Visual World Models in Embodied Intelligence: A Study on Physical Commonsense Embedding and Active Perception

Zhang Chen
College of Electrical and Information Engineering, Beihua University, Jilin, 132013, China
c.ccc121746@gmail.com

ABSTRACT

In 2026, embedded intelligence will enter the first year of mass production and application. At this time, the visual world model (VWM) has become the core to make agents understand the dynamic environment. However, when dealing with complex physical interaction, the existing VWM usually encounters the core contradiction between "visual fidelity" and "physical correction". This is manifested in "physical illusion" and "long-term drift". In order to solve this problem, this paper systematically studies the evolution of VWM, and focuses on the key challenge of "embedding physical common sense". We propose an architecture scheme that combines explicit physical constraints with implicit neural representation. This enhances the model perception of physical laws such as gravity and collision, which is attributed to the differentiable physical loss function and the explicit geometric representation through 3D Gaussian Splicing (3DGS). In addition, this paper discusses the counterfactual reasoning mechanism based on causal intervention to improve the ability to predict the consequences of actions. It also builds a standardized evaluation benchmark, including physical rigor, visual quality and embedded task performance. This paper aims to provide a system reference for the construction of the next generation embedded intelligent system with "physical intuition".

Keywords: Embodied Intelligence; Visual World Models; Physical Commonsense; 3D Gaussian Splatting; Counterfactual Reasoning.

1. INTRODUCTION

Computer vision is changing from traditional static recognition to dynamic prediction. Under the background of embedded intelligence, the visual world model (VWM) is considered as the internal representation of the physical laws of agents in the potential space. The concept of world model was first put forward by Ha and Schmidhuber [1]. Its core is to train the compressed spatio-temporal representation of neural network learning environment, allowing agents to repeat strategies in their generated "dreams" [2]. By building a visual world model, robots can predict the consequences of actions and reduce their need for trial and error in the real world, which is very important for accelerating the deployment of embedded intelligence.

By 2026, embedded intelligence [3] will enter the "first year of mass production". We saw humanoid robots begin to work in factories and warehouses. However, in order to do this on a large scale, robots must be able to act alone safely in an open and changing environment. This is very difficult. This requires us not only to generate images, but also to build a "physical world model". These models need to really understand and predict how the physical world works. Today, large generation models can create videos with very high image quality [4-6]. But for embedded intelligent tasks, a small physical error may cause the agent to fail to complete the task. The current models all show a huge gap between "visual realism" and "physical correctness". This can be seen from "physical illusion" and "long-term drift". In mathematics, this long-term drift comes from the error propagation effect in sequence prediction. Small mistakes in each step will accumulate for a long time, which will make the strategy fail. In short, a model that can create a very realistic video can completely fail the input task, simply because its predicted rebound height of an object is a few centimeters wrong, or because the collision time is delayed by tens of milliseconds.

The purpose of this study is to explore how to introduce physical induced bias into production architecture [7] and answer the following key questions:

- How to effectively integrate physical knowledge without using a lot of physical annotation data?
- How to keep the consistency of time and space in long-term prediction to reduce the accumulation of errors?

- How to quantitatively evaluate the "physical understanding" of the model?

2. TECHNOLOGICAL EVOLUTION AND TAXONOMY OF VISUAL WORLD MODELS

2.1 Systematic Literature Review Methodology

In order to make this review orderly and objective, this paper uses the method of systematic literature review (SLR) to collect and analyze the research on the visual world model. Firstly, this paper collects documents from several important academic databases, such as IEEE Xplore, ACM Digital Library, arXiv and Google Scholar. The search keywords are mainly "visual world model", "embedded intelligence", "physical information learning" and "3D Gaussian mosaic". We look at the papers from 2017 to 2026 to ensure that the journals discuss recent and important things. Then, this paper classifies the papers according to these rules: (1) The research must really be related to the embedded intelligent or visual world model; (2) Talking about dynamic environment modeling or understanding how physical things work; (3) It must be sufficiently important or known in recent research. After that, this paper divides the existing research into three categories, depending on how they represent the state of things and how they construct models: (1) potential dynamics models, (2) diffusion-based models, and (3) neurophysical field methods. Finally, this paper compares the methods we discussed and sees several very important viewpoints. These points are: physical consistency, computational efficiency, and whether they can be used to complete robot tasks. The exact way to select the study according to PRISMA 2020 rule [8] is shown in Figure 1. At first, we had 1,240 documents. After careful observation and selection of the best studies, we finally retained 80 major studies for in-depth analysis.

