret are richly encoded in language but have no direct sensory equivalent. Moreover, language provides an interface to collective human memory (e.g., documented observations, scientific discoveries, engineering failures), which are difficult, if not impossible, to derive from raw perceptual input alone [42]. Indeed, models trained on text can write software [43] or solve Olympiad-level math problems [8], whereas those trained on raw visual and motion data alone have been more suited for physical navigation [45] or manipulation tasks [2].

Thus, the path towards general-purpose world modeling must leverage all modalities of experience, whether it be text, images, videos, touch, audio, or more. Crucially, these modalities are not interchangeable, but reflect different layers of experience (e.g., video captures spatiotemporal dynamics in the embodied and physical world, while language encodes abstract concepts and social norms). Overemphasizing certain modality over others, such as video over text, reflects a limitation or bias over the very fundamental perception of what a world model is and what it can do. A successful world model must therefore learn from this stratified structure of experience to generalize across diverse tasks. Ignoring any level, be it low-level perception or high-level abstraction, risks omitting crucial information needed for intelligent behavior.

4.2 Representation: Continuous? Discrete? Or Both?

Do not use discrete tokens. The world state should be represented by continuous embeddings to enable gradient-based optimization.

Do humans perform gradient optimization (e.g., SGD) over continuous neural signals or pattern search (e.g., kNN) over discrete concepts during reasoning? We don't know for sure. What we do know is that reasoning can be cognitive or physiological, or both, and it might be unlikely to have one algorithm fit for all.

While continuous representations allow for smoother gradient flow, the argument overlooks the inherent noise and high variability associated with continuous sensory inputs, which makes them brittle for reasoning. Human cognition has evolved to counter this variability by categorizing raw perception into *discrete concepts* [6], which are what we typically encode in language, symbols, and structured thoughts (Figure 5, left).

Vocabulary-based tokens are thus not a liability, but an asset: they offer a stable, composable medium for representing concepts at various levels of abstraction. They form the foundation for designing and building language-based AI systems of today, such as the LLMs, which ground reasoning on sequence of discrete words which are human tokens that correspond to diverse perceptions from the universe (e.g., physical, mental, or social worlds), and allow a form of long-term memory to be employed by implementing (ideally, dynamically controllable) context length. Although not entirely accurate, it is reasonable to consider the space of language as a man-made (through evolution and learning) *latent space* of the perceived and describable universe that humans live – a substantial subspace of the whole universe. Benefitting from massive text-based pretraining, an LLM can learn to simulate contents in this latent space formed by natural language. Indeed, recent work that represents world states in natural language has seen success for reasoning and planning in a wide

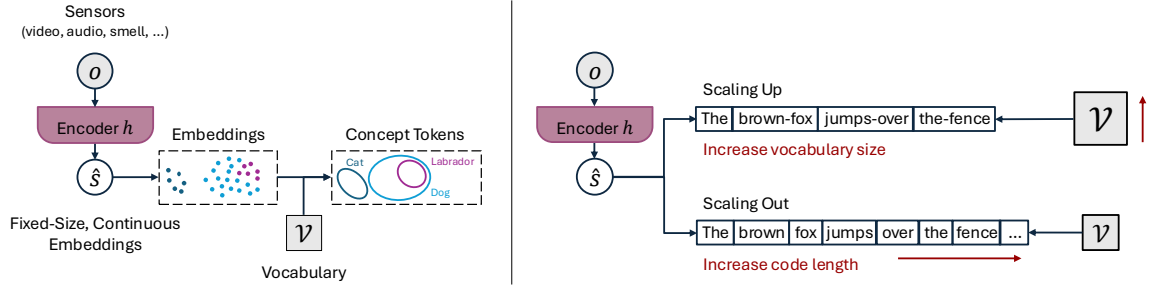


Figure 5: Vocabulary-based tokens is an effective way to categorize perceptual inputs into discrete concepts for reasoning (left). We may scale up or scale out discrete code to deal with increasing data complexity (right). Thm.1 shows either is effective, but scaling out is more efficient.

range of practical tasks [18, 12]. Complementing natural language tokens, modern techniques like VQ-VAE [39] allow us to further convert sensory data (e.g., images or audio) into discrete tokens while preserving structure and semantics.

While such discrete representations are expected to offer stability and symbolic structure like natural language tokens, a natural concern is whether they can faithfully capture the richness of high-dimensional, continuous sensory data due to the risk of information loss through distillation. This concern grows with the complexity of the world: Will discrete tokens be sufficient for distinguishing between subtly different world states? Indeed, the world often contains deeper layers of meaning than what is directly observable through sensory input (e.g., a puppet’s movements may reflect the hidden intentions of the puppet master). Capturing such latent structure requires representations that can scale in expressive capacity.

In an attempt to understand more deeply the potential and limitation of using discrete tokens, we present a theoretical result showing that discrete representations can, in principle, preserve arbitrarily fine distinctions between real-valued inputs, provided we scale them appropriately. Specifically, we consider two intuitive strategies for increasing representational capacity:

- **Learning a larger modality tokenizer** (scale up): Keep the number of tokens fixed, increase the vocabulary size to allow each token to encode a finer-grained chunk of information.
- **Finding a longer language expression** (scale out): Keep the vocabulary fixed, increase the sequence length and combine more tokens to express more complex inputs.

As we show in Theorem 1 below, it is more efficient to scale out by increasing the length of the encoding.

Theorem 1 (Completeness of Language Representation). *Assume real inputs $\mathbf{x} = [x_1, \dots, x_T]$, where $x_t \in \mathbb{R}^D$ and $\|x_t\| < K$. For any $\epsilon > 0$, there exists a language $L_\epsilon = (\mathcal{V}, N, f_\epsilon)$ with vocabulary \mathcal{V} , maximal sentence length $N < \infty$, and a mapping function $f_\epsilon : \mathbb{R}^{TD} \rightarrow \mathcal{V}^N$ such that for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^{TD}$, $\|\mathbf{x} - \mathbf{x}'\| > \epsilon \Rightarrow f_\epsilon(\mathbf{x}) \neq f_\epsilon(\mathbf{x}')$.*

Explanation. *If you have a sequence of continuous sensor readings or data points, no matter how small a difference you want to be able to distinguish between two sequences, you can always create a language (a system of words or symbols) that can represent these sequences uniquely.*

Proof Sketch. We will prove the contrapositive that $f_\epsilon(\mathbf{x}) = f_\epsilon(\mathbf{x}') \Rightarrow \|\mathbf{x} - \mathbf{x}'\| \leq \epsilon$. Specifically, we propose two ways to scale the discrete code:

- **Case 1** (Learning a Larger Modality Tokenizer). Keep the code length constant at T , increase the vocabulary size to $M_\epsilon = \lceil \sqrt{TDK\tilde{\epsilon}^{-1}} \rceil^D$ (i.e., **scaling up**).

- **Case 2** (Finding a Longer Language Expression). Keep the vocabulary size constant at M , increase the maximum sentence length to $N_\epsilon = TD \lceil \log_M \sqrt{TD} \tilde{K} \epsilon^{-1} \rceil$ (i.e., **scaling out**).

□

The detailed proof can be found in Appendix A. As the proof shows, complete representations are achievable with vocabulary-based discrete tokens. However, the way to scale the representation matters (Figure 5, right). In Case 1, the vocabulary size must grow in $\mathcal{O}((TD)^D)$, i.e., exponentially with the input size we’d like to capture, which is likely not sustainable. In Case 2, on the other hand, the sequence length need only increase in the order of $\mathcal{O}(TD \log TD)$, which is much more manageable. So in theory, scaling out the token sequences, which can be implemented with an enhanced LLM (with visual tokenizers and vocabulary mergers or switchers), offers a more flexible and efficient pathway to capturing complex structure in data generally. In practice, the model may also represent complex inputs efficiently by dynamically resizing the dictionary, and describe novel inputs by growing the vocabulary with more new observations.

In summary, given that discrete and continuous latent representations offer complementary level of abstraction, representation power, and operationalizability, we advocate for the approach of *mixed representations*, which leverages discrete tokens for more robust, interpretable, and symbolic forms of reasoning, while continuous embeddings still play a role in capturing fine-grained sensory nuance. While this form of representation is still in its early stages, recent work has demonstrated its promise for generalization in world modeling [48] and other forms of reasoning [e.g., 26].

4.3 Architecture: Autoregressive Generation is Not the Enemy

Abandon autoregressive, generative models; adopt joint embedding predictive architecture (JEPA) to avoid exponential compounding of token generation error, and absorb signal variability.

Proponents of JEPA advocate for a non-autoregressive, non-generative, encoder-encoder framework that predicts the next latent state directly, sidestepping the need to reconstruct raw observations. However, as we discuss in this section, the architecture of this framework remains fundamentally autoregressive and generative.

