Concept Paper

Not peer-reviewed version

Three-World Hierarchy for General Neural-Network-in-the-Loop Stochastic Dynamical System

HongSheng Qi *

Posted Date: 14 August 2025

doi: 10.20944/preprints202508.1078.v1

Keywords: world model; three-world hierarchy; Frechet distance; dynamic time wrapping

Concept Paper

Three-World Hierarchy for General Neural-Network-in-the-Loop Stochastic Dynamical System

HongSheng Qi [1,2]

[1] College of Civil Engineering and Architecture, Zhejiang University. 866 Yuhangtang Road, Hangzhou, 310058, China; qihongsheng@zju.edu.cn

[2] Department of Civil and Environmental Engineering, Pennsylvania State University, 212 Sackett Building, University Park, PA, USA. 16802

Abstract

The remarkable success of neural networks, both in theory and practice, has led to their increasing integration with the physical world, making physically interactive world models a tangible reality. The interaction between neural networks and the physical world is managed through human-crafted algorithms, reinforcement learning, and, more recently, world models. However, the standard neural network workflow-defining the network, training with data, and deploying to the real world-faces significant challenges when dealing with these complex, stochastic dynamical systems in real-world settings. These challenges include hallucination, out-of-distribution issues, and long-tail events. This manuscript proposes a novel hierarchical framework composed of three distinct levels of "worlds" to comprehensively describe general neural-network-in-the-loop stochastic dynamical systems: the data world, model world, and real world. Furthermore, to quantify the divergence between these multi-modal worlds, we introduce a new distance measurement called Fréchet World Distance (FWD). FWD generalizes the conventional Fréchet distance to accommodate dynamic and multi-modal settings, providing a crucial tool for analyzing and optimizing the interaction between neural networks and the physical environment.

Keywords: world model; three-world hierarchy; Frechet distance; dynamic time wrapping

1. Introduction

The remarkable success of large language models (LLMs) like GPT has propelled artificial intelligence (AI) to the forefront of both academia and industry. AI's evolution now extends significantly beyond its initial core task of natural language processing, with advanced neural models impacting diverse fields such as robotics, autonomous driving, and remote sensing.

Increasingly, sophisticated deep learning models are being deployed in the physical world. Historically, interaction between early neural networks and their environments relied on human-crafted algorithms. This evolved with the introduction of the reinforcement learning (RL) paradigm, enabling models to learn interactions directly from their environments. However, as the information processing capabilities of neural models have expanded to encompass multiple modalities (e.g., video, audio, image, LiDAR) beyond just language, RL appears increasingly insufficient for future applications. Since 2025, the concept of "world models" has emerged, playing a pivotal role in the development of the next generation of neural networks (Ha & Schmidhuber, 2018).

The term "world model" is still evolving, and its interpretation remains diverse. Various approaches are being explored to achieve successful interaction between the physical world and neural networks. These include multimodal LLMs (Suzuki & Matsuo, 2022), Joint Embedding Predictive Architectures (JEPA) and their variants (LeCun, 2022; Fei et al., 2023), simulation platforms (Guan et al., 2024), and video world models (Zhen et al., 2024). The very term "world" can sometimes be ambiguous, particularly given the lack of a current consensus on the future evolutionary path of neural networks.

The success of modern neural networks, however, is largely **scale-driven**. In stark contrast to their widespread deployment, the theoretical understanding and analytical tools for investigating neural network properties remain underdeveloped. This often leads to neural networks being described as "**black boxes**." The significant gap between practical application and theoretical comprehension poses numerous potential risks for the development of Artificial General Intelligence (AGI) and Artificial Super Intelligence (ASI). These risks include:

- Hallucination: Models can generate unrealistic or factually inconsistent outputs that diverge from real-world phenomena.
- Misalignment: Models may not perform as intended, necessitating further training or recalibration to align with desired behaviors.
- Security Risks: Models might fail to adequately cope with complex or unforeseen environmental conditions. A common example is the need for human intervention (takeover) in autonomous vehicles when the system encounters challenging scenarios.

The aforementioned risks necessitate a deep understanding of the mutual interaction between models (represented as complex neural networks) and the physical world. We use the term "**neural network in the loop**" to describe this critical combination of models and the physical environment. To clarify this "in-the-loop" concept, this manuscript outlines a three-world hierarchy that distinguishes different levels of the physical environment. We will elaborate on this framework in the next section. Here's a summary of our key definitions:

- The "model world" is defined as the state space of the trained neural network model. This model world's state space is distinct from that of the real world.
- The "data world" is spanned by the data used for training and validation, encompassing both real-world and synthetic data.

2. Three-World Hierarchy

Figure 1 illustrates our proposed three-level world hierarchy, comprising the Empirical World ($\mathbb{L}_1$), the Model World ($\mathbb{L}_2$), and the Data World ($\mathbb{L}_3$). Their definitions and interrelationships are detailed below.

- Empirical World ($\mathbb{L}_1$)

The Empirical World ($\mathbb{L}_1$) represents the uppermost layer of this hierarchy, encompassing all phenomena and states that can be perceived, predicted, or interacted with in reality. It is synonymous with objective reality.

- Model World ($\mathbb{L}_2$)

The Model World ($\mathbb{L}_2$) constitutes the operational space formed by our models, primarily complex neural networks, though it may also integrate theoretically driven models like Newtonian mechanics. Each model is characterized by its specific definition, data structure declarations, and inherent configurations, which collectively determine the scope of its sensory perception and actions. Compared to the Empirical World ($\mathbb{L}_1$), the Model World exhibits two key differences:

    - States Not Covered by the Model ($\mathbb{L}_1 \setminus \mathbb{L}_2$): This set difference represents aspects of the Empirical World that fall outside the model's perceptual or operational capabilities. Several factors contribute to this:
        - Undefined Sensor Modalities: The model's sensors may not be designed to capture certain states. For example, an autonomous vehicle lacking an infrared detector cannot directly sense pavement temperature, even if it could be indirectly estimated.
        - Performance Limitations: The model's capabilities may be insufficient to resolve certain details within its defined state space. For instance, an autonomous vehicle's detectors might have limited resolution, preventing the detection of nanoscale pavement details.

- Out-of-Definition States: The state space is simply beyond the model's fundamental definition. A standard large language model, for example, cannot inherently perceive audio.
- Model-Generated States Not Existing in Reality ($\mathbb{L}_2 \setminus \mathbb{L}_1$): This set difference signifies instances where the model generates outputs that have no corresponding existence in the Empirical World. These outputs are often referred to as hallucinations. Examples include a large language model generating a non-existent reference or a text-to-image model producing physically impossible or "weird" images. Numerous underlying reasons contribute to the generation of these non-existent states.

- Data World ($\mathbb{L}_3$)

The Data World ($\mathbb{L}_3$) is spanned by the datasets used to train deep neural network models. This world comprises both data collected from the real world (Empirical World, $\mathbb{L}_1$) and human-crafted or synthetic data.

As observed, these three distinct worlds share overlapping state spaces while also possessing unique characteristics. To comprehensively represent them, we define $\mathbb{L}_0 = \mathbb{L}_1 \cup \mathbb{L}_2 \cup \mathbb{L}_3$. When considering the stochastic dynamics of a neural-network-in-the-loop system, it's beneficial to introduce the concepts of forward and backward dynamics from the perspective of this three-world hierarchy. Discrepancies inevitably exist among these three worlds. Given the system's stochastic nature, any initial state (e.g., state A in Figure 5-b) can evolve into multiple future states. Once an initial state is defined, the potential system evolution trajectories are encapsulated within a "forward dynamics set," visualized as a cone in Figure 5-b. Conversely, a specific final state (e.g., state D in Figure 5-b) can be reached from numerous historical states. Thus, for a given terminal state D, the possible system trajectories form a "backward dynamics set," also represented as a cone in Figure 5-b. A typical observed data trajectory (such as the path from A to D in Figure 5-b) is effectively bounded by an infinite number of these forward and backward dynamics sets. Furthermore, at any given data point along an observed trajectory (e.g., data point B), the system retains the capacity to evolve to alternative states (e.g., state E in Figure 5-b). This illustrates that the dynamics observed within the data world ($\mathbb{L}_3$) (Figure 5-a) represent only a subset of the broader empirical world ($\mathbb{L}_1$) dynamics.