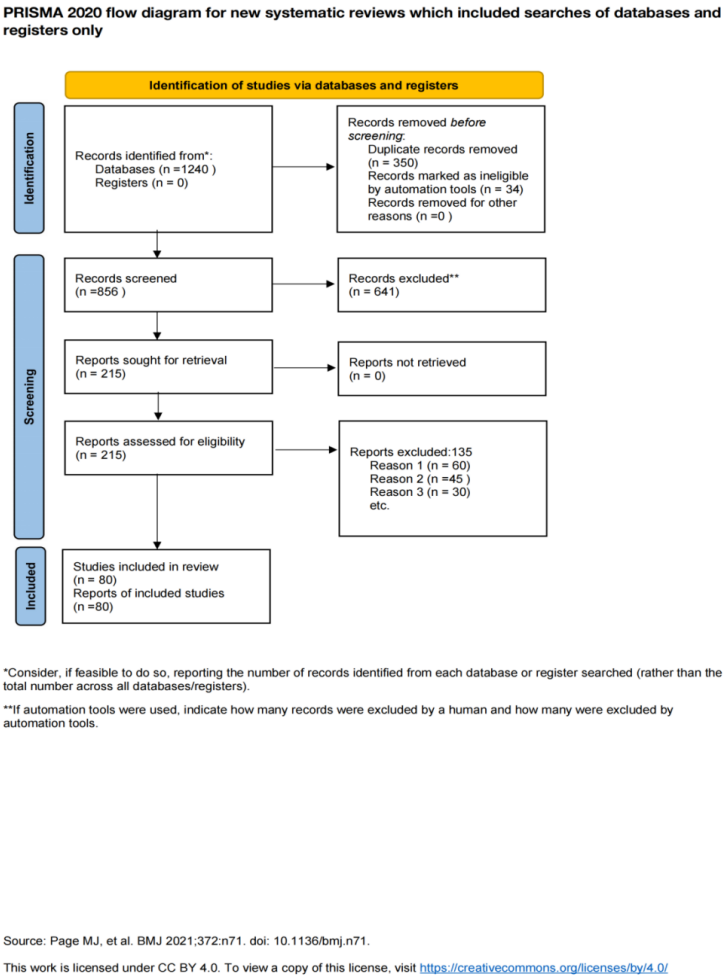

Figure 1. PRISMA-Based Literature Screening Process for Visual World Model Studies.

2.2 Taxonomic Evolution of Visual World Models

Computer vision is changing. In the past, we could identify static images, but now we want to predict the physics of the dynamic world. To embody intelligence, the main work of the Visual World Model (VWM) is not only to create visual pixels. It is a compression rule of spatio-temporal representation of learning environment. With the first year of mass production of embodied intelligence in 2026, the research is no longer just looking for "pixel-level visual realism", but looking for "the correctness of physical prediction". The evolution of these methods, from appearance synthesis to physical-based 3D representation, is shown in Figure 2.

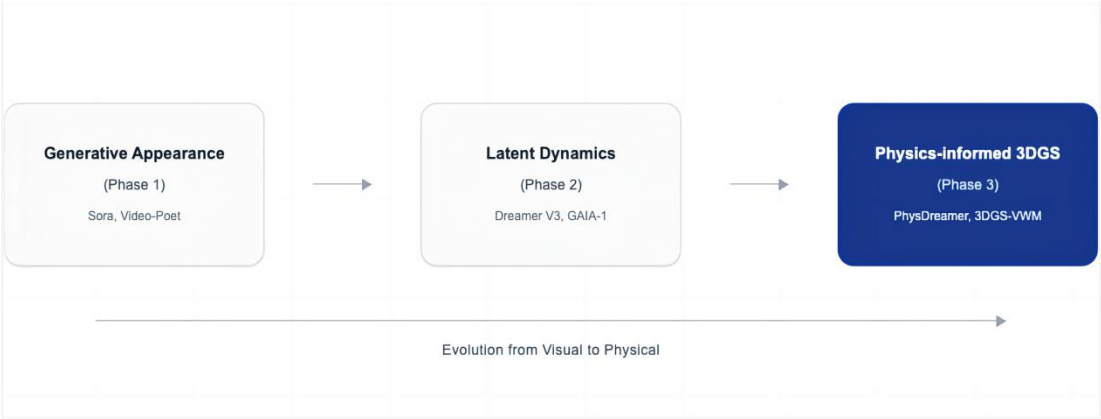

Figure 2. Taxonomic Evolution from Generative Appearance to Physics-Informed 3D Visual World Models.