Formally, JEPA defines two core functions (Figure 6, left):

$$\hat{s} = h(o), \quad \hat{s}' = f(\hat{s}, a),$$

where h is an encoder from observations to latent states, and f is the world model which predicts the next latent state given the current state and an action. Recursive application of these two operators defines a latent transition model that is effectively autoregressive and generative, even though it symbolically lacks an explicit probabilistic decoder to generate what can be compared against real next observation data. (It does not mean such comparisons are avoided, as the second encoder at the output end, in fact, indirectly still makes that comparison, but with poorer mathematical controllability, as we discuss in the next section.) More precisely, JEPA can be viewed as specifying a degenerate conditional distribution, denoted informally as below:

$$p_f(\hat{s}' | \hat{s}, a) = \delta(\hat{s}' - f(\hat{s}, a)),$$

where $\delta(\cdot)$ is the Dirac delta function centered at the deterministic prediction. Thus, JEPA is not generative in the **probabilistic** sense (i.e., it does not model uncertainty or samples from a distribution over outcomes), but it is generative in the **functional** sense of recursively simulating the evolution of the latent states over time, and is therefore subject to the same issues of autoregressive models. This is not to say, however, that autoregressive models are inherently flawed due to error accumulation. Many real-world systems (e.g., three-body problem, fluid dynamics, or financial

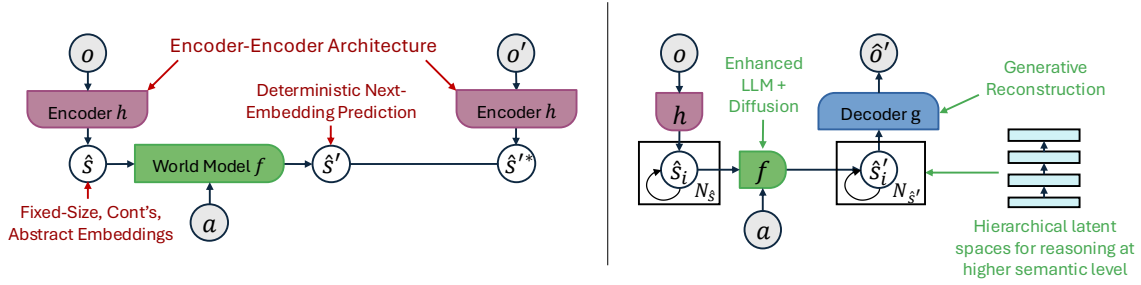


Figure 6: Comparison of the critiqued JEPA world model architecture (left) and our proposed Generative Latent Prediction (GLP) architecture (right), which includes a diffusion-based next-embedding predictor for modeling continuous perceptual dynamics and an enhanced LLM backbone for discrete long-horizon structures.

markets) are fundamentally chaotic, with small deviations growing exponentially over time [29]. In such settings, exact prediction is impossible regardless of model class. However, well-structured autoregressive models (e.g., Kalman filters for continuous cases and HMMs for discrete cases) can still learn useful, abstract properties of the system (e.g., whether water will spill, the direction of price movement) that are often surprisingly stable and predictable – an insight grounded in ergodic theory and statistical mechanics [27].

A common concern with encoder-decoder architectures (which defines an additional function $\hat{o}' = g(\hat{s}')$ with g as a decoder from latent states back to observations) is that they may compel the model to reconstruct aspects of the environment that are either inherently unpredictable or irrelevant to task performance. Examples often cited include fine-grained visual details, inconsequential events, or out-of-scene content, which may mislead the model into learning unstable or spurious correlations. Proponents of encoder-only architectures thus suggest that by avoiding this reconstruction step, the resulting WM can focus more selectively on predictable and task-relevant elements. While this motivation is understandable, it remains unclear whether removing the decoder is the effective remedy. In such an architecture as JEPA, supervision occurs solely in the latent space rather than observation space, trading off challenges of pixel-level variability for the risk of indefinability: the predicted latents are not directly grounded in observable data, which makes it difficult to diagnose whether the model is learning meaningful dynamics or collapsing to trivial solutions, an issue we discuss formally in §4.4.

It is possible that the degradation of next-state prediction performance in face of data signal variability stems less from the presence of generative decoders than from the use of continuous embeddings itself, which squeezes massive amounts of information into a limited subspace with fixed dimension (§4.2). Additional instability may arise from the energy-based loss functions typically used for next-latent prediction, which often require heuristics-based regularizers whose behaviors are difficult to understand and control. Furthermore, the challenge of signal variability may be particularly acute in visual-related domains, while many downstream reasoning tasks (including, for example, autonomous driving) may not demand pixel-perfect simulation of the visual world. Therefore, rather than abandoning generative modeling to avoid signal variability, an alternative and well-established strategy is to adopt **hierarchical abstraction** through what we call a Generative Latent Prediction (GLP) architecture (Figure 6, right). Instead of modeling the full world at a single level of detail, GLP decomposes the problem across multiple layers of latent prediction, each specialized for different representational granularities, whether it be continuous perceptual features or discrete conceptual tokens. This allows each layer to operate at an appropriate level of abstraction while remaining generative and predictive. For instance:

- At the lowest level, **next-embedding predictors** (e.g., latent diffusion models) can handle stochasticity and fine-grained variation in raw, continuous perceptual data (e.g., pixels, audio, proprioception). These models incorporate generative mechanisms (e.g., encoder-decoder architecture) that directly ground predictions in observable data, which leads to stronger supervision as we show in §4.4.
- At the intermediate level, a **next-token predictor** (e.g., autoregressive Transformer decoder) can reason over discrete modality tokens derived via VQ-VAE-style encoders, capturing the symbolic and compositional structure.
- At the highest level, an **large language model (LLM)** operating in a “thought space” composed of language tokens can support long-horizon planning, mental simulation, and counterfactual reasoning. Together with the intermediate level, these two levels of discrete reasoning can be jointly implemented through an enhanced LLM architecture performing next-token prediction.

The GLP paradigm not only supports structured, abstract reasoning through next-latent prediction, but also preserves the capacity for detailed reconstruction of the input world, enabling generative supervision and external use. This not only mitigates the compounding of prediction errors by isolating low-level variability within the bottom encoder-decoder layer, but also enables more expressive reasoning and generalization at higher layers of abstraction. Importantly, it allows the model to flexibly mix continuous embeddings for perceptual nuance with discrete tokens for abstract structure, which aligns with our discussion of representation in §4.2. As we further elaborate in §4.4, this encoder-world-model-decoder design leads to stronger supervision and more stable training dynamics than encoder-only approaches like JEPA.

4.4 Objective: Learning in Data Space, or Latent Space?

Abandon probabilistic, data-reconstruction objectives; adopt energy-based, latent-reconstruction objectives for tractability.

A key claim underlying the JEPA framework is that reconstructing raw observations (e.g., pixels in a video) is unnecessary, and that learning in latent space is more effective. This has given rise to a preference for **latent reconstruction objectives**, which bypass the decoder and directly supervise transitions between encoded states. Formally, given encoder h and world model f , the latent reconstruction loss is defined as:

$$\mathcal{L}_{\text{latent}}(h, f) = \mathbb{E}_{(o, a, o') \sim \mathcal{D}} [\|f(h(o), a) - h(o')\|], \quad (7)$$

where the model predicts the next latent state \hat{s}' and compares it to the encoded form of the next observation, without reconstruction o' itself.

Despite its apparent simplicity, this objective is prone to collapse, as we show in Proposition 1: the model can trivially minimize the loss by mapping all observations to a constant vector and learning an invariant transition. To counteract this tendency, JEPA-style systems often require complex regularizers (e.g., maximizing the information $I(\hat{s})$ of latent states). These regularizers, however, are often hard to tune and difficult to understand, which can make training brittle and limits scalability. By contrast, the **generative reconstruction loss** grounds the learning objective in observable data by introducing a decoder g , and supervising the predicted next observation directly as below:

$$\mathcal{L}_{\text{gen}}(h, f, g) = \mathbb{E}_{(o, a, o') \sim \mathcal{D}} [\|g \circ f(h(o), a) - o'\|]. \quad (8)$$

Indeed, the generative loss \mathcal{L}_{gen} anchors the learned representation to the structure of the sensory world, and thus avoids the collapse suffered by the latent loss $\mathcal{L}_{\text{latent}}$, as we show in Proposition 2.

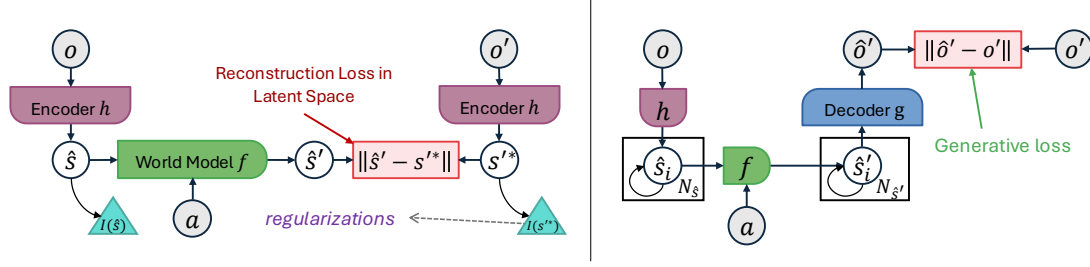


Figure 7: Comparison of the critiqued world model objective based on reconstruction in latent space (left) and our proposed alternative of generative data reconstruction objectives (right).

