NeuIDO: Neural Intrinsic Dynamics Operator for Physics-Informed 4D World Models

Jiajing Lin[1], Xin Zhang[2], and Jianhua Sun*[1]

[1] School of Artificial Intelligence, Shanghai Jiao Tong University, China
[2] School of Informatics, Xiamen University, China

Abstract. World models aim to capture environmental dynamics and predict future trajectories, showing growing potential for embodied intelligence. Physics-informed 4D generation integrates physical simulation to predict 3D object interactions, offering a promising pathway toward world models. However, this paradigm relies on manually imposed dynamical assumptions rather than internalizing world dynamics, and thus still leaves a gap toward a true world model. To bridge this gap, we propose NeuIDO, a novel world dynamics modeling framework that learns a unified intrinsic dynamics representation from visual observations, advancing physics-informed 4D generation toward a world model. Specifically, we formulate world modeling as a neural operator learning problem and introduce a two-stage training strategy to learn a generalizable mapping from the visual observation distribution to the intrinsic dynamics distribution. Building on this observation-dynamics mapping, NeuIDO enables zero-shot dynamics inference directly from videos and can be further aligned with complex real-world dynamics via few-shot adaptation. Extensive experiments demonstrate that NeuIDO effectively unifies the intrinsic dynamics underlying diverse visual observations into a shared representation and rapidly infers dynamics in novel scenes.

Keywords: World Model · 4D Generation · Neural Operator

1 Introduction

World models [3,11,24] provide agents with an "imagination space" for planning and decision-making, and have shown great potential in applications such as autonomous driving [16,40] and robotics [9,44]. Most current world models [2,12,14] learn environment dynamics from large-scale 2D data (e.g., images or videos) and predict future states via 2D frame generation. Despite progress under this purely data-driven paradigm, the absence of explicit physical constraints and limited 3D spatial reasoning (e.g., inaccurate object layouts and relative relations) often leads to predictions that violate fundamental physical principles [20] (e.g. conservation laws), limiting reliability for embodied decision-making.

* Corresponding author: Jianhua Sun (gothic@sjtu.edu.cn).

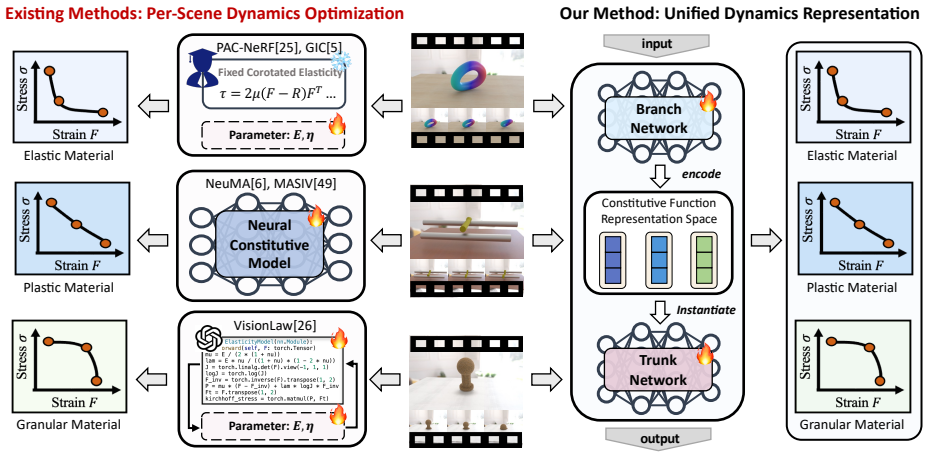

Fig. 1: Existing approaches typically perform per-scene dynamics optimization, which is computationally expensive and generalizes poorly to unseen scenes. In contrast, *NeuIDO* learns a unified intrinsic dynamics representation that enables efficient zero-shot generalization, moving physics-informed 4D generation toward world modeling.

In contrast, physics-informed 4D generation [29, 50] explicitly incorporates physical simulators to evolve 3D scenes over time. Such a paradigm enables physically consistent movement prediction and provides an explicit 3D spatiotemporal representation, thus offering a promising avenue to mitigate the limitations of purely data-driven world models. However, physical simulation relies on access to the system's intrinsic dynamics, whose core lies in the constitutive law [7] with specified material parameters that govern the object's response under applied forces. These laws are typically imposed through hand-crafted assumptions.

Humans can intuitively infer an object's intrinsic dynamics through vision and predict its future evolution. To achieve such intuitive physical inference, recent methods have attempted to integrate differentiable renderers, such as Neural Radiance Fields (NeRF) and 3D Gaussian Splatting (3DGS) [21, 37], with differentiable simulators [41] in an end-to-end paradigm. Specifically, some methods [5, 25, 52] estimate unknown material parameters through visual supervision. However, they rely on hand-crafted constitutive laws, making it difficult to align with the complex dynamics of the real world. Furthermore, another class of methods [6, 27, 55] attempts to learn neural or symbolic constitutive models from visual observations. However, these approaches independently infer object intrinsic dynamics from scratch for each scene, preventing the learning of a shared cross-scene dynamics structure and limiting efficient generalization to unseen scenes. As a result, they remain short of a true world model.