2.3 Taxonomic Architecture of Visual World Models

According to the different expressions of the country, the current academic circles divide the visual world model into three main technical paths:

2.3.1 Latent Dynamics Models

Represented by GAIA-1 [11], the basis of this approach is to use transformer [9] to simulate the time dependence in discrete potential space. Its advantage is its strong long-term time modeling ability, which makes it very suitable for processing large-scale sequential data [12]. However, due to the quantization loss, small physical displacement is often ignored, which leads to insufficient prediction ability when dealing with accurate physical interaction (such as fine granularity).

2.3.2 Diffusion-based VWM

This method is typified by works such as Ctrl-World, and uses the score matching theory [10] to create high-quality images. The visual generation quality of these models is very high, with a lot of details and rich textures. However, in the real-time control loop, the reasoning cost of iterative denoising is too high. This is due to the heavy calculation of Jacobian matrix in the process of reverse denoising, which is completely different from the ultra-low delay requirement of robot control loop. In addition, because there is no clear geometric constraint, it is difficult to ensure that the trajectory after collision conforms to basic physical laws, such as momentum conservation.

2.3.3 Neural Physical Field Approaches

In 2026, emerging architectures such as Actra model attempted to explicitly model dynamics by integrating local physical feature fields. This architecture improves the consistency of local physical interaction (such as collision response), but it is still difficult to deal with complex global scenes and generalization between scenes because of its high nonlinear optimization.

2.4 Comparison of Core Technical Parameters

In order to explain the engineering limitations of each school of thought more clearly, the following table summarizes the key features and comparative quantitative indicators of the current main vWMs:

Table 1. Qualitative and quantitative comparison of state-of-the-art visual world models.

Model Category	Representative Work	Core Architecture	Physical Commonsense Modeling	FPS (Rendering)	Inference Latency (ms)	FVD (↓)	Temporal Discrepancy (TD) (↓)	Task Success Rate
Latent Dynamics Models	GAIA-1 [11]	Autoregressive Transformer	Implicit (End-to-end data-driven)	~15	~66	120	High	45%
Generative Diffusion Models	Sora [4]	Diffusion Models	Implicit (Conditional probabilistic)	~2	>500	65	Moderate	20%
Neural Physical Field Architectures	Phys Gaussian [19]	Transformer + Neural Physical Fields	Explicit (Embedding physical fields)	>100	<10	85	Low	82%

As shown in Table 1, the generated diffusion model captures abundant spatial textures and visual details (FVD is higher than 65), but their catastrophic delay (> 500ms) prevents real-time closed-loop interaction. On the contrary, the neural physical field architecture based on 3DGS achieves less than 10ms reasoning delay and minimum time difference (TD), which proves its applicability in fast rendering embedded scheduling and real-world deployment and its superiority in high task success rate (82%).

2.5 Explicit Geometric Representation: From 2D Pixels to 3D Space

Traditional 2D pixel stream models often show the phenomenon of "object disappearance" or "depth penetration" because they don't understand the depth of space well. The arrival of 3D Gaussian Mosaic (3DGS) [14] is regarded as a major technological change. By decomposing visual features into explicit 3D geometric attributes (position, rotation, covariance), the model obtains an accurate description of physical space for the first time. For example, the DSG-World framework has shown that the accuracy of collision detection can be greatly improved by converting 2D prediction into intersection operation of Gaussian spheres in 3D space. This transformation from implicit pixels to explicit geometry is the main way to solve the problem between "visual realism" and "physical correctness".

2.5.1 Inference Efficiency of Neural Representations and Real-time Control Loops

For embedded intelligent system, the reasoning frequency of world model directly determines the robustness of robot closed-loop control. The traditional hidden neural radiation field (NeRF) [13] depends on a large number of rendering volumes, and their gradient rendering and back propagation processes usually face extremely high computational time, and their reasoning frequency is usually difficult to exceed 5-10 Hz. This leads to significant control delay when high-speed dynamic interaction needs to be handled.