Proposition 1 (Collapse of Latent Reconstruction Loss). *Given \mathcal{O} , \mathcal{S} , $\mathcal{A} \subseteq \mathbb{R}^d$ and functions $h : \mathcal{O} \rightarrow \mathcal{S}$, $f : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$, and latent reconstruction loss:*

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

There exists (h^, f^*) and $c \in \mathcal{S}$, such that $h^*(o) = c$ for all $o \in \mathcal{O}$ and $f^*(c, a) = c$ for all $a \in \mathcal{A}$, such that:*

$$\mathcal{L}_{\text{latent}}(h^*, f^*) = \min_{h, f} \mathcal{L}_{\text{latent}}(h, f)$$

Explanation. *If you have an encoder and a world model, and you train it using the latent reconstruction loss, there is a cheat configuration for the model to minimize the loss while learning nothing about the true dynamics.*

Proof Sketch. If we construct such a degenerate solution (h^*, f^*) , this solution satisfies $\mathcal{L}_{\text{latent}}(h^*, f^*) = 0$, which is a global minimum as $\mathcal{L}_{\text{latent}}(h, f) \geq 0$ for all (h, f) . \square

Proposition 2 (Non-Collapse of Generative Loss). *Given functions $h : \mathcal{O} \rightarrow \mathcal{S}$, $f : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$, $g : \mathcal{S} \rightarrow \mathcal{O}$, and generative loss:*

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

Assuming $\exists (o_1, a_1, o_2), (o_3, a_3, o_4) \in \mathcal{D}$ such that $o_2 \neq o_4$, then given (h', f') , fixed g' and $c \in \mathcal{S}$ such that $h'(o) = c \forall o \in \mathcal{O}$ and $f'(c, a) = c \forall a \in \mathcal{A}$, there exists (\tilde{h}, \tilde{f}) such that:

$$\mathcal{L}_{\text{gen}}(\tilde{h}, \tilde{f}, g') < \mathcal{L}_{\text{gen}}(h', f', g')$$

Explanation. *If you add a decoder to the model and train it with the generative loss, assuming the data contains different next-observation targets, there will always be another set of encoder and world model that gets lower loss than the previous cheat configuration.*

Proof Sketch. Given degenerate solution (h', f') and fixed g' , construct (\tilde{h}, \tilde{f}) to be equal to (h', f') at every point except (o_1, a_1, o_2) and (o_3, a_3, o_4) where $o_2 \neq o_4$ and the constant-valued (h', f') will get non-zero loss. Instead, set (\tilde{h}, \tilde{f}) to perfectly fit these two targets, so they will get zero loss. Thus we have:

$$\mathcal{L}_{\text{gen}}(\tilde{h}, \tilde{f}, g') < \mathcal{L}_{\text{gen}}(h', f', g')$$

\square

Details of the proofs of both propositions are available in Appendices B and C, respectively.

Beyond the issue of collapsing, a more fundamental structural limitation of the latent reconstruction objective is that it essentially acts as a loose surrogate for observation-level consistency, as we show in Theorem 2. This means that minimizing $\mathcal{L}_{\text{latent}}$ does not, in general, guarantee consistency with what the agent would observe in the world, which can lead to misaligned or brittle representations. In general-purpose settings, we argue that anchoring to the next observation o' via generative loss provides a more stable and mechanistically interpretable training signal.

Theorem 2 (Latent reconstruction is an upper-bounded surrogate for generative reconstruction). *Given sufficiently powerful encoder $h : \mathcal{O} \rightarrow \mathcal{S}$ and decoder $g : \mathcal{S} \rightarrow \mathcal{O}$, such that for all latent states $\hat{s} \in \mathcal{S}$, the roundtrip reconstruction error satisfies $\|h \circ g(\hat{s}) - \hat{s}\| \leq \epsilon$ for some small $\epsilon > 0$. For world model $f : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ and transition data $(o, a, o') \sim \mathcal{D}$, define the latent-space loss $\mathcal{L}_{\text{latent}}$ and the generative loss \mathcal{L}_{gen} as below:*

$$\mathcal{L}_{\text{latent}} = \|f(h(o), a) - h(o')\|, \quad \mathcal{L}_{\text{gen}} = \|g \circ f(h(o), a) - o'\|.$$

Assume that encoder h , decoder g , and world model f induce the following conditional distributions:

$$\hat{s} \mid o \sim \mathcal{N}(h(o), I), \quad \hat{o} \mid \hat{s} \sim \mathcal{N}(g(\hat{s}), I), \quad \text{and} \quad \hat{s}' \mid \hat{s}, a \sim \mathcal{N}(f(\hat{s}, a), I).$$

Then, the following inequality holds:

$$\mathcal{L}_{\text{latent}} \leq \mathcal{L}_{\text{gen}} + \epsilon,$$

with $\mathcal{L}_{\text{latent}} = \mathcal{L}_{\text{gen}}$ when $h \circ g(\hat{s}) = \hat{s}$ for all $\hat{s} \in \mathcal{S}$ and $\text{supp}(\mathcal{D}) \subseteq \text{Im}(g)$.

Explanation. *If your encoder and decoder approximately undo each other, then the JEPA latent reconstruction loss is upper-bounded by the generative loss plus a small reconstruction error. These losses are only equal when the encoder-decoder pair perfectly invert each other, which is unrealistic in practice. As such, minimizing the latent loss does not guarantee consistency with observed data, which is required for minimizing the generative loss (Figure 8).*

Proof Sketch. Observe that the two losses $\mathcal{L}_{\text{latent}}$ and \mathcal{L}_{gen} are scaled KL divergences of the Gaussian prediction distributions in the latent and observation spaces, respectively. Apply the encoder h to both the observation prediction and true observation distributions in \mathcal{L}_{gen} , and the **data processing inequality** states that the augmented KL divergence is upper-bounded by the original \mathcal{L}_{gen} . After that, apply the triangle inequality to show that $\mathcal{L}_{\text{latent}}$ is upper-bounded by the sum of this augmented KL (upper-bounded by \mathcal{L}_{gen}) and the roundtrip reconstruction error (upper-bounded by ϵ), thus completing the proof. (We include the detailed proof in Appendix D.) \square

In practice, ϵ is small (as is typical for strong modern autoencoders), so $\mathcal{L}_{\text{latent}} \leq \mathcal{L}_{\text{gen}}$ usually holds, meaning that the former can miss semantically important mistakes that the latter will penalize. Additionally, the use of less-understood regularizers in conjunction with $\mathcal{L}_{\text{latent}}$ as the objective makes its outcome even more difficult to assess without necessary boundary conditions imposed by observation data.

In conclusion, our argument is not that world models must operate in pixel space, but that they should learn from it. Framing the distinction as *next-representation prediction versus next-observation prediction* creates a false dichotomy that can lead to theoretical ambiguities and practical instability. The purpose of predicting the next observation is to ensure that the predicted latent representations are meaningfully grounded in the real world, whether conceptually or physically. Conversely, reliable prediction in latent space depends on continual validation through observable data. Mathematically, any latent representation of real-world signal intrinsically suffers from issues of identifiability and stability. As such, alignment and calibration with real data are essential to ensure the representations remain meaningful and robust. Generative reconstruction objectives

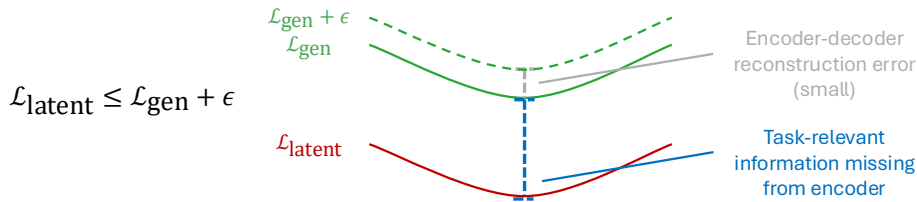


Figure 8: As Thm. 2 shows, the JEPA *latent* reconstruction loss ($\mathcal{L}_{\text{latent}}$) is upper bounded by the generative *data* reconstruction loss (\mathcal{L}_{gen}) plus a small encoder-decoder reconstruction error (ϵ). ϵ is small in practice, meaning $\mathcal{L}_{\text{latent}} \leq \mathcal{L}_{\text{gen}}$ usually holds. Minimizing $\mathcal{L}_{\text{latent}}$, therefore, does not guarantee consistency with observed data, which is required for minimizing \mathcal{L}_{gen} .

tether the learned representations to the observable world, providing richer and more stable learning signal that supports meaningful distinctions, general usability, and human interpretability. These properties are critical for downstream usage, whether it be planning trajectories or training agents through reinforcement learning, which we discuss more in §4.5 below.

4.5 Usage: MPC or RL?

Abandon reinforcement learning (RL); adopt model-predictive control (MPC) for fewer trials required during training.