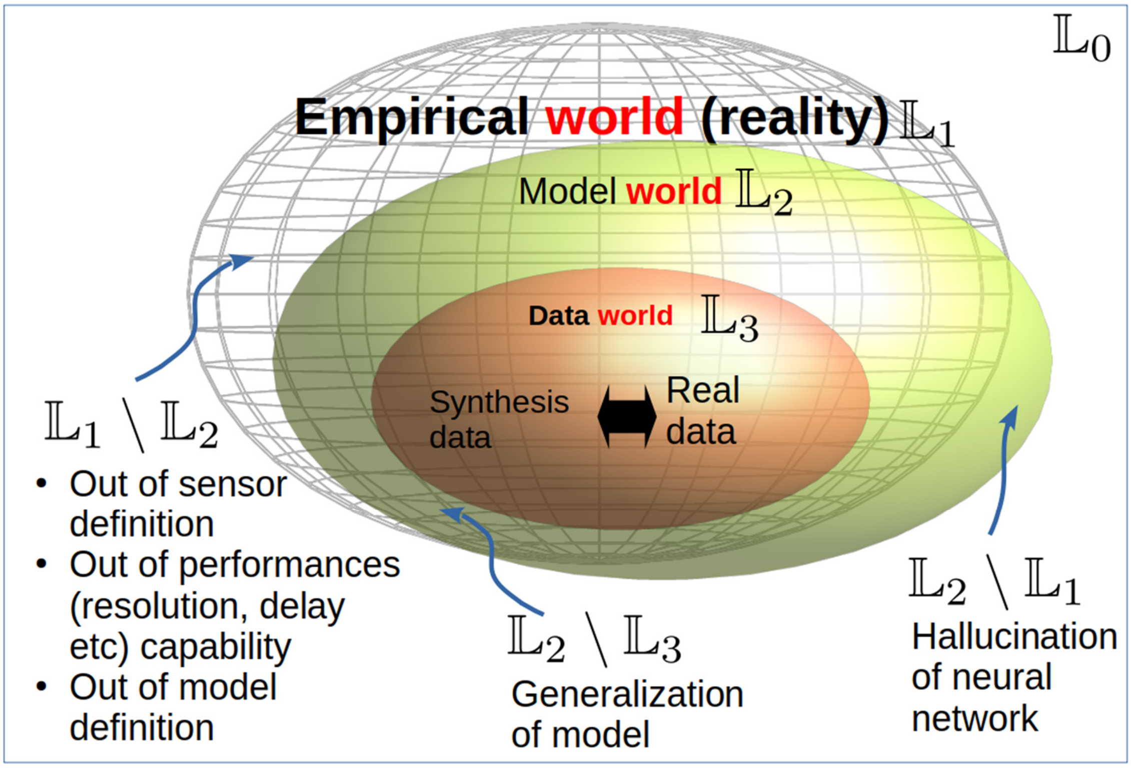

$\mathbb{L}_0$
Empirical world (reality) $\mathbb{L}_1$
Model world $\mathbb{L}_2$
Data world $\mathbb{L}_3$
Synthesis data
Real data
$\mathbb{L}_1 \setminus \mathbb{L}_2$
• Out of sensor definition
• Out of performances (resolution, delay etc) capability
• Out of model definition
$\mathbb{L}_2 \setminus \mathbb{L}_3$
Generalization of model
$\mathbb{L}_2 \setminus \mathbb{L}_1$
Hallucination of neural network

Figure 1. Three-world hierarchy.

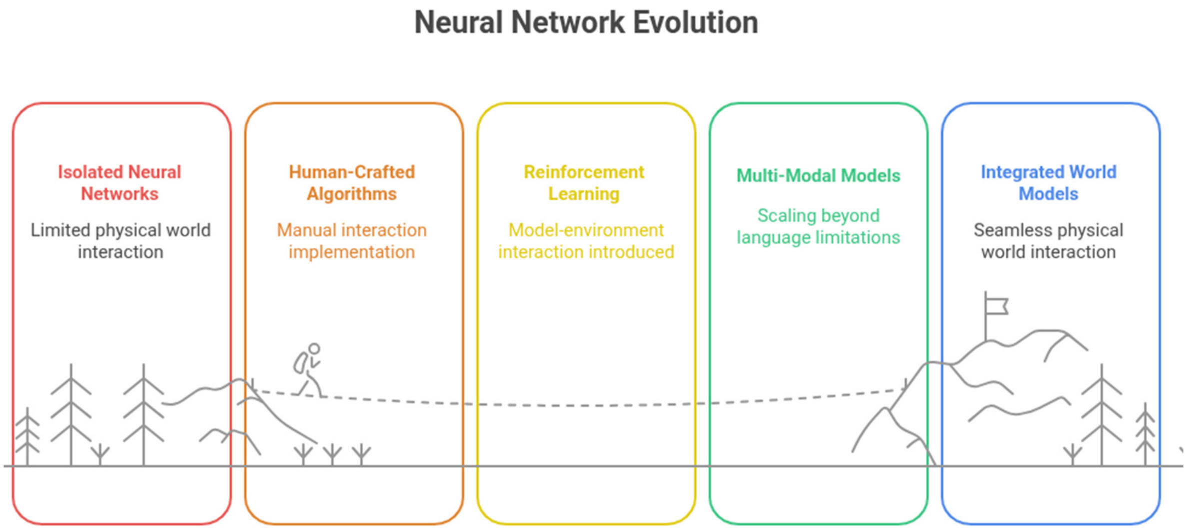

Neural Network Evolution
Isolated Neural Networks
Limited physical world interaction
Human-Crafted Algorithms
Manual interaction implementation
Reinforcement Learning
Model-environment interaction introduced
Multi-Modal Models
Scaling beyond language limitations
Integrated World Models
Seamless physical world interaction

Figure 2. Neural network evolution with physical worlds.