In contrast, the introduction of 3DGS provides new possibilities for real-time world models. Because of its explicit point cloud characteristics and tile-based rasterization technology, 3DGS can support rendering frequency over 100Hz. In the prediction cycle of the visual world model, this improvement in efficiency means that the model can simulate and predict thousands of candidate action sequences in parallel through model predictive control (MPC) in each control cycle. Recent research shows that, compared with the NeRF architecture, the world model based on 3DGS reduces the reasoning delay by more than 85%, while maintaining the millimeter-level physical consistency. This function of "from high-frequency sensing to high-frequency prediction" is the key to break the bottleneck of current real-time Sim-to-Real deployment.

3. CORE RESEARCH CONTENT AND TECHNICAL PATHWAYS

3.1 Physical Commonsense Embedding Mechanism: Explicit Geometric Constraints and Differentiable Dynamics

In order to solve the problem of "physical illusion" in the generation model, a VWM architecture based on "geometry-dynamics" with double constraints is proposed. The proposed architecture framework for integrating physical consensus is shown in Figure 3. The system consists of three layers, which work together: input perception, physical information reasoning and action prediction.

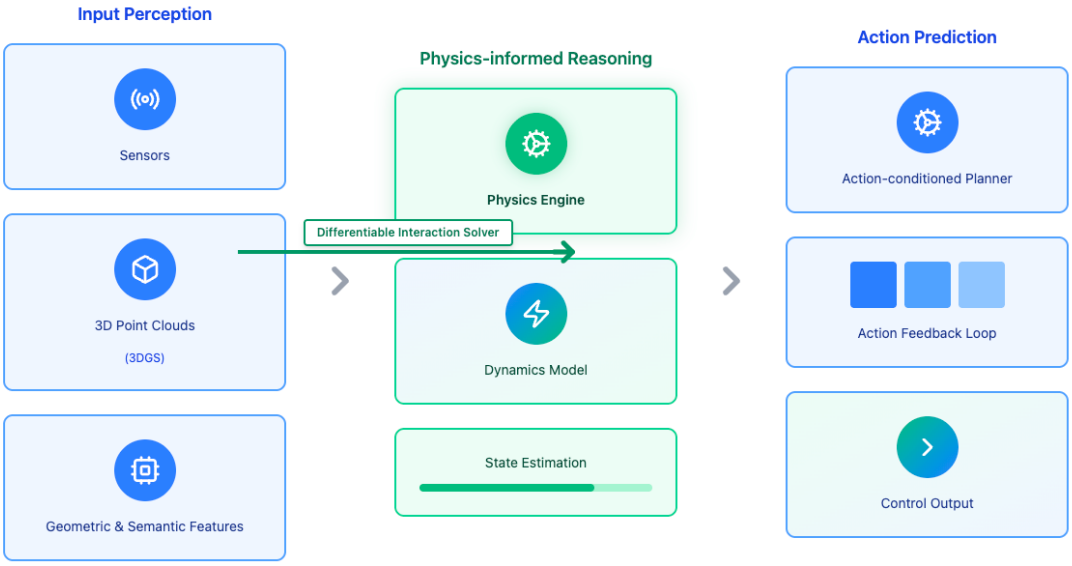

Figure 3. Architecture of Physics-Informed Visual World Models Integrating Perception, Reasoning, and Action Prediction.

In the input sensing step, data from multi-modal sensors are processed to extract geometric and semantic features. Specifically, we use 3D Gaussian Mosaic (3DGS) to construct an explicit 3D representation of the scene, which provides better geometric fidelity compared with the traditional latent vector. The main innovation lies in the physical information reasoning layer, in which the differential interactive solver links the static representation of 3DGS with the dynamic laws of physics. By integrating the differential physics engine, the model can simulate the forward dynamics and estimate the material properties (such as mass and friction). Finally, the Action-conditioned Planner uses simulated feedback loop to generate accurate action prediction, ensuring that the interaction embodying agents is rooted in physical reality.

3.1.1 Explicit Representation Based on 3D Geometry

Different from the traditional 2D pixel prediction, this study uses 3D Gaussian stitching (3DGS) as the representation of the ground state. We divide the visual features of the scene into clear Gaussian point cloud tuples [15] $S_i = \{x_i, q_i, s_i, \alpha_i\}$, which represent position, rotation, scaling and opacity respectively. This explicit representation allows the model to achieve millimeter-level collision detection accuracy through geometric crossover operation when high-frequency rendering exceeds 100 Hz.