Beyond training world models, there has also been debate over whether model-predictive control (MPC) is favored over reinforcement learning (RL) in using the WM for reasoning, due to sample efficiency and safety advantages [13]. Here we describe a typical MPC setup (Figure 9, left) which is often adopted by recent work [35, 3]: at timestep t during inference, the agent infers its current latent state $\hat{s}_t = h(o_t)$, proposes an initial sequence of actions (a_t, \dots, a_{T-1}) until some decision horizon T , and uses the world model to predict the corresponding next-state sequence $(\hat{s}_{t+1}, \dots, \hat{s}_T)$. These simulated states can then be evaluated using a cost function $C(g, \hat{s})$ for goal g (e.g., L2 distance between \hat{s} and encoded goal $\hat{s}_g = h(g)$), based on which the agent may propose the next action with lower cost. Decision-making thus amounts to finding the action sequence that minimizes the cost function, formalized as below:

$$(a_t^*, \dots, a_{T-1}^*) = \arg \min_{a_t, \dots, a_{T-1}} \sum_{k=t}^{T-1} C(g, f(\hat{s}_k, a_k)).$$

In practice, the (continuous) action optimization is often performed using traditional numerical algorithms (e.g., MPPI [46] and CEM [31]) involving decision horizons of 1-20 steps and 100s upon 1000s of action samples, and the agent executes the first action in the final action sequence a_t^* before replanning in the next step $t+1$. The appeal of MPC lies in learning from offline trajectories $(o_1, a_1, \dots, o_T) \sim \mathcal{D}$ without potentially unsafe exploration in the real world, as well as potential for higher-quality decision-making from world-model-based simulation.

However, MPC can suffer from practical limitations. Simulation of latent trajectories using the world model, for instance, must be performed repeatedly at every timestep during inference, leading to high computational overhead and making it difficult to respond effectively in fast-changing environments. Beyond computational efficiency, MPC typically plans only a few steps ahead (e.g., up to 10-20 steps) in terms of searching horizon. This limits the extent of its foresight, as long planning horizons (e.g., 100s of steps) can be difficult due to the exploding number of trajectories and world-model errors. As the horizon increases, MPC also becomes more difficult to implement and optimize, since the proposal distribution must sample entire action sequences at once over the full planning

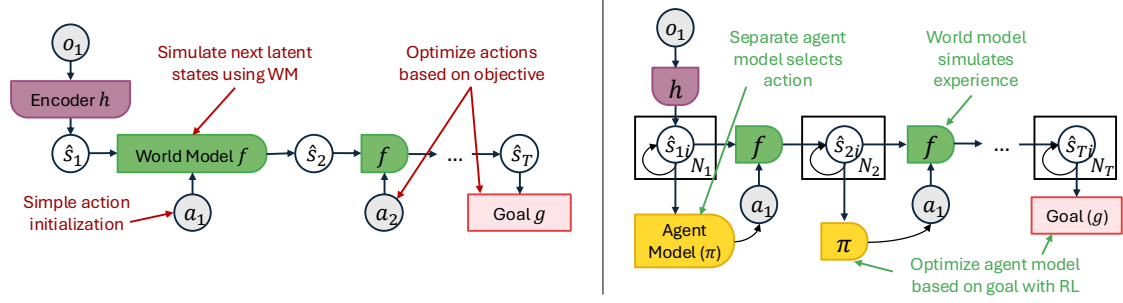


Figure 9: Comparison of the critiqued approach to using world model (left) vs. our proposed alternative (right).

horizon. This is why MPC often relies on relatively simple proposal distributions, such as uniform random sampling or multivariate Gaussians. Indeed, MPC so far has shown promise primarily in simplified settings (e.g., Go) where environment dynamics are simple and slower decision-making is rewarded, but struggles to extend to real-world tasks (e.g., customer service), which typically involve complex dynamics and require a mixture of short- and long-term decision-making.

On the other hand, RL is a general, flexible, and scalable approach to training agents without restrictions on the decision-making method or search horizon. In particular, one can replace the true universe with a world model f for exploration and learning (as discussed in §2.2). Below we describe an RL setup (Figure 9, right) where the agent interacts with the world model instead of the environment [17]: in each time step t with world state representation \hat{s}_t (which may be encoded from some observation data o_t or completely imagined from scratch), the agent π takes action $a_t \sim p_\pi(a_t | \hat{s}_t)$ and the world model f simulates the next state $\hat{s}_{t+1} \sim p_f(\hat{s}_{t+1} | \hat{s}_t, a_t)$. This may repeat until some rollout horizon T or in a never-ending manner. Computing the reward at each step $r(g, \hat{s}_t)$ based on goal g , the optimal agent π_f^* may thus learn by maximizing the expected discounted cumulative reward (with well-formed discount schedule $\{\gamma_k\}_{k=t}^\infty$ to ensure numerical stability) as defined below:

$$\begin{aligned} \pi_f^*(\hat{s}_t) &= \arg \max_{\pi} \mathbb{E}_{\pi, f} \left[\sum_{k=t}^{\infty} \gamma_k r(g, \hat{s}_k) \mid \hat{s}_t \right] \\ &= \arg \max_{\pi} \lim_{T \rightarrow \infty} \sum_{(a_t, \hat{s}_{t+1}, \dots, \hat{s}_T)} \sum_{k=t}^T \gamma_k r(g, \hat{s}_k) \prod_{i=t}^T \underbrace{p_\pi(a_i | \hat{s}_i)}_{\text{select action}} \underbrace{p_f(\hat{s}_{i+1} | \hat{s}_i, a_i)}_{\text{simulation with world model}}. \end{aligned} \quad (9)$$

Operationally, as we show above, both MPC and RL can use world models, the former only for decision-making while the latter also for learning. We recognize the latter as part of a broader paradigm: **learning from experience** [20]. In this framework, the agent model continually interacts with and learns from an infinite space of imagined worlds, simulated by a world model. The countless hypothetical trajectories can then be used to train the agent via RL, imitation learning, or other learning signals that make full use of all experiences. These updates can occur entirely offline, using batches of rollouts from the world model rather than interacting with the real environment.

Compared to MPC which is computationally expensive at decision-making time, RL with world model (as in Equation 9) shifts part of the computational cost to the training phase. Instead of planning from scratch at each step, the world model is used offline to train a policy network that can later be reused for fast action selection at every state. Crucially, Both RL and MPC along with the world model may be included as components inside the agent model that must carry both deliberate planning and reactive actioning, while another fast policy can still learn to react

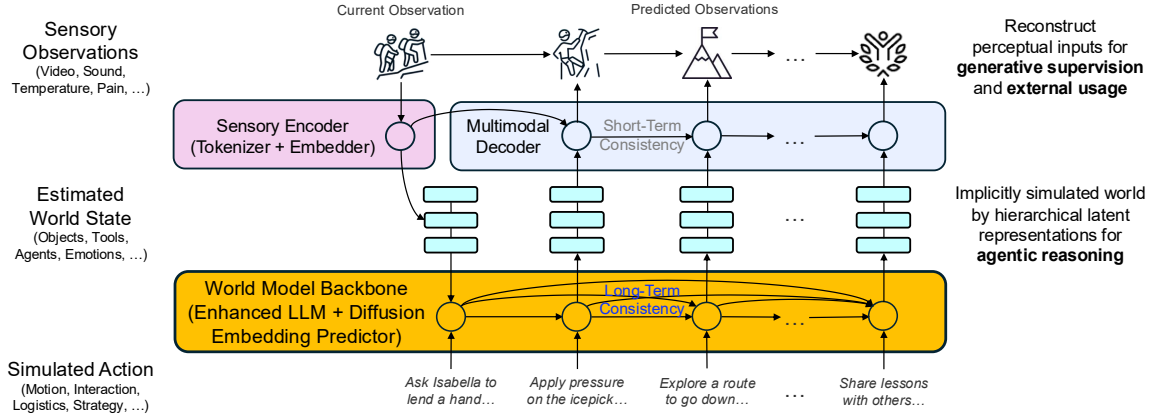


Figure 10: Illustration of the architecture of PAN-World, our proposed framework for general-purpose world modeling. At each time point, PAN-World estimates the world state \hat{s} based on the current sensory observation o , predicts the future world states \hat{s}' based on proposed actions a for agentic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

swiftly when needed. Whereas recent work like o1, o3, and R1 [16] can be seen as special cases in math and coding where model-free policy-based methods enables fast-reacting behaviors, our view is to generalize this pattern: Agents should both reason with and learn from the worlds that they simulate, allowing flexible decision-making, continual improvement, and emergence of intelligence with experience.

In summary, as demonstrated above, unlike MPC, RL can learn a policy function that reflects long-term cumulative rewards, enabling more strategic reasoning over extended time horizons. This makes it applicable in practical settings like goal-conditioned robotic manipulation, multi-turn dialog systems, or autonomous driving.

5 The PAN World Model

Drawing from the critiques made on existing world model frameworks, we arrive at the following conclusion regarding design principles for a general-purpose WM: 1) use data from all modalities of experience; 2) employ a mixed continuous and discrete representation; 3) adopt a hierarchical generative model paradigm with an extended-LLM backbone (for discrete concept-based reasoning), as well as a generative embedding predictive module (for continuous gradient-based reasoning), as the reasoning engines; 4) train over a generative loss grounded in observation data; and 5) apply world model to simulate experiences for training agents using reinforcement learning. We conclude our critiques of WM with a brief preview of an new architecture, PAN – a Physical, Agentic, and Nested world model, based on the aforementioned designed principles. Details and preliminary results of PAN will be presented in [19].