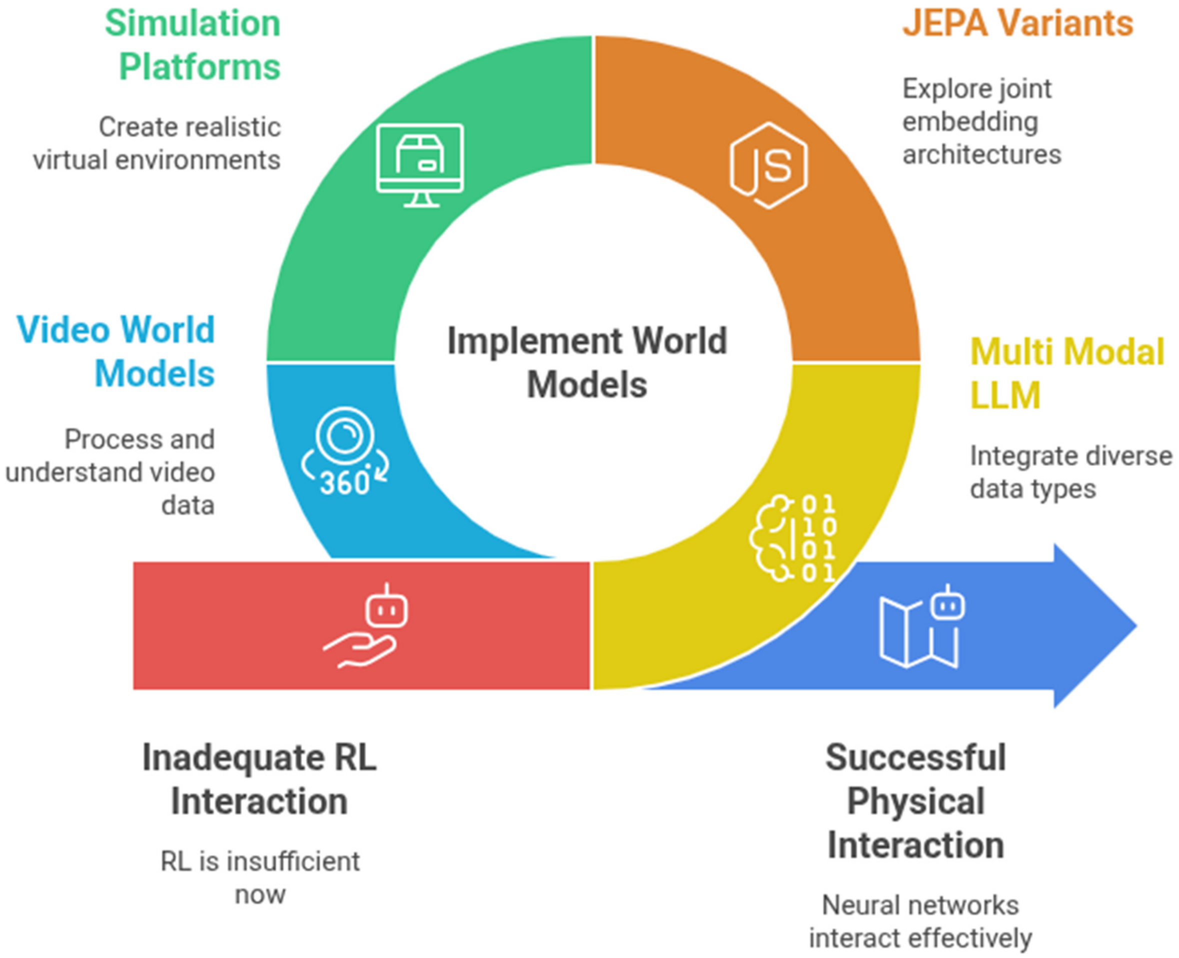
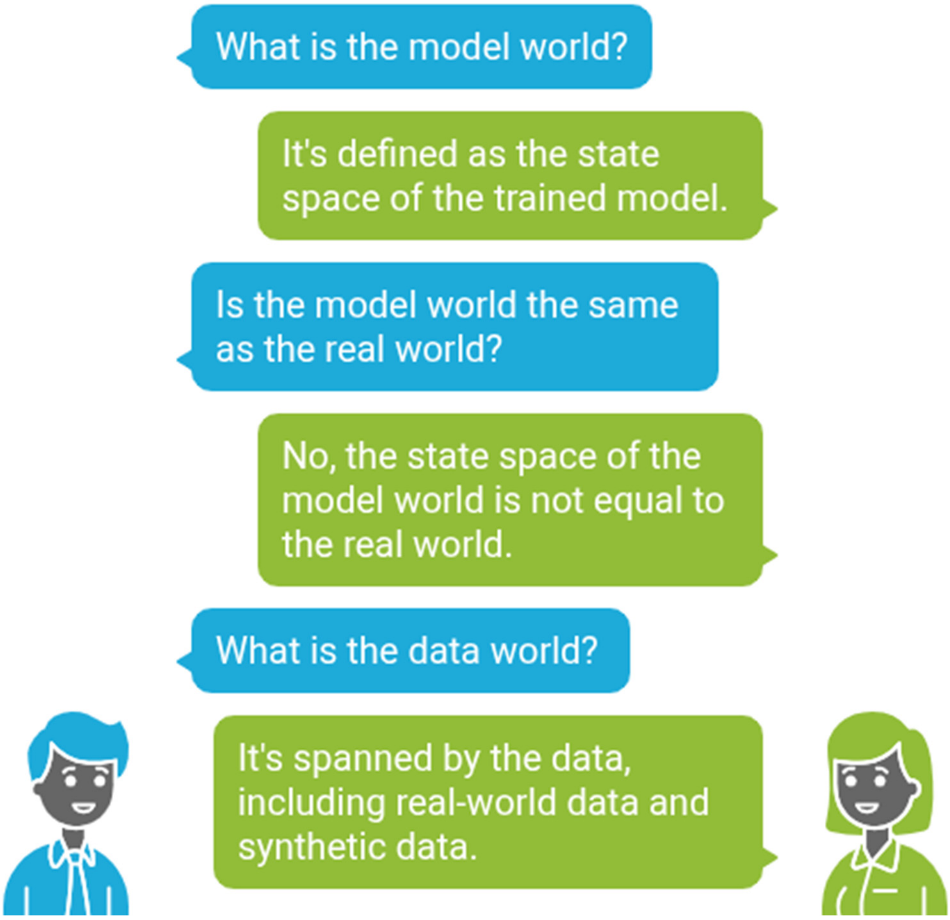

Simulation Platforms
Create realistic virtual environments
JEPA Variants
Explore joint embedding architectures
Video World Models
Process and understand video data
Implement World Models
Multi Modal LLM
Integrate diverse data types
360
Inadequate RL Interaction
RL is insufficient now
Successful Physical Interaction
Neural networks interact effectively

Figure 3. Approaches toward the neural-network-in-the-loop system.



What is the model world?

It's defined as the state space of the trained model.

Is the model world the same as the real world?

No, the state space of the model world is not equal to the real world.

What is the data world?

It's spanned by the data, including real-world data and synthetic data.

Figure 4. Key points of the three-world hierarchy.

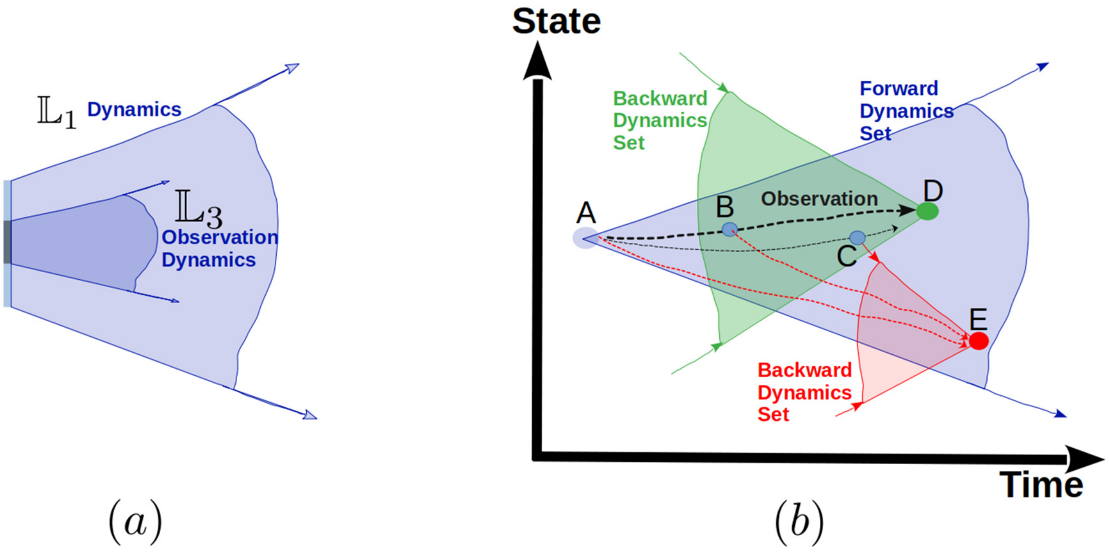


Figure 5. Time evolution of the world.

3. Frechet World Distance for Multi-Modal Configurations

Having outlined the three-level world hierarchy, we need tools to quantify the relationships between these distinct levels. We draw inspiration from Fréchet distance (which measures the distance between distributions) and Dynamic Time Warping (which considers temporal dimensions) to introduce a new concept: **Fréchet World Distance (FWD)**. FWD extends these established metrics to address both single-modal scenarios (e.g., only visual data) and multi-modal worlds (e.g., visual, audio, and language data). To further account for the temporal evolution within these worlds, we also propose forward and backward Fréchet World Distance as more detailed measurements for understanding world models.