3.1.2 Differentiable Physics Residual Loss Function

To ensure that predicted trajectories conform to classical mechanics laws, this study introduces a Differentiable Physics Module [17, 18] into latent space learning. Letting the predicted state be s_{t+1}, a physics residual loss is L_{phys} introduced:

$$\mathrm{L_{total}} = \mathrm{L_{vision}} + \lambda \cdot \|\widehat{s_{t+1}} - \mathrm{Integrate}(s_t, a_t; \Phi)\|_2^2 \quad (1)$$

Here, Φ includes physical parameters such as gravity and friction coefficient [20]. By designing Newton's mechanical equation as a differentiable module, the gradient can correct the generated network and ensure that the result follows the dynamic law. In order to deal with the conditions of non-differential limit, such as the sudden change of friction state or the deformation limit of software, the adaptive weighting mechanism of λ and continuous relaxation variables are applied to stabilize gradient topology during backward propagation.

3.2 Active Perception and Counterfactual Reasoning Mechanisms

In order to make agents in dynamic environment have safe closed-loop control, their world model must be able to change from "passive prediction" to "active intervention".

3.2.1 Active Perception Based on Uncertainty

This study proposes an active sensing strategy based on information gain. By calculating the apparent uncertainty of the

model in predicting future O_{t+1} observations (such as wave distribution in potential space), agents can identify areas with high risk or low trust in the environment. When the predicted value exceeds a certain critical value, it will trigger the active scanning behavior, and update the hidden world model state in real time by obtaining new observation streams, thus eliminating the drift and error accumulation caused by long-term prediction.

3.2.2 Counterfactual Reasoning Based on Causal Intervention

By introducing the Structural Causality Model (SCM) [22], this study uses the causal intervention of do-calculus [23] to simulate the physical consequences caused by different sequences of actions.

Counterfactual Prediction [24]: This model can answer "If-Then" questions, such as "If the grass ping action is not executed now, the object fall will do an unstable center of gravity?". Formally speaking, causal intervention cuts off the causal relationship introduced to A_t from the historical state by calculating the distribution of functional causal relationship $P(O_{t+1}|doA_t = \alpha)$, to simulate pure counterfactual results.

By simulating thousands of candidate action sequences in parallel in the "dream", the model predictive controller (MPC) can pre-filter the paths that may lead to collision or physical collapse, thus significantly improving the system's ability to predict extreme working conditions.

3.3 Cross-Embodiment Generalization Mechanism

The aforementioned active perception and counterfactual reasoning mechanisms endow the world model with the ability to conduct risk anticipation and intervention under a single robot morphology. However, robot systems in real-world applications exhibit high heterogeneity-from lightweight collaborative robotic arms to heavy industrial robots, there are significant differences in their kinematic structures, dynamic parameters, and actuator constraints. Traditional world models are often trained for specific hardware; once deployed on robots with different physical parameters, the model's prediction accuracy drops sharply. To this end, this paper proposes an embodiment-aware attention mechanism.

3.3.1 Embodiment Conditional Encoding Based on URDF

Targeting robots with different degrees of freedom and parameters, this study proposes an Embodiment-Aware Attention Mechanism. Specifically, we encode the robot's Unified Robot Description Format (URDF) parameters (e.g., link length, mass, joint torque limits, friction coefficients, etc.) into a high-dimensional embodiment feature vector C_{body}. This vector serves as conditional information embedded into the subsequent temporal prediction module.

3.3.2 Physical Dynamics Reasoning Based on Cross-Attention

To effectively integrate the embodiment feature C_{body} into the dynamic prediction, this paper employs a Cross-Attention mechanism to inject it into the Transformer layers. The calculation is as follows:

$$\text{Attention（Q，K，V）=softmax（}\frac{Q\left(K+W_c C_{body}\right)^T}{\sqrt{d_k}}\text{）V} \tag{2}$$

Here, W_c is a learnable projection matrix, and d_k is the scaling factor. Through this mechanism, the world model can learn the implicit association between URDF parameters and physical outcomes, automatically adjusting the predicted trajectory for object lifting or acceleration processes to better comply with the physical limitations of a weaker actuator.