5.1 A Motivating Usecase

A truly versatile and generalizable WM must be grounded in tasks that reflect the full complexity of real-world reasoning demands. These may include variations in data modality (e.g., verbal, visual, sensory), spatio-temporal scope (from one second in a room to days in a whole country), action granularity (e.g., fine motor control, bodily movement, expressive gestures), and decision scale (from

immediate actions, tactics, to long-term strategies). While many existing WMs are demonstrated on simplified, toy tasks (e.g., manipulating kitchen tools) and simple scenarios (seconds to minutes of video in 3D worlds), such settings fall short of capturing the richness of real-world agentic experience. A WM designed around these tasks, therefore, is unlikely to scale to complexities required for real-world applications. For instance, a WM that only enables tool manipulation in a kitchen is inadequate for planning and executing an end-to-end dinner service in a restaurant.

In contrast, PAN is motivated by a more complex and realistic use case: a mountaineering expedition. In this setting, the WM must internalize multimodal sensory inputs and simulate future world states in service of a demanding, structured task. This task naturally decomposes into multiple interconnected sub-tasks at different levels: high-level decisions like gear selection, route and segment planning, navigation, weather assessment, pacing, etc., low-level actions like climbing, roping, and precise motor control in response to terrain and surface conditions, and social coordination with teammates through verbal and non-verbal communication, among others.

The mountaineer’s sensory experience includes not only sights and sounds – snowfields, cliffs, a partner calling out from ahead – but also tactile and motion signals such as wind, cold, and muscular strain. The actionable world states that drive purposeful reasoning, such as terrain affordances, team dynamics, or latent risks, exist at multiple levels of abstraction beneath them. PAN thus begins by ingesting this continuous stream of multimodal signals: inputs from vision, sound, temperature, motion, potentially even pain, which may each be relevant for different tasks but together constitute a holistic reality.

5.2 The PAN Architecture

Following a mixed representation and multi-scale reasoning principle, PAN processes multimodal sensory inputs o using its Sensory Encoder (h), which maps inputs via both discrete and continuous pathways to capture complementary aspects of the world. On one hand (Figure 10), a tokenizer hierarchically maps raw signals into discrete tokens grounded in PAN’s vocabulary, which spans multiple levels of abstraction. These tokens may consist of abstract tokens learned via a VQ-VAE-style approach [39] as well as concrete words drawn from natural language. The representation may include a flexible number of such tokens to compactly reflect deep layers of world information: *Where am I? Who is with me? What tools do I have? What is my emotional state?* As discussed in §4.2, this form of representation can be sufficient to capture relevant information, even for continuous data like video. On the other hand (Figure 6, right), PAN may also encode low-level details into continuous latent embeddings to capture the full nuanced perceptual experience where necessary. Together, these tokens and embeddings form a layered estimate of the world state $\hat{s} = \{\hat{s}_i\}_{i=1}^N$ over which PAN performs simulation and purposeful reasoning.

Given a proposed action a (e.g., “buckle the carabiner to my harness”), PAN predicts the next world state \hat{s}' (e.g., a conceptual state like “I am safely anchored”, or a physical state like “that rope is tightening”) using a World Model Backbone (f) built on an enhanced LLM and a diffusion-based next-latent-embedding predictor (Figure 10). This design is a concrete instantiation of the GLP architecture introduced earlier in §4.3 (Figure 6, right). The LLM-based backbone reasons over both natural language tokens and a learned conceptual vocabulary – some explicit (e.g., a particular shape of icepicks), others implicit or emergent (e.g., feelings that arise when sharing hard-earned knowledge). This supports broad generalization across domains [12]. During both training and inference, the model can also dynamically extend its vocabulary by introducing new tokens or merging existing ones to maximize the prediction quality.

On the other hand, the diffusion-based embedding predictor is responsible for fast, low-level, and often subconscious reasoning that are critical for embodied responses yet difficult to express in language. This module simulates detailed perceptual experiences, such as whether a foothold is secured, or how the body might shift its weight during a climb [48]. A Learned Switch allows PAN to predict the next world state hierarchically ($\{\hat{s}'_i\}_{i=1}^{N'}$) by adaptively combining the LLM-based

backbone, multiple vocabularies, and the diffusion-based embedding predictor, depending on task demands. These mechanisms enable PAN-WM to scale across spatio-temporal scopes and action granularities as is required for general usability – from concrete physical scenarios like mountain climbing and social interaction, to abstract, far-reaching strategic consequences like nationwide policy changes.

To supervise its predictions, and to allow the trained WM to interface with external agents (or human) who may use its outputs, PAN reconstructs the next observation \hat{o}' using a Multimodal Decoder (g) and compares it to the actual observation o' . Crucially, the decoder’s outputs are not limited to videos, but includes a full sensory experience, which may include sound, temperature, motion, pain, other embodied signals, and/or even text. As discussed in §4.3 and §4.4, this generative supervision grounds the predicted world state \hat{s}' in sensory reality, ensuring that the representation retains all possible information while allowing residual variability to be absorbed by the decoder g . This approach contrasts sharply with models trained on next-representation prediction (e.g., V-JEPA 2 [3]), which supervise the world model purely in latent space. The latter objectives are, at best, loose surrogates of generative objectives and prone to representation collapse or unidentifiability, as they lack grounding in real sensory input. Formally, PAN models the conditional distribution of the next observation o' given the current observation o and proposed action a as below:

$$\begin{aligned}
& p_{\text{PAN}}(o' \mid o, a) \\
&= \sum_{\hat{s}, \hat{s}'} \underbrace{p_h(\hat{s} \mid o)}_{\text{encoder}} \underbrace{p_f(\hat{s}' \mid \hat{s}, a)}_{\text{world model}} \underbrace{p_g(o' \mid \hat{s}')}_{\text{decoder}} \\
&= \sum_{\hat{s}, \hat{s}'} \prod_{i=1}^N \underbrace{p_h(\hat{s}_i \mid \hat{s}_{<i}, o)}_{\text{hierarchical world state inference}} \prod_{j=1}^{N'} \underbrace{p_f(\hat{s}'_j \mid \hat{s}'_{<j}, \hat{s}, a)}_{\text{switch-based next-state prediction}} p_g(o' \mid \hat{s}')
\end{aligned}$$

Overall, with its hierarchical, multi-level, and mixed representation architecture, and an encoder-decoder pipeline that threads perception o , action o , belief \hat{s}_i , simulated belief \hat{s}'_i , and simulated worlds o' , PAN represents a general-purpose generative model for simulating actionable real-world possibilities for an agent to perform purposeful reasoning, as we will briefly elude to in §5.4. PAN does not sidestep the variability in raw perceptual input, but instead modularizes and organizes it. This enables richer internal simulation of every layer of experience for more powerful agent reasoning and planning.

5.3 Training the PAN World Model

It should be obvious from the mountaineering example that simply watching videos is not enough to learn all the reasoning capability needed to accomplish the final goal, which can take days of time and hundreds upon thousands of actions and steps from the onset, and is built on rich background knowledge about geography, climate, equipments, sports, and even history. The training of PAN-WM shall use a divide-and-conquer approach that begins with pretraining each of its modules independently through self-supervision (e.g., LLM for text data, and diffusion model for video data). These modality- and level-specific modules are then aligned or integrated during the post-training phase using multimodal data, cascaded embedding, and gradient propagation. Modules operating over continuous embeddings can be trained using standard gradient-based optimization techniques. In contrast, components using discrete tokens may benefit from gradient-free methods similar to reinforcement learning [16]. As proven in §4.4, generative, data-reconstruction-based objectives are grounded in observed data and provide a stable and reliable learning signal for the entire system.

A key strength of the PAN architecture is its data efficiency, because of its use of a multi-scale and stratified view of the world. In the mountaineering task, when reasoning about navigation and

path-finding, the world states do not need to include snow or rock surface details at pixel level, whereas when deciding where to lay hands or feet during claiming, the world states can ignore geographical contexts. Therefore it is not necessary for WMs simulating highly complex possibilities to be contingent on data that capture all such complexity all at once (e.g., videos that visually cover mountaineering at all levels), but to take advantage of data of different kinds offering information at different levels (e.g., travel book for trail guide and map reading, indoor video for rock climbing and gear usage). After all, it is unrealistic to expect a large corpus of videos that comprehensively cover all aspects of alpine climbing. Many general capabilities (e.g., social reasoning, travel planning, cold weather survival) can be learned from abundant language data. Only directly embodied skills (e.g., foot placement, rock climbing technique) require physical data like videos or proprioception, which can be obtained in controlled or simulated environments.

Indeed, PAN’s pretrain-then-align/integrate strategy enables sensory information (e.g., from a video diffusion model) to be grounded within higher-level, richer contexts through LLMs, thereby facilitating cross-modal generalization. At the same time, abstract knowledge embedded in LLMs can be anchored to concrete, embodied experiences, increasing the precision and realism of the system’s reasoning [49]. The result is a WM that, like humans, derives commonsense understanding from a diverse set of experiences. Consequently, it does not require exhaustive training data for each specific task (e.g., mountaineering or autonomous driving), but can instead draw on conceptual knowledge acquired from many domains. We believe this kind of general-purpose WM is well-suited to simulating experience for agent decision-making and/or training, as elaborated below.

5.4 Towards Agentic Reasoning with PAN

Recall in §2 we outlined an agent architecture for simulative reasoning using the world model. PAN fits naturally into this paradigm, functioning not merely as a video generator, but as a rich internal sandbox for simulation, experimentation, and foresight.