To overcome the above challenges, we propose *NeuIDO*, a novel world dynamics modeling framework that learns a unified intrinsic dynamics representation from multi-scene video observations and supports efficient intrinsic dynamics inference for unseen scenes—thereby advancing physics-informed 4D generation

toward a world model. Inspired by the Universal Approximation Theorem (UAT) for neural operators—which states that neural operators can approximate continuous mappings between infinite-dimensional function spaces—we cast world dynamics modeling as an operator learning problem. Specifically, we learn a mapping from an input (e.g., videos) function space to an output (e.g., constitutive laws) function space, enabling diverse intrinsic dynamics across scenes to be modeled within a unified framework. To reduce the complexity of operator learning, we further propose a two-stage training strategy: (i) pretraining the structural representation space of the constitutive function from large-scale known laws to inject dynamical priors; and (ii) aligning videos with this learned representation space so that each observation is interpreted as a manifestation of intrinsic dynamics. Leveraging this observation–dynamics operator mapping, *NeuIDO* enables real-time zero-shot intrinsic dynamics inference from videos. Furthermore, to better capture complex real-world dynamics, it supports efficient few-shot adaptation via lightweight fine-tuning on visual observations. Our contributions are summarized as follows:

- We propose a world dynamics modeling framework that learns a unified intrinsic dynamics representation from visual observations across multiple scenes, advancing physics-informed 4D generation toward a true world model.
- We formulate world dynamics modeling as an operator learning problem and address optimization challenges with a two-stage strategy: (i) structured dynamics representation learning and (ii) observation-to-dynamics alignment.
- Extensive experiments on synthetic and real-world datasets demonstrate that *NeuIDO* effectively models shared cross-scene dynamics structure and achieves real-time zero-shot generalization with efficient few-shot adaptation.

2 Related Work

2.1 Physics-Informed 4D Interaction

In recent years, 3D representations [21, 37, 49] have increasingly been coupled with physics simulators [35,38,41] to enable interactive 4D dynamics [26,28,29]. PIE-NeRF [10] couples NeRF with a mesh-free elastodynamics simulator, PhysGaussian [50] combines 3DGS with an MPM simulator [41], and VR-GS [19] enables real-time VR interaction through fast XPBD-based simulation [35]. However, these approaches rely on expert-specified constitutive laws and carefully tuned parameters. PAC-NeRF [25] and GIC [5] perform differentiable simulation-based identification from multi-view videos. Another line of work, such as PhysDreamer, DreamPhysics, Physics3D, and PhysFlow [18, 31, 32, 52], distills dynamic priors from video diffusion models to guide estimation. Along this direction, OmniPhysGS [30] further assigns each Gaussian kernel a constitutive model from a predefined set. Nevertheless, these methods still rely on expert-defined constitutive laws, limiting their ability to capture complex real-world dynamics. To reduce this dependency, NeuMA [6] and MASIV [55] optimize neural constitutive models to align simulations with visual observations, though the

learned models are implicit and lack interpretability. VisionLaw [27] further advances this direction by discovering explicit constitutive-law expressions through bilevel optimization. Despite these advances, existing methods typically learn a separate intrinsic dynamics for each scene.

2.2 Data-Driven World Model

World models [11, 13, 48] aim to understand and predict environment dynamics, enabling applications such as autonomous driving [46, 54], robotics [42, 43, 47, 51, 57], and VR/AR [58]. Existing approaches broadly fall into three families: SSM-based, JEPA-based, and Transformer-based methods. World Models [11] learns compact latent representations and RNN-based latent dynamics from high-dimensional observations, serving as an early precursor to later RSSM-based world models. Dreamer family [12, 14, 15] extends this paradigm to model-based reinforcement learning via latent imagination. JEPA [24] learns predictive world representations by predicting embeddings of masked regions rather than reconstructing pixels. I-JEPA [1] and V-JEPA [2] apply this idea to images and videos, respectively, through context-to-masked-block prediction without hand-crafted augmentations. Transformer-based world models [8, 36, 39] replace recurrent dynamics with attention-based sequence modeling [45] to better capture long-horizon dependencies. More recently, large-scale generative world simulators such as Sora [3] and Genie [4] learn controllable environment dynamics from video, enabling long-horizon generation, while 3D-aware generative world models [22] (e.g., Genie 2, Explorer, LWMs) further couple generative modeling with neural scene representations to construct consistent 3D worlds. Despite their effectiveness, existing world models largely follow a purely data-driven paradigm and lack explicit physical constraints, which often leads to physically inconsistent predictions. In this work, we instead explicitly model world dynamics, moving toward a physics-informed 4D world model.