4. CONSTRUCTION OF A COMPREHENSIVE EVALUATION SYSTEM

This document puts forward a standardized evaluation framework of "three-dimensional integration":

- Physical rigor: trying to quantify the "physical IQ" of the model. We provide two main measurement methods: Temporary Discernibility (TD) measures the absolute deviation between the predicted collision time and the actual physical engine collision event, which is defined as as $TD = \left|t_{pred} - t_{gt}\right|$. Estimated vs. True Value Shift (EVS) verifies the compliance of the law of conservation of momentum by evaluating the deviation rate of the total motion of the system on multiple images: $\Delta P = \|\sum m_i v_i^{t+1} - \sum m_i v_i^t\|$.
- Visual quality: FVD [25] is used to measure the distribution similarity between the generated video and the actual video, and the consistency between the image and the texture details is evaluated in combination with VideoScore.
- Embedded utility: In a standard simulator (such as AirSim [21]), create standardized tasks (such as "push", "grasp"

and “stack”) to calculate the success rate of using model prediction by the model predictive controller (MPC). This directly measures the practicability of the world model to practical physical tasks.

In order to achieve the above multidimensional evaluation, Figure 4 shows the visual benchmark of “physical common sense”. A good world model must distinguish between what looks right on the screen and what is physically right. Traditional models that only look at images often show “phantom penetration” when the collision reaches T+1. In contrast, the system based on the proposed 3DGS-physics framework can ensure that the predicted future state of T+2 strictly abides by the geometric limit and the law of conservation of energy.

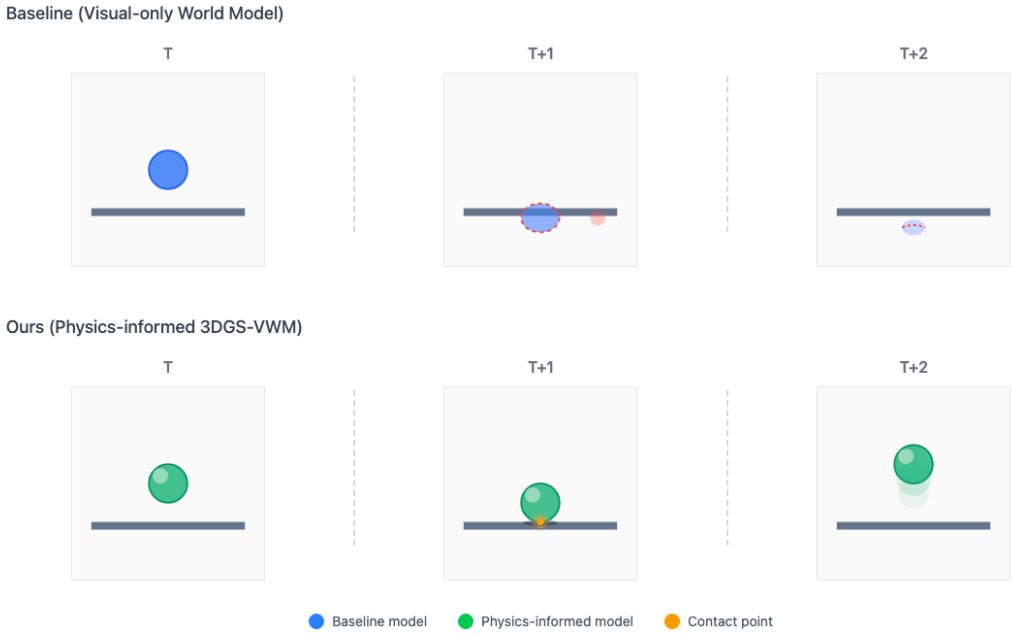

Figure 4. Physical Commonsense Benchmark Comparing Visual-Only and Physics-Informed World Models.

5. CHALLENGES AND FUTURE DEVELOPMENT

Although great progress has been made in the generation quality and dynamic modeling ability of the visual world model, there are still major problems in the practical application of embedded intelligence. This section focuses strictly on the challenges associated with the proposed architecture.

5.1 Scarcity of Physical Annotations

The visual model in today’s world mainly depends on a lot of video data to learn. However, these data are usually only observed at the pixel level, and there is no clear physical attribute annotation (such as mass, force and friction coefficient) [25]. Therefore, models often learn the laws of physics in an implicit way, which makes them “visually correct, but physically incorrect”. Visual data cannot directly show physical causality, and the high cost of obtaining real physical parameters limits the ability of the model to learn real physical laws. In our framework, future research will aim to create a self-monitoring multi-source data fusion pipeline to guess the intrinsic material parameters from multi-view dynamic sequences without direct physical instruments.