As illustrated in Figure 11, a PAN-Agent, prompted by a goal and in reception of continuous stream of perception from the real world, is expected to come up with actions, plans (sequence of actions), or strategies (plans in light of counterfactual situations), which would involve using the PAN-WM to precompute and cache a diverse set of possible world states, plausible actions within those states, and their simulated outcomes [9]. At decision time, rather than only relying on performing expensive real-time simulations, the agent may consult this cache and select actions based on current beliefs and expected rewards. This decoupling of simulation from action selection allows the agent to reason more deliberately, adaptively, and selectively, avoiding the rigidity of purely reactive policy in end-to-end RL and the computational burden of constant forward rollout in MPC. The result is an agent that more closely mirrors human-like cognition – planning ahead, navigating uncertainty, and acting with both flexibility and foresight. We believe this represents a promising step towards agents with richer forms of agency – one capable not only of simulative reasoning, but also of choosing among imagined futures with intent. Such agents may ultimately approach the adaptability, resilience, and autonomy characteristic of human intelligence.

6 Conclusion

We have examined the foundations, debates, and practical challenges in the pursuit of general-purpose world modeling. Our intent in writing this critique is to inspire further discussions and deeper reflections on the following fundamental questions: *what is indeed a world model?*, *what is a world model for*, and *how to build a world model of practical and general utility?*

We argue that a WM is not about video or virtual reality generation, but is about simulating all possibilities in real world; and such outcome is not for visual pleasure, but for purposeful reasoning; and current paradigms and efforts towards this end remain primitive.

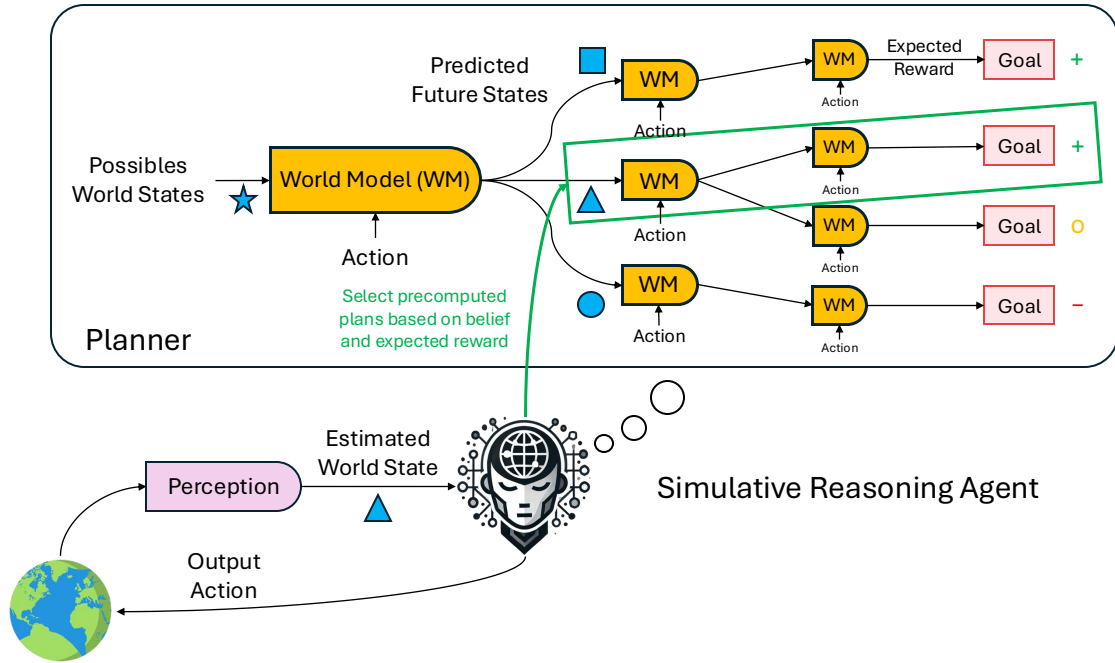


Figure 11: Illustration of our proposed simulative reasoning agent powered by the PAN World Model. Unlike traditional RL agents that rely on reactive policies, or model-precitive control (MPC) agents that expensively simulates futures at decision time, this agent leverages a cache of precomputed simulations 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.

It is our wish that, by offering critical, but analytical and constructive dissections of some of the most popular thoughts on how to build World Models; and by presenting our alternative proposal – the PAN architecture, we can spark further advancements in both theory and implementation of stronger world models.

The PAN world model we previewed is a framework designed for simulating all possible worlds and enabling agent reasoning and planning. By combining multimodal data, hierarchical representations, multi-level generative modeling, and observation-grounded objectives, PAN supports long-horizon simulation and flexible decision-making across tasks in the physical and cyber domains.

Looking ahead, the PAN framework opens several promising directions: scaling from single-agent to multi-agent simulations (e.g., collective behaviors of a business, a society, consequences to public health), extending across time scales (e.g., from milliseconds to millennia), improving simulation fidelity across modalities, and enabling agent learning directly through imagined experience. As world models increasingly serve as the substrate for reasoning, imagination, and action, we believe that frameworks like PAN, with its experience grounding, multi-layer abstraction, and empirical scalability, offers a compelling foundation for the development of robust, general-purpose AI.

References

- [1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.
- [2] Figure AI. Helix: A vision-language-action model for generalist humanoid control, February 2025. Accessed: 2025-05-01.
- [3] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zhohlov, 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.
- [4] Linden J. Ball. *Hypothetical Thinking*, page 514–528. Cambridge Handbooks in Psychology. Cambridge University Press, 2020.
- [5] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.
- [6] Lisa Feldman Barrett. *How emotions are made: The secret life of the brain*. Pan Macmillan, 2017.
- [7] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. 2024. URL <https://openai.com/research/video-generation-models-as-world-simulators>, 3:1, 2024.
- [8] Yuri Chervonyi, Trieu H Trinh, Miroslav Olšák, Xiaomeng Yang, Hoang Nguyen, Marcelo Mene-gali, Junehyuk Jung, Vikas Verma, Quoc V Le, and Thang Luong. Gold-medalist performance in solving olympiad geometry with alphageometry2. *arXiv preprint arXiv:2502.03544*, 2025.
- [9] Brandon Chiou, Mason Choey, Mingkai Deng, Jinyu Hou, Jackie Wang, Ariel Wu, Frank Xu, Zhiting Hu, Hongxia Jin, Li Erran Li, Graham Neubig, Yilin Shen, and Eric P. Xing. Reasoneragent: A fully open source, ready-to-run agent that does research in a web browser and answers your queries, February 2025.
- [10] Terrence W Deacon. *The symbolic species: The co-evolution of language and the brain*. WW Norton & Company, 1998.
- [11] Decart, Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. 2024.
- [12] Mingkai Deng, Jinyu Hou, Zhiting Hu, Graham Neubig, Hongxia Jin, Yilin Shen, and Eric P. Xing. Simura: Towards general goal-oriented agent via simulative reasoning architecture with llm-based world model, 2 2025.
- [13] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion, 2017.
- [14] Rafael C Gonzales and Paul Wintz. *Digital image processing*. Addison-Wesley Longman Publishing Co., Inc., 1987.
- [15] Google DeepMind. Veo. <https://deepmind.google/models/veo/>, 2025. Latest release version: Veo 3 (May 20, 2025), accessed: 2025-06-10.

- [16] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [17] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.
- [18] Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.
- [19] Zhiting Hu and Eric Xing. Pan: A physical, agentic, and nested world model for general and super intelligence. Manuscript in preparation, 2025.
- [20] Zhiting Hu and Eric P Xing. Toward a standard model of machine learning. *arXiv preprint arXiv:2108.07783*, 2021.
- [21] Philip N Johnson-Laird. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250, 2010.
- [22] Anssi Kanervisto, Dave Bignell, Linda Yilin Wen, Martin Grayson, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Tabish Rashid, Tim Pearce, Yuhao Cao, et al. World and human action models towards gameplay ideation. *Nature*, 638(8051):656–663, 2025.
- [23] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.
- [24] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.
- [25] Willis E. McNelly, editor. *The Dune Encyclopedia*. Berkley Publishing Group, New York, 1984. Published by arrangement with the author. Designed by Jeremiah B. Lighter.
- [26] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431, 2024.
- [27] Donald S Ornstein and Benjamin Weiss. Statistical properties of chaotic systems. *Bulletin of the American Mathematical Society*, 24(1):11–116, 1991.
- [28] Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 2: A large-scale foundation world model. 2024.
- [29] Lawrence Perko. *Differential equations and dynamical systems*, volume 7. Springer Science & Business Media, 2013.
- [30] Steven Pinker. The language instinct (1994/2007). 2007.
- [31] Reuven Y. Rubinstein. Optimization of computer simulation models with rare events. *European Journal of Operational Research*, 99(1):89–112, 1997.
- [32] Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025.