3 Methodology

3.1 Problem Setup and Overview

In this paper, we aim to learn a unified representation that captures the diverse intrinsic dynamics underlying visual observations across multiple scenes, while enabling efficient dynamics inference in novel scenarios. Formally, let $K_V \subset \mathbb{R}^{d_v}$ denote the domain of video functions, and let $\mathcal{C}(K_V)$ be the corresponding space of continuous video functions. In practice, we consider a compact subset $\mathcal{V} \subset \mathcal{C}(K_V)$ representing the set of valid video observations that sufficiently capture object motions governed by scene-specific dynamics. We consider a set of N dynamic scenes, each providing calibrated multi-view videos $\mathbf{u} \in \mathcal{V}$. Let $\mathcal{C}(K_C)$ be the constitutive function space defined on the constitutive-input domain $K_C \subset \mathbb{R}^{d_c}$. Our goal is to learn an operator:

$$\widehat{\mathcal{G}} : \mathcal{V} \to \mathcal{C}(K_C) \tag{1}$$

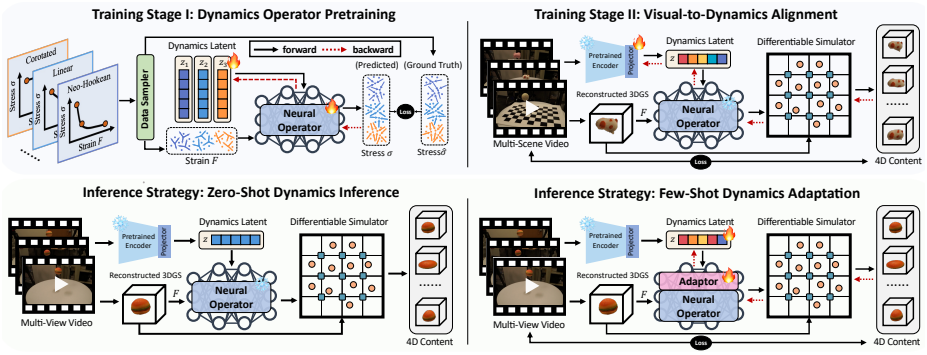

Fig. 2: Pipeline of NeuIDO. In Stage I, we pretrain the operator on a collection of known constitutive laws to learn a structured representation space of constitutive functions, injecting rich dynamics priors. In Stage II, we train a projector to align visual observations with this representation space, establishing a connection between videos and intrinsic dynamics. Under the zero-shot paradigm, the operator immediately infers plausible intrinsic dynamics for unseen scenes. Under the few-shot paradigm, lightweight adaptation further aligns the operator to out-of-distribution dynamics.

which maps the video function space to the constitutive function space, such that diverse scene dynamics are embedded into a unified representation. In Sec. 3.2, we formulate a neural intrinsic dynamics operator. In Sec. 3.3, we propose a two-stage training strategy to learn the unified dynamics representation. In Sec. 3.4, we present zero-shot and few-shot inference strategies for generalization to novel scenes. In addition, we provide a theoretical analysis in the Appendix to support the feasibility of our method for modeling world dynamics. The overall architecture of *NeuIDO* is illustrated in Fig. 2.

3.2 Unified Intrinsic Dynamics Representation Framework

Intrinsic Dynamics Operator Parameterization. We parameterize the mapping $\widehat{\mathcal{G}}$ from observation to dynamics using a neural intrinsic dynamics operator. Formally, $\widehat{\mathcal{G}}$ acts on an input video function $\mathbf{u}$ to output a continuous constitutive function, $\widehat{\mathcal{G}}(\mathbf{u})$. This function can then be evaluated at any given constitutive input $\boldsymbol{\xi} \in K_C$ (e.g., deformation gradient $\mathbf{F}$) to yield the specific physical response, $\widehat{\mathcal{G}}(\mathbf{u})(\boldsymbol{\xi})$. Specifically, inspired by DeepONet [33], *NeuIDO* is formulated as a branch–trunk operator architecture. The branch network uses a dynamics encoder $\mathbf{E}$ to extract $\mathbf{z} \in \mathbb{R}^d$ from the input video, such that $\mathbf{z} = \mathbf{E}(\mathbf{u})$, capturing the intrinsic dynamics governing the observed motion. This latent representation is then mapped to the coefficient vector of the operator expansion through a multilayer perceptron (MLP), producing $\boldsymbol{\lambda} \in \mathbb{R}^D$. The trunk network parameterizes shared basis functions $\{\phi_k(\cdot)\}_{k=1}^D$, and evaluates them at $\boldsymbol{\xi}$ to output the vector $\{\phi_k(\boldsymbol{\xi})\}_{k=1}^D$. The video-dependent constitutive function $\widehat{\mathcal{G}}(\mathbf{u})$ is then represented as a linear combination of these basis functions weighted by

the coefficients.