3.1. Conventional Frechet Distance

We firstly brief the definition of the FD. The FD measure the similarity between two distributions. One popular application of FD is to quantify the quality of a generative model. Suppose the distribution from the outputs of a generative model and reference distribution are $\mathrm{P_G}$ and P_R. 'G' refer to generative and 'R' refers to reference. Then the FD is defined as:

$$\mathrm{d}(\mathrm{P_G}, \mathrm{P_R}) = \min_{\mathrm{X \sim P_G, Y \sim P_R}} (E|X - Y|^2) \quad (1)$$

The above results are difficult to derive for a general case. However, if both distributions are of Gaussian type, then:

$$\mathrm{d}(\mathrm{P_G}, \mathrm{P_R}) = |\mu_R - \mu_G|^2 + Tr\left(\Sigma_R + \Sigma_G - 2(\Sigma_R \Sigma_G)^{1/2}\right) \quad (2)$$

μ_R and μ_G are mean vector of the two distributions, while Σ_R and Σ_G are co-variance matrices.

3.2. Frechet World Distance for Unimodal C

Firstly we only consider single modal world. The modal is embedded in a context representation C. Given the state at moment t $\mathrm{X_t}$ in world $\mathbb{L}_2$ (the model world), the neural network (which is trained with dataset from world $\mathbb{L}_3$) output the action $\mathrm{y_t}$ with some desired objective, as given in Figure 6. The state $\mathrm{X_t}$ and $\mathrm{y_t}$ leads the evolution of the system state to a new state $\mathrm{S}_{t+1}^{\mathbb{L}_2}$. At the same time, the nn sense the state X_{t+1}. Note that the difference between $\mathrm{S}_{t+1}^{\mathbb{L}_2}$ and $\mathrm{S}_{t+1}^{\mathbb{L}_1}$ are characterized by two set-difference operation of $\mathbb{L}_1 \backslash \mathbb{L}_2$ and $\mathbb{L}_2 \backslash \mathbb{L}_1$.

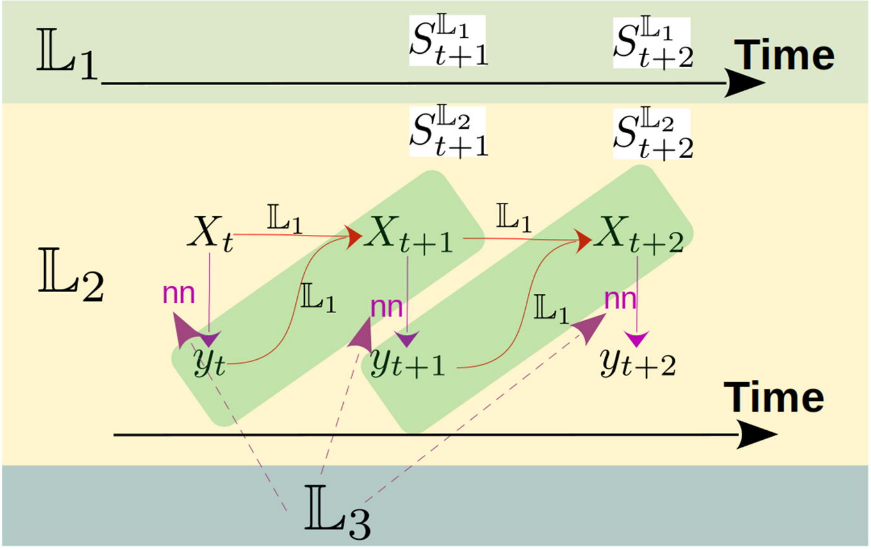


Figure 6. Time evolution of the world.

Combining the above description, we have two trajectories from two worlds:

$$\begin{cases} \boldsymbol{S}^{\mathbb{L}_1}: S_0^{\mathbb{L}_1} \to S_1^{\mathbb{L}_1} \to S_2^{\mathbb{L}_1} \ \ldots\ldots \\ \boldsymbol{S}^{\mathbb{L}_2}: S_0^{\mathbb{L}_2} \to S_1^{\mathbb{L}_2} \to S_2^{\mathbb{L}_2} \ldots\ldots \end{cases} \quad (3)$$

The distribution of the initial state for the two worlds are represented as $\pi_0^{\mathbb{L}_1}$ and $\pi_0^{\mathbb{L}_2}$. As the system is stochastic, the distribution of the state at moment t is $\mathrm{p}\left(S_t^{\mathbb{L}_1}\right)$ for $\mathbb{L}_1$ and is $\mathrm{p}\left(S_t^{\mathbb{L}_2}\right)$ for $\mathbb{L}_2$. Considering the time horizon between $[0, T]$ and combining the idea from dynamic time wrapping, we define a FWD of the power q for the two worlds $\mathbb{L}_1$ and $\mathbb{L}_2$:

$$\mathrm{FWD}_{\mathrm{q}}^{C}(\mathbb{L}_1,\mathbb{L}_2)=\min_{\lambda\in\mathcal{A}(S^{\mathbb{L}_1},S^{\mathbb{L}_2})}\left(\sum_{(i,j)\in\lambda}\left(FD\left(\mathrm{p}\left(S_i^{\mathbb{L}_1}\right),\mathrm{p}\left(S_j^{\mathbb{L}_2}\right)\right)\right)^{q}\right)^{1\backslash q}\qquad(4)$$

λ is a alignment path of length K index pairs $\left((i_0,j_0),\ldots.(i_{K-1},j_{K-1})\right)$, and $\mathcal{A}(S^{\mathbb{L}_1},S^{\mathbb{L}_2})$ is all admissible alignment path between two trajectories $S^{\mathbb{L}_1}$ and $S^{\mathbb{L}_2}$. An alignment path should satisfy the following conditions:

- Beginning and ending of the sequence are matched together:
  - ■ $\lambda_0=(0{,}0)$
  - ■ $\lambda_{\mathrm{K}-1}=(T-1,T-1)$
- The sequence is monotonically increasing in both i and j:
  - ■ $\mathrm{i}_{k-1}\leq\mathrm{i}_{\mathrm{k}}\leq i_{k-1}+1$
  - ■ $\mathrm{j}_{k-1}\leq\mathrm{j}_{\mathrm{k}}\leq j_{k-1}+1$

3.3. Frechet World Distance for Multi Modal

The above formulation Eq. 4 is developed for single modal trajectory. However, many models are trained via multi-modal information. As an example, autonomous vehicles end2end model are trained with video, LiDAR, Radar etc. Therefore, a world is characterized via the information from many modals. To extend the Frechet world distance to multi-modal setting, we combine the idea from contrastive learning. Suppose there are N modals $\mathrm{C_i}$, $\mathrm{i}=1{,}2,\ldots\mathrm{N}$. For each modal $\mathrm{C_i}$, we apply an modal-specific encoder $\mathrm{ENC}_{C_i}(\mathrm{S}_{\mathrm{t}}^{\mathbb{L}_{\mathrm{k}}})$ to extract and encode the information at moment t for world $\mathbb{L}_{\mathrm{k}}$. The output $\mathrm{U}_{\mathrm{t,i}}^{\mathbb{L}_{\mathrm{k}}}=\mathrm{ENC}_{C_i}(\mathrm{S}_{\mathrm{t}}^{\mathbb{L}_{\mathrm{k}}})$ resides at a shared vector space across all modals. As the multi-modal state space $\mathrm{S}_{\mathrm{t}}^{\mathbb{L}_{\mathrm{k}}}$ follows the distribution $\mathrm{p}\left(S_t^{\mathbb{L}_{\mathrm{k}}}\right)$, and thus the modal specific distribution is $\mathrm{p}\left(\mathrm{U}_{\mathrm{t,i}}^{\mathbb{L}_{\mathrm{k}}}\right)$. We use this distribution to extend the FWD. For a set of modals $\{C_k\}$, the FWD is defined as follows:

$$\mathrm{FWD}_{\mathrm{q}}^{\{C_k\}}(\mathbb{L}_1,\mathbb{L}_2)=\min_{\mathrm{u,v}\in\{C_k\}}\min_{\lambda\in\mathcal{A}(S^{\mathbb{L}_1},S^{\mathbb{L}_2})}\left(\sum_{(i,j)\in\lambda}\left(FD\left(\mathrm{p}\left(\mathrm{U}_{\mathrm{t,u}}^{\mathbb{L}_1}\right),\mathrm{p}\left(\mathrm{U}_{\mathrm{t,v}}^{\mathbb{L}_2}\right)\right)\right)^{q}\right)^{1\backslash q}\qquad(5)$$

The above definition iterate over the possible modal u from the first world $\mathbb{L}_1$, and the target modal v from the second world $\mathbb{L}_2$, and calculate the FWD distance respectively.