- [33] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.
- [34] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dhharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.
- [35] Vlad Sobal, Wancong Zhang, Kynghyun Cho, Randall Balestriero, Tim GJ Rudner, and Yann LeCun. Learning from reward-free offline data: A case for planning with latent dynamics models. *arXiv preprint arXiv:2502.14819*, 2025.
- [36] Ralph E Strauch. Battle simulation for command and control training. 1982.
- [37] Qiao Sun, Xin Huang, Junru Gu, Brian C Williams, and Hang Zhao. M2i: From factored marginal trajectory prediction to interactive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6543–6552, 2022.
- [38] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.
- [39] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
- [40] Nicole Van Hoeck, Patrick D Watson, and Aron K Barbey. Cognitive neuroscience of human counterfactual reasoning. *Frontiers in human neuroscience*, 9:420, 2015.
- [41] Balakrishnan Varadarajan, Ahmed Hefny, Avikalp Srivastava, Khaled S Refaat, Nigamaa Nayakanti, Andre Cornman, Kan Chen, Bertrand Douillard, Chi Pang Lam, Dragomir Anguelov, et al. Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 7814–7821. IEEE, 2022.
- [42] Lev S Vygotsky. *Thought and language*, volume 29. MIT press, 2012.
- [43] Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. Openhands: An open platform for ai software developers as generalist agents. In *The Thirteenth International Conference on Learning Representations*, 2024.
- [44] Yao Wang, Jörn Ostermann, and Ya-Qin Zhang. *Video processing and communications*, volume 1. Prentice hall Upper Saddle River, 2002.
- [45] Waymo. Waymo Driver: Self-Driving Car Technology for a Reliable Ride, 2025. Accessed: 2025-05-01.
- [46] Grady Williams, Andrew Aldrich, and Evangelos Theodorou. Model predictive path integral control using covariance variable importance sampling. *arXiv preprint arXiv:1509.01149*, 2015.
- [47] World Labs Technical Staff. Generating worlds. <https://www.worldlabs.ai/blog>, 2025. Accessed: 2025-06-10.
- [48] Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024.

- [49] Jiannan Xiang, Tianhua Tao, Yi Gu, Tianmin Shu, Zirui Wang, Zichao Yang, and Zhiting Hu. Language models meet world models: Embodied experiences enhance language models. *Advances in neural information processing systems*, 36:75392–75412, 2023.
- [50] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

A Proof for Theorem 1

Proof. We will prove the contrapositive that $f_\epsilon(\mathbf{x}) = f_\epsilon(\mathbf{x}') \Rightarrow \|\mathbf{x} - \mathbf{x}'\| \leq \epsilon$.

Based on the conditions, we have it that $\|\mathbf{x} - \mathbf{x}'\| = \sum_{t=1, d=1}^{T,D} (x_{td} - x'_{td})^2 \leq \epsilon = \tilde{\epsilon}^2$, which is satisfied when $|x_{td} - x'_{td}| \leq \frac{\tilde{\epsilon}}{\sqrt{TD}}$.

Because $\|x_t\| \leq K := \frac{\tilde{K}^2}{4}$, we have that $|x_{td}| \leq \frac{\tilde{K}}{2} \forall d \in [1, D]$. Therefore, to satisfy the condition that $\|\mathbf{x} - \mathbf{x}'\| \leq \epsilon$, we need to divide the interval $[-\frac{\tilde{K}}{2}, \frac{\tilde{K}}{2}]$ into subintervals with length at most $\frac{\tilde{\epsilon}}{\sqrt{TD}}$, of which there are $\lceil \sqrt{TD} \tilde{K} \tilde{\epsilon}^{-1} \rceil$ of them. Further, because there are TD such intervals to divide, the support of our code should be at least as large as the number of required subintervals as below:

$$|\mathcal{V}|^M \geq \lceil \sqrt{TD} \tilde{K} \tilde{\epsilon}^{-1} \rceil^{TD}$$

In the two cases below, we will attempt to construct this support by scaling the vocabulary size $|\mathcal{V}|$ and the code length M individually. It is technically also possible to scale both at the same time, but we focus on the existence for this proof, and leave more efficient solutions to future work.

Case 1 (Modality Tokenizer): Take $N = T$, and construct a vocabulary \mathcal{V} with size $M_\epsilon = \lceil \sqrt{TD} \tilde{K} \tilde{\epsilon}^{-1} \rceil^D$. We then define the encoding function for x as $f_\epsilon(x) = (c_1, \dots, c_T)$ where $c_t \in \mathcal{V}$. We define the index of c_t as a D -digit number of base- M_ϵ

$$(d_1(c_t), d_2(c_t), \dots, d_D(c_t)),$$

where $d_n(c_t) = \lfloor \frac{x_{tn} + \tilde{K}/2}{\tilde{\epsilon}/\sqrt{TD}} \rfloor$ is the n -th digit. Note that this is well defined because $d_n(c_t) \in [0, \lceil \sqrt{TD} \tilde{K} \tilde{\epsilon}^{-1} \rceil]$, and thus enables the representation of $\lceil \sqrt{TD} \tilde{K} \tilde{\epsilon}^{-1} \rceil^D$ unique values for c_t using D digits. This way, given $f_\epsilon(x) = f_\epsilon(x')$, since $c_t = c'_t \forall t \in [1, T]$, we have $d_n(c_t) = d_n(c'_t) \forall n \in [1, D]$, which implies that $|x_{tn} - x'_{tn}| \leq \tilde{\epsilon}/\sqrt{TD}$. Thus, we have

$$\|x - x'\| = \sum_{t=1, d=1}^{T,D} (x_{td} - x'_{td})^2 \leq \sum_{t=1, d=1}^{T,D} \frac{\tilde{\epsilon}^2}{TD} = \epsilon \quad (10)$$

Case 2 (Finding a Language Expression): Assume that the vocabulary has a fixed size $|\mathcal{V}| = N$ (which applies to a wide range of natural and machine languages). Take sequence length $N_\epsilon = TD \lceil \log_M \sqrt{TD} \tilde{K} \tilde{\epsilon}^{-1} \rceil$ and define

$$f_\epsilon(\mathbf{x}) = (c_1, \dots, c_{N_\epsilon}), \quad c_t \in \mathcal{V},$$

We map each x_{td} to a number $u_{td} \in [0, 1]$ represented by L digits of base- N , where the n -th digit $d_n(u_{td}) = c_{s+n}$. In this manner, x_{td} will correspond to subsequence $(c_{s+1}, c_{s+2}, \dots, c_{s+L})$, where $s = (td - 1) \lceil \log_M \sqrt{TD} \tilde{K} \tilde{\epsilon}^{-1} \rceil$ and $L = \lceil \log_N \sqrt{TD} \tilde{K} \tilde{\epsilon}^{-1} \rceil$. Thus, we have that

$$x_{td} = \left(\sum_{n=1}^L \frac{d_n(u_{td})}{M^n} - \frac{1}{2} \right) \tilde{K}. \quad (11)$$

Now, given \mathbf{x}, \mathbf{x}' where $f_\epsilon(\mathbf{x}) = f_\epsilon(\mathbf{x}')$, we have $f_\epsilon(\mathbf{x})_{td} = f_\epsilon(\mathbf{x}')_{td}$ and $d_n(u_{td}) = d_n(u'_{td})$, so u_{td} and u'_{td} must lie in the same interval formed by numbers for which all digits d_n are identical as below:

$$[u : d_n(u) = v_n, n = 1, \dots, L] = \left[\sum_{n=1}^L \frac{v_n}{M^n}, \sum_{n=1}^L \frac{v_n}{M^n} + \frac{1}{M^L} \right) \quad (12)$$

Thus we have

$$\begin{aligned} |x_{td} - x'_{td}| &= \left| \left(\sum_{n=1}^L \frac{d_n(u_{td})}{M^n} - \frac{1}{2} \right) \tilde{K} - \left(\sum_{n=1}^L \frac{d_n(u'_{td})}{M^n} - \frac{1}{2} \right) \tilde{K} \right| = \left| \left(\sum_{n=1}^L \frac{d_n(u_{td})}{M^n} - \sum_{n=1}^L \frac{d_n(u'_{td})}{M^n} \right) \tilde{K} \right| \\ &\leq \left| M^{-L} \tilde{K} \right| = \left| M^{-\lceil \log_M \sqrt{TD} \tilde{K} \tilde{\epsilon}^{-1} \rceil} \tilde{K} \right| \leq \left| M^{-\log_M \sqrt{TD} \tilde{K} \tilde{\epsilon}^{-1}} \tilde{K} \right| \leq \left| \frac{\tilde{\epsilon}}{\sqrt{TD}} \right| \end{aligned}$$

And by extension $\|\mathbf{x} - \mathbf{x}'\| \leq \epsilon$ □

B Proof for Proposition 1

Proof. We will prove the proposition by constructing such a (h^*, f^*) pair such that $\mathcal{L}_{\text{latent}}(h^*, f^*) = \min_{h, f} \mathcal{L}_{\text{latent}}(h, f)$. $((h^*, f^*))$ is the global minimum of $\mathcal{L}_{\text{latent}}$.

This could be simply done by finding arbitrary $c \in \mathbb{R}^n$ and letting $h^*(o) = c$ for all $o \in \mathcal{O}$ and $f^*(c, a) = c$ for all $a \in \mathcal{A}$. Then, we have that $\mathcal{L}_{\text{latent}}(h^*, f^*) = 0$. Since $\mathcal{L}_{\text{latent}}$ is a distance measurement, we have that $\mathcal{L}_{\text{latent}} \geq 0$, which means that (h^*, f^*) is the global minimum. □

C Proof for Proposition 2

Proof. We prove the proposition by showing that, with the generative loss \mathcal{L}_{gen} , there is no solution that can be both degenerate and optimal at the same time. More specifically, for any degenerate solution (h', f') , and fixed g , there must be a way to construct another solution (\tilde{h}, \tilde{f}) which yields a lower loss value.