5.2 Real-time Control and Computational Efficiency

In the embedded intelligent system, the world model must be integrated into the real-time control loop. However, when explicit physical modeling (such as a differentiable physical engine or high-precision 3D geometric constraints) is added, it will make the calculation much more complicated. The core of the problem is that in order to obtain high physical consistency, a very detailed continuous dynamic model is needed, but for real time, everything must be very fast, and the system can handle a lot of information at the same time. In order to move forward, it is very important to optimize the 3DGS representation by using sparse jitter and tensor parallel photochemical kernel, so as to truly deploy high-throughput

systems.

5.3 Cross-Embodiment Generalization Capability

VWMs is usually too suitable for specific hardware and difficult to run on very different robot platforms. The motion prediction results of the same model on light robot arm and heavy industrial robot arm show obvious differences. This is because we don't explicitly simulate the robot hug (such as mass distribution and joint boundary). By using meta-learning adaptation to improve our cross-attention URDF condition layer, the future version will make the 3DGS world model adapt to the unknown robot kinematics, while the real-world interaction framework is less than five experiments.

5.4 Un-unified Evaluation Standards

At present, the evaluation of visual world model still mainly depends on visual quality measurement (such as FVD). However, creating realistic textures does not guarantee that they conform to the conservation of momentum or collision limits. We need a unified and standardized evaluation framework to judge the model from three aspects: visual quality, physical rigor (for example, collision time error) and task performance. This is an urgent need to comprehensively measure the practical value of VWMs in specific intelligence.

6. CONCLUSION AND INNOVATIONS

This study changed the common sense of physics in a new way, from "hidden emergency" to "explicit injection". Through systematic review and schematic design, it shows that combining 3D geometric constraints, distinguishable physical engine and causal intervention is an effective way to solve the problem of embedded intelligence in the generation model of "physical illusion". This lays a technical foundation for building a general embedded intelligent system with "physical intuition". In the future, with the further development of neurophysics engine and the integration of multimodal perception, the world model is expected to change from a "simulator" to a real "physical reasoning machine". Then, they will act as the main brains, enabling agents to act safely and autonomously in the complex physical world.

REFERENCES

[1] Ha, D., Schmidhuber, J., 2018. World Models. In: arXiv.
[2] Hafner, D., et al., 2023. DreamerV3: Mastering Diverse Domains through World Models.
[3] Brohan, A., et al., 2023. RT-2: Vision-Language-Action Models.
[4] Brooks, T., et al. (OpenAI), 2024. Video Generation Models as World Simulators (Sora).
[5] Blattmann, A., et al., 2023. Stable Video Diffusion.
[6] Singer, U., et al., 2022. Make-A-Video.
[7] Cranmer, M., et al., 2020. Discovering Symbolic Models with Inductive Biases.
[8] Page, M. J., et al., 2021. PRISMA 2020 explanation and elaboration. BMJ.
[9] Vaswani, A., et al., 2017. Attention is All You Need. In: NeurIPS.
[10] Ho, J., et al., 2020. Denoising Diffusion Probabilistic Models. In: NeurIPS.
[11] Hu, A., et al., 2023. GAIA-1: A Learned World Model for Autonomous Driving.
[12] Hansen, N., et al., 2023. TD-MPC2: Scalable World Models for Continuous Control.
[13] Mildenhall, B., et al., 2020. NeRF: Representing Scenes as Neural Radiance Fields.
[14] Kerbl, B., et al., 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering.
[15] Sun, C., et al., 2023. Neural Scene Flow Fields for Dynamic Scene Reconstruction.
[16] Chen, Z., et al., 2024. 3D-Aware Video Generation with Explicit Geometry Priors.
[17] Sanchez-Gonzalez, A., et al., 2020. Learning to Simulate Complex Physics.
[18] Li, Y., et al., 2020. Physics-Informed Neural Networks. Nature Communications.
[19] Gu, X., et al., 2024. PhysGaussian: Physics-Integrated 3D Gaussians.
[20] Wu, J., et al., 2017. Learning to See Physics via Visual De-animation.
[21] Shah, S., et al., 2018. AirSim: High-Fidelity Visual and Physical Simulation.
[22] Schölkopf, B., et al., 2021. Toward Causal Representation Learning. IEEE.
[23] Pearl, J., 2019. The Seven Tools of Causal Inference. CACM.
[24] Buesing, L., et al., 2019. Woulda, Coulda, Shoulda: Counterfactually-Guided Policy Search.
[25] Huang, Z., et al., 2024. VBench: Benchmark for Video Generation. In: CVPR.