4. Related Work

4.1. World Model

World model is designed to mimic the human brain mechanism "based on what they are able to perceive with their limited senses" (Ha & Schmidhuber, 2018; Assran et al., 2025; Feng et al., 2022; Fu et al., 2023; Ha & Schmidhuber, 2018; Hong et al., 2023; Jiang et al., 2024; Karypidis et al., 2025a, 2025b; Ren et al., 2025; Rimon et al., 2024; Wang et al., 2023; Zhen et al., 2024; Zürn et al., 2024). The agent model includes three components, Vision (V), Memory (M), and Controller. The goals of the world model include understanding and predicting the abstraction of the world and generating/simulate the detail of the world. The world model concept then has been under discussion and become an overloaded word. It is extended beyond the pixel space and language space (i.e., LLM) to video (Ren et al., 2025), LiDAR (Zyrianov et al., 2024), Occupancy (Zheng et al., 2023), etc. Later one, the multi-modal world is conceived (Ashuach et al., 2023; Radford et al., 2021).

Except perception, simulation and prediction of the world, one core function of the world model is to plan the actions embodied environments (Zhen et al., 2024). The interaction between neural network represented model and the world can be achieved via a series of interaction tokens (Zhen et al., 2024), or standalone sub-neural-network modules (Ha & Schmidhuber, 2018). To achieve realistic interaction, the dynamics of the world should be learned, for instance, via the latent dynamics model (Ren et al., 2025; Sobal et al., 2025), or the reinforcement learning (Rimon et al., 2024). World models still is on its early stage. Further investigation is required to study its properties, for instance ergodicity (Bilaloglu et al., 2023), reachability (Fu et al., 2023), etc.

4.2. JEPA and Its Variants

In his seminar paper (LeCun, 2022), LeCun claimed that "best ML systems are still very far from matching human reliability in real-world tasks such as driving". And he listed three challenges of ML systems: represent the world, learn to predict, and learn to act; reason and plan in ways that are compatible with gradient-based learning; learn to represent percepts and action plans in a hierarchical manner. Corresponding to the above claims, he proposed a framework termed JEPA (Joint-Embedding Predictive Architecture) that encapsulate the world model as its components. Other modules include configurator module, configurator module, cost module, short-term memory module and actor module. Given the architecture, the perception-action loop have two modes: mode 1 and mode 2. Since then, the JEPA are extended to many field (Assran et al., 2023a, 2023b, 2025; Bardes et al., 2023a, 2023b, 2024), encompassing many modals (Bardhan et al., 2025; Chen et al., 2025; Dong et al., 2024; Fei et al., 2023, 2024; Fu et al., 2024; Ghaemi et al., 2025;) including LiDAR (Zhu et al., 2025), audio (Fei et al., 2023), point cloud (Saito et al., 2025), etc (Girgis et al., 2025; Guetschel et al., n.d.; Hu et al., 2024; Kalapos & Gyires-Tóth, 2024; Kenneweg et al., 2025; LeCun, 2022; Li et al., 2024;).

Although JETA in its essence is not a generative model, its combination with generative models and also reinforcement learning (Kenneweg et al., 2025), imitation learning (Vujinović & Kovačević, n.d.), is appealing (Chen et al., 2025).

JEPA with its variants (Mahowald et al., 2023; Mo & Yun, 2024; Riou et al., 2024; Saito et al., 2025; Thimonier et al., 2025; Vujinović & Kovačević, n.d.; Weimann & Conrad, 2024; Zhu et al., 2025) has been successfully applied to different physical world such as collider physics (Bardhan et al., 2025).

4.3. Frechet Distance

The Fréchet distance (FD) is a measure of similarity between curves that takes into account the location and ordering of the points along the curves. The Fréchet distance quantifies the similarity between two curves. Consider two points, each tracing a finite, curvilinear path, such as a person and their dog on a leash. Both entities can adjust their speeds along their respective paths, but movement is restricted to a forward direction. The Fréchet distance is then defined as the minimum leash length required for both to complete their traversals from start to finish without the leash ever becoming taut. This metric is symmetric; the calculated distance remains constant irrespective of which curve is designated as the primary or secondary path.

FD has been successfully extended in the generative modeling (Bringmann et al., 2019; Ciesielski & Lewicki, 2021; Conradi et al., 2024; Dodson, 2011; Driemel & Har-Peled, 2013;). The seminal work is FID (Fréchet inception distance) that measure the quality of generative image neural network models with a reference image set. The Fréchet Inception Distance (FID), or specialized variants thereof, serves as an evaluation metric across diverse domains (Eftekharinasab, 2022; Farahbakhsh Touli, 2020; Gui et al., 2023; Kilgour et al., 2018; Li et al., 2024; Ramos et al., 2018;). For instance, the Fréchet Audio Distance (FAD) assesses music enhancement algorithms, while the Fréchet Video Distance (FVD) evaluates generative models of video. Similarly, the Fréchet ChemNet Distance (FCD) is employed to evaluate AI-generated molecules. However, it should be noted that, only in limited circumstance, the FD can be derived as a analytical form (Reimering et al., 2018; Retkowski et al., 2024; Soloveitchik et al., 2021a, 2021b; Vodolazskiy, 2021).

4.4. Potential Harms of Neural Networks Model in Physical World

Although the neural networks gain great success, its deployment into real world still faces considerable challenges. Given the fact that the practice is way ahead of the theory investigation, the understanding of neural networks still is limited. The magic of neural network are characterized by scale up law: the scale up in data, network size and training. During the test and utilization of the models, researchers find the negative aspects of the models. These negative aspects include reversal curse (Berglund et al., 2023; Zhu et al., 2024), hallucination (Ainslie et al., 2023; Aithal et al., 2024a, 2024b; Bai et al., 2024; Chrysos et al., n.d.; Du et al., 2023; Gao et al., 2024; Jesson et al., 2024; Kim et al., 2024; Lauscher et al., 2025; Oorloff et al., 2025; Rani et al., 2024; Rathkopf, 2025; Sahoo et al., 2024; Shazeer, 2020; Tauman Kalai & Vempala, 2023; H. Wang et al., 2024; Z. Wang et al., 2024; Wei et al., 2024; Xu et al., 2024; Zhang & Sennrich, 2019), interpretability (Conmy et al., 2023; Dunefsky et al., 2024; Gurnee et al., 2023; Kissane et al., 2024; Kramár et al., 2024; Makelov et al., 2024; Marks et al., 2024; Rajamanoharan et al., 2024; Sharkey et al., 2025) etc. Hence it is crucial to build a safe, responsible AI (Batool et al., 2023; Ghamisi et al., 2024; Goellner et al., 2024; Pi, 2023; Wang et al., 2024) stacks that can work as desired in physical world, and still have the capabilities to explore the world in ergodic way.

5. Discussion and Conclusion

In conclusion, the imminent convergence of neural networks with the physical world necessitates a deeper theoretical understanding to mitigate inherent risks. Currently, the practical application of neural networks outpaces our theoretical grasp, leading to potential vulnerabilities in their development and deployment. To address this, we propose a three-tiered world hierarchy: the data world, the model world, and the real world. Furthermore, the introduction of Fréchet World Distance offers a robust metric for quantifying the discrepancies between single and multi-modal representations within these interconnected domains. This framework is crucial for fostering safer and more reliable advancements in neural network technology as it becomes increasingly integrated with our physical reality.

Future works including:

1) Develop three-worlds hierarchy JEPA models to encapsulate complex dynamics. In the JEPA model, the cost modules can be implemented via the Frechet world distance.
2) Develop models that detect the hallucination of the neural network models. The hallucination is defined over the difference operation between model world and the real world. Given that the FWD can measure the difference between different levels of world, the FWD can serve as objective that train the models with minimal hallucination.