More formally, without loss of generality, assume that we have:

- Some arbitrary constant $c \in \mathbb{R}^n$;
- $h' : \mathcal{O} \rightarrow \mathcal{S}$ such that $h'(o) = c$ for all $o \in \mathcal{O}$;
- $f' : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ such that $f'(c, a) = c$ for all $a \in \mathcal{A}$;
- A fixed decoder $g : \mathcal{S} \rightarrow \mathcal{O}$;
- A dataset $\mathcal{D} \subset \mathcal{O} \times \mathcal{A} \times \mathcal{O}$ containing at least two distinct triplets:

$$(o_1, a_1, o_2), \quad (o_3, a_3, o_4) \in \mathcal{D} \text{ with } o_2 \neq o_4.$$

Assuming all the function families are sufficiently expressive. Then, under the degenerate solution (h', f') , for any $(o, a, o') \in \mathcal{D}$, we have: $g \circ f'(h'(o), a) = g(f'(c, a)) = g(c)$. Therefore, the generative loss becomes:

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

Note that this loss is non-zero if there exists at least one $(o, a, o') \in \mathcal{D}$ such that $g(c) \neq o'$, which must be the case here because $o_2 \neq o_4$ while $g(c)$ is fixed across all inputs.

We now construct an alternative pair (\tilde{h}, \tilde{f}) such that:

- For all $(o, a, o') \in \mathcal{D} \setminus \{(o_1, a_1, o_2), (o_3, a_3, o_4)\}$, define $\tilde{h}(o) = c$ and $\tilde{f}(c, a) = c$, i.e., \tilde{h}, \tilde{f} behave identically to the degenerate solution.
- For the two special cases, define $s_1, s_2 \in \mathcal{S}$ such that $g'(s_1) = o_2$ and $g'(s_2) = o_4$. (This is possible since g' is fixed and assumed expressive enough to approximate o_2 and o_4 .) Then, set $\tilde{h}(o_1) = \hat{s}_1$, $\tilde{f}(\hat{s}_1, a_1) = s_1$, and $\tilde{h}(o_3) = \hat{s}_2$, $\tilde{f}(\hat{s}_2, a_3) = s_2$, where $\hat{s}_1, \hat{s}_2 \in \mathcal{S}$ are any distinct intermediate representations to encode o_1 and o_3 .

Now, consider the loss under this construction:

- At (o_1, a_1, o_2) : $\tilde{h}(o_1) = \hat{s}_1$, $\tilde{f}(\hat{s}_1, a_1) = s_1$, $g'(s_1) = o_2$, so loss is $\|o_2 - o_2\| = 0$.
- At (o_3, a_3, o_4) : similarly, loss is zero for \tilde{h} and \tilde{f} .
- While for h' and f' , there must be one of (o_1, a_1, o_2) and (o_3, a_3, o_4) at which the loss is strictly positive. In other words, $\|g(c) - o_2\| + \|g(c) - o_4\| > 0$
- At all other datapoints, the loss is unchanged (equal to the degenerate solution's loss) because we did not modify the outputs for those inputs.

Hence, the total loss under (\tilde{h}, \tilde{f}) is strictly less than under (h', f') :

$$\mathcal{L}_{\text{gen}}(\tilde{h}, \tilde{f}, g) < \mathcal{L}_{\text{gen}}(h', f', g)$$

This indicates that the degenerate solution must not be optimal, thus proving that it cannot be a global minimizer when g is trained or expressive and when the dataset includes different outputs for different inputs. \square

D Proof of Theorem 2

Proof. Because the conditional distributions in the statement are all isotropic Gaussians, we can use them to construct the following transition distributions (assuming \hat{q} is the empirical data distribution):

$$\begin{aligned} \text{Latent Prediction } p_{h \circ f} : \hat{s}' &| o, a \sim \mathcal{N}(f(h(o), a), I) \\ \text{Latent Target } \hat{q}_h : \hat{s}' &| o' \sim \mathcal{N}(h(o'), I) \\ \text{Observation Prediction } p_{h \circ f \circ g} : \hat{o}' &| o, a \sim \mathcal{N}(g \circ f(h(o), a), I) \end{aligned}$$

It emerges that the two loss functions described above can be expressed as scaled KL divergences of those distributions, as shown below:

$$\begin{aligned} \mathcal{L}_{\text{latent}} &= \|f(h(o), a) - h(o')\| = \sqrt{2D_{\text{KL}}(\hat{q}_h(\hat{s}' | o') \| p_{h \circ f}(\hat{s}' | o, a))}, \\ \mathcal{L}_{\text{gen}} &= \|g \circ f(h(o), a) - o'\| = \sqrt{2D_{\text{KL}}(\hat{q}(\hat{o}') \| p_{h \circ f \circ g}(\hat{o}' | o, a))}. \end{aligned}$$

Consider applying the encoder transition to \hat{q} and $p_{h \circ f \circ g}$. In the former case, we will recover the latent target \hat{q}_h , and in the latter case we will have performed a roundtrip from latent to observation, and back to latent, resulting in the following conditional distribution:

$$\text{Roundtrip Latent Prediction } p_{h \circ f \circ g \circ h} : \hat{s}' | o, a \sim \mathcal{N}(h \circ g \circ f(h(o), a), I).$$

Applying the **data processing inequality** gives the following relation:

$$D_{\text{KL}}(\hat{q}_h(\hat{s}' | o') \| p_{h \circ f \circ g \circ h}(\hat{s}' | o, a)) \leq D_{\text{KL}}(\hat{q}(\hat{o}') \| p_{h \circ f \circ g}(\hat{o}' | o, a)),$$

And because the left-hand side is $\frac{1}{2}\|h \circ g \circ f(h(o), a) - h(o')\|^2$ and the right-hand side is $\frac{1}{2}\mathcal{L}_{\text{gen}}^2$, we plug them into the inequality above and get:

$$\|h \circ g \circ f(h(o), a) - h(o')\| \leq \mathcal{L}_{\text{gen}}. \quad (13)$$

Next, because the encoder h and decoder g satisfy the roundtrip consistency of $\|h \circ g(\hat{s}) - \hat{s}\| \leq \epsilon$, plugging in $\hat{s} = f(h(o), a)$ gives:

$$\|h \circ g \circ f(h(o), a) - f(h(o), a)\| \leq \epsilon. \quad (14)$$

Finally, we apply the triangle inequality to upper-bound the latent loss as below:

$$\begin{aligned} \mathcal{L}_{\text{latent}} &= \|f(h(o), a) - h(o')\| \\ &\leq \underbrace{\|f(h(o), a) - h \circ g \circ f(h(o), a)\|}_{\leq \epsilon \text{ by Inequality 14}} + \underbrace{\|h \circ g \circ f(h(o), a) - h(o')\|}_{\leq \mathcal{L}_{\text{gen}} \text{ by Inequality 13}} \\ &\leq \mathcal{L}_{\text{gen}} + \epsilon \end{aligned}$$

When $h \circ g(\hat{s}) = \hat{s}$ for all $\hat{s} \in \mathcal{S}$, we have that $g : \mathcal{S} \rightarrow \text{Im}(g) \subseteq \mathcal{O}$ is an injective function which maps each point in \mathcal{S} to a unique point in its image $\text{Im}(g)$. On the other hand, h is the left inverse of g on $\text{Im}(g)$, which maps each $\hat{o} \in \text{Im}(g)$ to $\hat{s} \in \mathcal{S}$ s.t. $g(\hat{s}) = \hat{o}$. Thus h is bijective in $\text{Im}(g)$ by mapping each element $o \in \text{Im}(g)$ to exactly one $\hat{s} \in \mathcal{S}$.

This means h defines a parameter transformation T_h where $T_h(o) = h(o)$ for all $o \in \text{Im}(g)$. Because $\hat{q}(o')$ and $p_{h \circ f \circ g}(o' \mid o, a)$ are supported entirely inside $\text{Im}(g)$ and KL divergence is invariant under parameter transformation, we have:

$$\begin{aligned} \frac{1}{2}\mathcal{L}_{\text{gen}}^2 &= D_{\text{KL}}(\hat{q}(o') \parallel p_{h \circ f \circ g}(o' \mid o, a)) \\ &= D_{\text{KL}}(\hat{q}_h(\hat{s}' \mid o') \parallel p_{h \circ f \circ g \circ h}(\hat{s}' \mid o, a)) \\ &= \frac{1}{2}\|h \circ g \circ f(h(o), a) - h(o')\|^2 \\ &= \frac{1}{2}\|f(h(o), a) - h(o')\|^2 \quad (h \circ g(\hat{s}) = \hat{s}) \\ &= \frac{1}{2}\mathcal{L}_{\text{latent}}^2. \end{aligned}$$

Thus we have $\mathcal{L}_{\text{latent}} = \mathcal{L}_{\text{gen}}$. □