Disclaimer

The author hereby declares: 1) Originality: The submitted manuscript is an original work and has not been previously published, nor is it under consideration for publication elsewhere; 2) Authorship: All listed authors have contributed significantly to the research and manuscript preparation. No individuals who meet authorship criteria have been omitted; 3) Conflicts of Interest: Any potential conflicts of interest, financial or otherwise, have been disclosed in the manuscript or cover letter. 4) Ethical Compliance: The research complies with all relevant ethical guidelines, including (if applicable) approval from an institutional review board (IRB) or ethics committee. 5) Copyright & Permissions: All third-party materials (figures, tables, text excerpts) used in the manuscript have been properly cited, and necessary permissions have been obtained where required. 6) Data Availability: The data presented in this manuscript are accurate to the best of the author' knowledge, and fabricated or falsified data have not been included. 7) Liability: The author takes full responsibility for the content of the manuscript and any errors or omissions therein.

doi:10.20944/preprints202508.1078.v1

References

Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., & Sanghai, S. (2023). GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. arXiv E-Prints, arXiv:2305.13245. https://doi.org/10.48550/arXiv.2305.13245

Aithal, S. K., Maini, P., Lipton, Z. C., & Kolter, J. Z. (2024a). Understanding Hallucinations in Diffusion Models through Mode Interpolation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Eds.), Advances in Neural Information Processing Systems (Vol. 37, pp. 134614–134644). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2024/file/f29369d192b13184b65c6d2515474d78-Paper-Conference.pdf

Ashuach, T., Gabitto, M. I., Koodli, R. V., Saldi, G.-A., Jordan, M. I., & Yosef, N. (2023). MultiVI: deep generative model for the integration of multimodal data. Nature Methods, 20(8), 1222-+. https://doi.org/10.1038/s41592-023-01909-9

Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Mojtaba, Komeili, Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., Arnaud, S., Gejji, A., Martin, A., Hogan, F. R., Dugas, D., Bojanowski, P., Khalidov, V., … Ballas, N. (2025). V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. arXiv E-Prints, arXiv:2506.09985. https://doi.org/10.48550/arXiv.2506.09985

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., & Ballas, N. (2023a). Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. arXiv E-Prints, arXiv:2301.08243. https://doi.org/10.48550/arXiv.2301.08243

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., & Ballas, N. (2023b, April). Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. arXiv. https://doi.org/10.48550/arXiv.2301.08243

Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., & Shou, M. Z. (2024). Hallucination of Multimodal Large Language Models: A Survey. arXiv E-Prints, arXiv:2404.18930. https://doi.org/10.48550/arXiv.2404.18930

Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., & Ballas, N. (2024, February). Revisiting Feature Prediction for Learning Visual Representations from Video. arXiv. https://doi.org/10.48550/arXiv.2404.08471

Bardes, A., Ponce, J., & LeCun, Y. (2023a). MC-JEPA: A Joint-Embedding Predictive Architecture for Self-Supervised Learning of Motion and Content Features. arXiv E-Prints, arXiv:2307.12698. https://doi.org/10.48550/arXiv.2307.12698

Bardes, A., Ponce, J., & LeCun, Y. (2023b, July). MC-JEPA: A Joint-Embedding Predictive Architecture for Self-Supervised Learning of Motion and Content Features. arXiv. https://doi.org/10.48550/arXiv.2307.12698

Bardhan, J., Agrawal, R., Tilak, A., Neeraj, C., & Mitra, S. (2025, February). HEP-JEPA: A foundation model for collider physics using joint embedding predictive architecture. arXiv. https://doi.org/10.48550/arXiv.2502.03933

Batool, A., Zowghi, D., & Bano, M. (2023). Responsible AI Governance: A Systematic Literature Review. arXiv E-Prints, arXiv:2401.10896. https://doi.org/10.48550/arXiv.2401.10896

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Cooper Stickland, A., Korbak, T., & Evans, O. (2023). The Reversal Curse: LLMs trained on “A is B” fail to learn “B is A.” arXiv E-Prints, arXiv:2309.12288. https://doi.org/10.48550/arXiv.2309.12288

Bilaloglu, C., Löw, T., & Calinon, S. (2023, November). Whole-Body Ergodic Exploration with a Manipulator Using Diffusion. arXiv. https://doi.org/10.48550/arXiv.2306.16898

Bringmann, K., Künnemann, M., & Nusser, A. (2019). Walking the Dog Fast in Practice: Algorithm Engineering of the Fréchet Distance. arXiv E-Prints, arXiv:1901.01504. https://doi.org/10.48550/arXiv.1901.01504

Chen, D., Hu, J., Wei, X., & Wu, E. (2025, February). Denoising with a Joint-Embedding Predictive Architecture. arXiv. https://doi.org/10.48550/arXiv.2410.03755

Chrysos, G., Li, Y., Angelopoulos, A. N., Bates, S., Plank, B., & Khan, M. E. (n.d.). Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI. ICLR 2025 Workshop Proposals.

Ciesielski, M., & Lewicki, G. (2021). Admissibility of Frechet spaces. arXiv E-Prints, arXiv:2112.00349. https://doi.org/10.48550/arXiv.2112.00349

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). Towards Automated Circuit Discovery for Mechanistic Interpretability. arXiv E-Prints, arXiv:2304.14997. https://doi.org/10.48550/arXiv.2304.14997

Conradi, J., Driemel, A., & Kolbe, B. (2024). Revisiting the Fréchet distance between piecewise smooth curves. arXiv E-Prints, arXiv:2401.03339. https://doi.org/10.48550/arXiv.2401.03339

Dodson, C. T. J. (2011). Some recent work in Frechet geometry. arXiv E-Prints, arXiv:1109.4241. https://doi.org/10.48550/arXiv.1109.4241

Dong, Z., Li, R., Wu, Y., Nguyen, T. T., Chong, J. S. X., Ji, F., Tong, N. R. J., Chen, C. L. H., & Zhou, J. H. (2024, September). Brain-JEPA: Brain Dynamics Foundation Model with Gradient Positioning and Spatiotemporal Masking. arXiv. https://doi.org/10.48550/arXiv.2409.19407

Driemel, A., & Har-Peled, S. (2013). Jaywalking Your Dog: Computing the Fréchet Distance with Shortcuts. SIAM Journal on Computing, 42(5), 1830–1866. https://doi.org/10.1137/120865112

Du, L., Wang, Y., Xing, X., Ya, Y., Li, X., Jiang, X., & Fang, X. (2023). Quantifying and attributing the hallucination of large language models via association analysis. arXiv Preprint arXiv:2309.05217.

Dunefsky, J., Chlenski, P., & Nanda, N. (2024). Transcoders Find Interpretable LLM Feature Circuits. arXiv E-Prints, arXiv:2406.11944. https://doi.org/10.48550/arXiv.2406.11944

Eftekharinasab, K. (2022). Multiplicity Theorems for Frechet Manifolds. arXiv E-Prints, arXiv:2210.09270. https://doi.org/10.48550/arXiv.2210.09270

Farahbakhsh Touli, E. (2020). Frechet-Like Distances between Two Merge Trees. arXiv E-Prints, arXiv:2004.10747. https://doi.org/10.48550/arXiv.2004.10747

Fei, Z., Fan, M., & Huang, J. (2023). A-JEPA: Joint-Embedding Predictive Architecture Can Listen. arXiv E-Prints, arXiv:2311.15830. https://doi.org/10.48550/arXiv.2311.15830

Feng, Z., Guo, S., Tan, X., Xu, K., Wang, M., & Ma, L. (2022). Rethinking efficient lane detection via curve modeling. Computer Vision and Pattern Recognition.

Fu, Y., Peng, R., & Lee, H. (2023). Go Beyond Imagination: Maximizing Episodic Reachability with World Models. arXiv E-Prints, arXiv:2308.13661. https://doi.org/10.48550/arXiv.2308.13661

Fu, Y., Anantha, R., Vashisht, P., Cheng, J., & Littwin, E. (2024, October). UI-JEPA: Towards Active Perception of User Intent through Onscreen User Activity. arXiv. https://doi.org/10.48550/arXiv.2409.04081

Gao, N., Li, J., Huang, H., Zeng, Z., Shang, K., Zhang, S., & He, R. (2024). Diffmac: Diffusion manifold hallucination correction for high generalization blind face restoration. arXiv Preprint arXiv:2403.10098.

Ghaemi, H., Muller, E., & Bakhtiari, S. (2025, May). seq-JEPA: Autoregressive Predictive Learning of Invariant-Equivariant World Models. arXiv. https://doi.org/10.48550/arXiv.2505.03176

Ghamisi, P., Yu, W., Marinoni, A., Gevaert, C. M., Persello, C., Selvakumaran, S., Girotto, M., Horton, B. P., Rufin, P., Hostert, P., Pacifici, F., & Atkinson, P. M. (2024). Responsible AI for Earth Observation. arXiv E-Prints, arXiv:2405.20868. https://doi.org/10.48550/arXiv.2405.20868

Girgis, A. M., Valcarce, A., & Bennis, M. (2025, July). Time-Series JEPA for Predictive Remote Control under Capacity-Limited Networks. arXiv. https://doi.org/10.48550/arXiv.2406.04853

Goellner, S., Tropmann-Frick, M., & Brumen, B. (2024). Responsible Artificial Intelligence: A Structured Literature Review. arXiv E-Prints, arXiv:2403.06910. https://doi.org/10.48550/arXiv.2403.06910

Guan, Y., Liao, H., Li, Z., Hu, J., Yuan, R., Li, Y., Zhang, G., & Xu, C. (2024). World Models for Autonomous Driving: An Initial Survey. https://arxiv.org/abs/2403.02622

Guetschel, P., Moreau, T., & Tangermann, M. (n.d.). S-JEPA: Towards seamless cross-dataset transfer through dynamic spatial attention. https://doi.org/10.3217/978-3-99161-014-4-003

Gui, A., Gamper, H., Braun, S., & Emmanouilidou, D. (2023). Adapting Frechet Audio Distance for Generative Music Evaluation. arXiv E-Prints, arXiv:2311.01616. https://doi.org/10.48550/arXiv.2311.01616

Gurnee, W., Nanda, N., Pauly, M., Harvey, K., Troitskii, D., & Bertsimas, D. (2023). Finding Neurons in a Haystack: Case Studies with Sparse Probing. arXiv E-Prints, arXiv:2305.01610. https://doi.org/10.48550/arXiv.2305.01610

Ha, D., & Schmidhuber, J. (2018). World Models. arXiv E-Prints, arXiv:1803.10122. https://doi.org/10.48550/arXiv.1803.10122

Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., & Gan, C. (2023). 3D-LLM: Injecting the 3D World into Large Language Models. arXiv E-Prints, arXiv:2307.12981. https://doi.org/10.48550/arXiv.2307.12981

Hu, N., Cheng, H., Xie, Y., Li, S., & Zhu, J. (2024, September). 3D-JEPA: A Joint Embedding Predictive Architecture for 3D Self-Supervised Representation Learning. arXiv. https://doi.org/10.48550/arXiv.2409.15803

Jesson, A., Beltran Velez, N., Chu, Q., Karlekar, S., Kossen, J., Gal, Y., Cunningham, J. P., & Blei, D. (2024). Estimating the hallucination rate of generative ai. Advances in Neural Information Processing Systems, 37, 31154–31201.

Jiang, J., Hong, G., Zhou, L., Ma, E., Hu, H., Zhou, X., Xiang, J., Liu, F., Yu, K., Sun, H., Zhan, K., Jia, P., & Zhang, M. (2024). DiVE: DiT-based Video Generation with Enhanced Control. arXiv E-Prints, arXiv:2409.01595. https://doi.org/10.48550/arXiv.2409.01595

Kalapos, A., & Gyires-Tóth, B. (2024). CNN-JEPA: Self-Supervised Pretraining Convolutional Neural Networks Using Joint Embedding Predictive Architecture. 1111–1114. https://doi.org/10.1109/ICMLA61862.2024.00169

Karypidis, E., Kakogeorgiou, I., Gidaris, S., & Komodakis, N. (2025a). Advancing Semantic Future Prediction through Multimodal Visual Sequence Transformers. arXiv Preprint arXiv:2501.08303.

Kenneweg, T., Kenneweg, P., & Hammer, B. (2025). JEPA for RL: Investigating Joint-Embedding Predictive Architectures for Reinforcement Learning. arXiv E-Prints, arXiv:2504.16591. https://doi.org/10.48550/arXiv.2504.16591

Kilgour, K., Zuluaga, M., Roblek, D., & Sharifi, M. (2018). Fréchet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms. arXiv E-Prints, arXiv:1812.08466. https://doi.org/10.48550/arXiv.1812.08466

Kim, S., Jin, C., Diethe, T., Figini, M., Tregidgo, H. F., Mullokandov, A., Teare, P., & Alexander, D. C. (2024). Tackling structural hallucination in image translation with local diffusion. European Conference on Computer Vision, 87–103.

Kissane, C., Krzyzanowski, R., Bloom, J. I., Conmy, A., & Nanda, N. (2024). Interpreting Attention Layer Outputs with Sparse Autoencoders. arXiv E-Prints, arXiv:2406.17759. https://doi.org/10.48550/arXiv.2406.17759

Kramár, J., Lieberum, T., Shah, R., & Nanda, N. (2024). AtP*: An efficient and scalable method for localizing LLM behaviour to components. arXiv E-Prints, arXiv:2403.00745. https://doi.org/10.48550/arXiv.2403.00745

Lauscher, A., Glavaš, G., & others. (2025). How Much Do LLMs Hallucinate across Languages? On Multilingual Estimation of LLM Hallucination in the Wild. arXiv Preprint arXiv:2502.12769.

Oorloff, T., Yacoob, Y., & Shrivastava, A. (2025). Mitigating Hallucinations in Diffusion Models through Adaptive Attention Modulation. arXiv Preprint arXiv:2502.16872.

LeCun, Y. (2022). A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. Open Review, 62(1), 1–62.

Li, W., Wei, Y., Liu, T., Hou, Y., Li, Y., Liu, Z., Liu, Y., & Liu, L. (2024). Predicting Gradient is Better: Exploring Self-Supervised Learning for SAR ATR with a Joint-Embedding Predictive Architecture. ISPRS Journal of Photogrammetry and Remote Sensing, 218, 326–338. https://doi.org/10.1016/j.isprsjprs.2024.09.013

Li, Y., Gui, A., Emmanouilidou, D., & Gamper, H. (2024). Rethinking Emotion Bias in Music via Frechet Audio Distance. arXiv E-Prints, arXiv:2409.15545. https://doi.org/10.48550/arXiv.2409.15545

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models. arXiv E-Prints, arXiv:2301.06627. https://doi.org/10.48550/arXiv.2301.06627

Makelov, A., Lange, G., & Nanda, N. (2024). Towards Principled Evaluations of Sparse Autoencoders for Interpretability and Control. arXiv E-Prints, arXiv:2405.08366. https://doi.org/10.48550/arXiv.2405.08366

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., & Mueller, A. (2024). Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models. arXiv E-Prints, arXiv:2403.19647. https://doi.org/10.48550/arXiv.2403.19647

Mo, S., & Yun, S. (2024, May). DMT-JEPA: Discriminative Masked Targets for Joint-Embedding Predictive Architecture. arXiv. https://doi.org/10.48550/arXiv.2405.17995

Pi, Y. (2023). Beyond XAI:Obstacles Towards Responsible AI. arXiv E-Prints, arXiv:2309.03638. https://doi.org/10.48550/arXiv.2309.03638

doi:10.20944/preprints202508.1078.v1

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv E-Prints, arXiv:2103.00020. https://doi.org/10.48550/arXiv.2103.00020

Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramár, J., Shah, R., & Nanda, N. (2024). Improving Dictionary Learning with Gated Sparse Autoencoders. arXiv E-Prints, arXiv:2404.16014. https://doi.org/10.48550/arXiv.2404.16014

Ramos, P. L., Louzada, F., Ramos, E., & Dey, S. (2018). The Frechet distribution: Estimation and Application an Overview. arXiv E-Prints, arXiv:1801.05327. https://doi.org/10.48550/arXiv.1801.05327

Rani, A., Rawte, V., Sharma, H., Anand, N., Rajbangshi, K., Sheth, A., & Das, A. (2024). Visual hallucination: Definition, quantification, and prescriptive remediations. arXiv Preprint arXiv:2403.17306.

Rathkopf, C. (2025). Hallucination, reliability, and the role of generative AI in science. arXiv Preprint arXiv:2504.08526.

Reimering, S., Muñoz, S., & McHardy, A. C. (2018). A Fréchet tree distance measure to compare phylogeographic spread paths across trees. Scientific Reports, 8(1), 17000. https://doi.org/10.1038/s41598-018-35421-4

Ren, Z., Wei, Y., Guo, X., Zhao, Y., Kang, B., Feng, J., & Jin, X. (2025). VideoWorld: Exploring Knowledge Learning from Unlabeled Videos. arXiv E-Prints, arXiv:2501.09781. https://doi.org/10.48550/arXiv.2501.09781

Retkowski, J., Stępniak, J., & Modrzejewski, M. (2024). Frechet Music Distance: A Metric For Generative Symbolic Music Evaluation. arXiv E-Prints, arXiv:2412.07948. https://doi.org/10.48550/arXiv.2412.07948

Rimon, Z., Jurgenson, T., Krupnik, O., Adler, G., & Tamar, A. (2024). MAMBA: an Effective World Model Approach for Meta-Reinforcement Learning. arXiv E-Prints, arXiv:2403.09859. https://doi.org/10.48550/arXiv.2403.09859

Riou, A., Lattner, S., Hadjeres, G., Anslow, M., & Peeters, G. (2024, August). Stem-JEPA: A Joint-Embedding Predictive Architecture for Musical Stem Compatibility Estimation. arXiv. https://doi.org/10.48550/arXiv.2408.02514

Sahoo, P., Meharia, P., Ghosh, A., Saha, S., Jain, V., & Chadha, A. (2024). A comprehensive survey of hallucination in large language, image, video and audio foundation models. arXiv Preprint arXiv:2405.09589.

Saito, A., Kudeshia, P., & Poovvancheri, J. (2025, February). Point-JEPA: A Joint Embedding Predictive Architecture for Self-Supervised Learning on Point Cloud. arXiv. https://doi.org/10.48550/arXiv.2404.16432

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., Biderman, S., Garriga-Alonso, A., Conmy, A., Nanda, N., Rumbelow, J., Wattenberg, M., Schoots, N., Miller, J., Michaud, E. J., … McGrath, T. (2025). Open Problems in Mechanistic Interpretability. arXiv E-Prints, arXiv:2501.16496. https://doi.org/10.48550/arXiv.2501.16496

Shazeer, N. (2020). GLU Variants Improve Transformer. arXiv E-Prints, arXiv:2002.05202. https://doi.org/10.48550/arXiv.2002.05202

Sobal, V., Zhang, W., Cho, K., Balestriero, R., Rudner, T. G. J., & LeCun, Y. (2025). Learning from Reward-Free Offline Data: A Case for Planning with Latent Dynamics Models. arXiv E-Prints, arXiv:2502.14819. https://doi.org/10.48550/arXiv.2502.14819

Soloveitchik, M., Diskin, T., Morin, E., & Wiesel, A. (2021a). Conditional Frechet Inception Distance. arXiv E-Prints, arXiv:2103.11521. https://doi.org/10.48550/arXiv.2103.11521

Suzuki, M., & Matsuo, Y. (2022). A survey of multimodal deep generative models. Advanced Robotics, 36(5–6), 261–278. https://doi.org/10.1080/01691864.2022.2035253

Tauman Kalai, A., & Vempala, S. S. (2023). Calibrated Language Models Must Hallucinate. arXiv E-Prints, arXiv:2311.14648. https://doi.org/10.48550/arXiv.2311.14648

Thimonier, H., Costa, J. L. D. M., Popineau, F., Rimmel, A., & Doan, B.-L. (2025, May). T-JEPA: Augmentation-Free Self-Supervised Learning for Tabular Data. arXiv. https://doi.org/10.48550/arXiv.2410.05016

Vodolazskiy, E. (2021). Discrete Frechet distance for closed curves. arXiv E-Prints, arXiv:2106.02871. https://doi.org/10.48550/arXiv.2106.02871

Vujinović, A., & Kovačević, A. (n.d.). ACT-JEPA: Novel Joint-Embedding Predictive Architecture for Efficient Policy Representation Learning.

Wang, A., Datta, T., & Dickerson, J. P. (2024). Strategies for Increasing Corporate Responsible AI Prioritization. arXiv E-Prints, arXiv:2405.03855. https://doi.org/10.48550/arXiv.2405.03855

Wang, H., Cao, J., Liu, J., Zhou, X., Huang, H., & He, R. (2024). Hallo3d: Multi-modal hallucination detection and mitigation for consistent 3d content generation. Advances in Neural Information Processing Systems, 37, 118883–118906.

Wang, Y., He, J., Fan, L., Li, H., Chen, Y., & Zhang, Z. (2023). Driving into the Future: Multiview Visual Forecasting and Planning with World Model for Autonomous Driving. arXiv E-Prints, arXiv:2311.17918. https://doi.org/10.48550/arXiv.2311.17918

Wang, Z., Bingham, G., Yu, A. W., Le, Q. V., Luong, T., & Ghiasi, G. (2024). Haloquest: A visual hallucination dataset for advancing multimodal reasoning. European Conference on Computer Vision, 288–304.

Wei, J., Yao, Y., Ton, J.-F., Guo, H., Estornell, A., & Liu, Y. (2024). Measuring and reducing llm hallucination without gold-standard answers. arXiv Preprint arXiv:2402.10412.

Weimann, K., & Conrad, T. O. F. (2024, October). Self-Supervised Pre-Training with Joint-Embedding Predictive Architecture Boosts ECG Classification Performance. arXiv. https://doi.org/10.48550/arXiv.2410.13867

Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is Inevitable: An Innate Limitation of Large Language Models. arXiv E-Prints, arXiv:2401.11817. https://doi.org/10.48550/arXiv.2401.11817

Zhang, B., & Sennrich, R. (2019). Root Mean Square Layer Normalization. arXiv E-Prints, arXiv:1910.07467. https://doi.org/10.48550/arXiv.1910.07467

Zhen, H., Qiu, X., Chen, P., Yang, J., Yan, X., Du, Y., Hong, Y., & Gan, C. (2024). 3D-VLA: A 3D Vision-Language-Action Generative World Model. arXiv E-Prints, arXiv:2403.09631. https://doi.org/10.48550/arXiv.2403.09631

Zheng, W., Chen, W., Huang, Y., Zhang, B., Duan, Y., & Lu, J. (2023). OccWorld: Learning a 3D Occupancy World Model for Autonomous Driving. https://arxiv.org/abs/2311.16038

Zhu, H., Dong, Z., Topollai, K., & Choromanska, A. (2025, January). AD-L-JEPA: Self-Supervised Spatial World Models with Joint Embedding Predictive Architecture for Autonomous Driving with LiDAR Data. arXiv. https://doi.org/10.48550/arXiv.2501.04969

Zhu, H., Huang, B., Zhang, S., Jordan, M., Jiao, J., Tian, Y., & Russell, S. (2024). Towards a Theoretical Understanding of the "Reversal Curse" via Training Dynamics. arXiv E-Prints, arXiv:2405.04669. https://doi.org/10.48550/arXiv.2405.04669

Zürn, J., Gladkov, P., Dudas, S., Cotter, F., Toteva, S., Shotton, J., Simaiaki, V., & Mohan, N. (2024). WayveScenes101: A Dataset and Benchmark for Novel View Synthesis in Autonomous Driving. arXiv E-Prints, arXiv:2407.08280. https://doi.org/10.48550/arXiv.2407.08280

Zyrianov, V., Che, H., Liu, Z., & Wang, S. (2024). LidarDM: Generative LiDAR Simulation in a Generated World. https://arxiv.org/abs/2404.02903

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.