$$\widehat{\mathcal{G}}(\mathbf{u})(\boldsymbol{\xi}) = \sum_{k=1}^{D} \lambda_k \, \phi_k(\boldsymbol{\xi}). \tag{2}$$

Under this formulation, the trunk network learns a shared basis spanning the constitutive function space, while the branch network predicts scene-specific coefficients that instantiate the constitutive relation for each observation. This design enables *NeuIDO* to effectively fit the mapping between visual observations and intrinsic dynamics, while exploiting the generalization of neural operators to enable efficient inference in unseen scenarios.

Trunk Network: Intrinsic Dynamics Basis Representation. The trunk network parameterizes a shared set of constitutive basis functions $\{\phi_k(\cdot)\}_{k=1}^{D}$ across different scenes. For a query point $\boldsymbol{\xi} \in K_C$, we evaluate these basis functions at ξ to obtain their basis values $\{\phi_k(\boldsymbol{\xi})\}_{k=1}^{D}$, which are then linearly combined with the branch-predicted coefficients in Eq. 2. In differentiable MPM, two types of constitutive relations must be defined: an elastic component that describes the reversible material response under deformation, and a plastic component that captures irreversible evolution beyond the elastic limit. Accordingly, we parameterize two trunk networks to match these two constitutive mechanisms.

(i) Elastic trunk network. The elastic trunk takes the deformation gradient $\mathbf{F}$ as input and outputs stress basis values $\{\boldsymbol{\phi}_k^E(\mathbf{F})\}_{k=1}^{D}$. Given on the visual observation $\mathbf{u}$, the elastic constitutive function $\widehat{\mathcal{G}}^E(\mathbf{u})$ is represented as:

$$\widehat{\mathcal{G}}^E(\mathbf{u})(\mathbf{F}) = \sum_{k=1}^{D} \lambda_k^E \, \phi_k^E(\mathbf{F}). \tag{3}$$

The coefficient vector $\boldsymbol{\lambda}^E$ is obtained by mapping the dynamics latent through the elastic MLP head of the branch network.

(ii) Plastic trunk network. Similarly, the plastic trunk takes $\mathbf{F}$ as input and outputs corrected deformation-gradient basis values $\{\phi_k^P(\mathbf{F})\}_{k=1}^{D}$. The scene-specific plastic constitutive function $\widehat{\mathcal{G}}^P(\mathbf{u})$ is parameterized in a residual form:

$$\widehat{\mathcal{G}}^P(\mathbf{u})(\mathbf{F}) = \mathbf{F} + \sum_{k=1}^{D} \lambda_k^P \, \phi_k^P(\mathbf{F}). \tag{4}$$

The coefficient vector $\boldsymbol{\lambda}^P$ is obtained by mapping the same dynamics latent through the plastic MLP head of the branch network.

Branch Network: Intrinsic Dynamics Latent Identification. The branch network aims to extract scene-dependent dynamics features from visual observations and transform them into coefficient vectors that, together with the trunk network, instantiate the constitutive operators. To this end, we introduce a dynamics encoder $\mathbf{E}$ that maps an input video $\mathbf{u} \in \mathbb{R}^{L \times H \times W \times 3}$ to a compact

dynamics latent $\mathbf{z} \in \mathbb{R}^d$, which captures the underlying dynamics governing the observed motion. Concretely, $\mathbf{E}$ consists of three stages: (i) we compute the optical flow $\mathbf{u}_f \in \mathbb{R}^{(L-1)\times H\times W\times 3}$ between consecutive frames to emphasize motion cues; (ii) we employ the pretrained optical flow encoder from Tora [53] to embed the flow fields into a spatiotemporal feature representation $\mathbf{z}^0 \in \mathbb{R}^{\frac{L-1}{4}\times\frac{H}{8}\times\frac{W}{8}\times 3}$, which captures motion patterns across frames; (iii) a projector network maps $\mathbf{z}^0$ into the compact dynamics latent $\mathbf{z}$. The dynamics latent $\mathbf{z}$ is then mapped by two MLP heads to produce elastic and plastic coefficient vectors, $\boldsymbol{\lambda}^E \in \mathbb{R}^D$ and $\boldsymbol{\lambda}^P \in \mathbb{R}^D$, respectively, which are used to weight the corresponding trunk-evaluated bases.

3.3 Unified Intrinsic Dynamics Representation Learning

Due to video supervision being sparse and indirect with respect to intrinsic dynamics, learning the observation-to-dynamics operator in an end-to-end manner is challenging. In contrast, numerous known constitutive relations provide strong supervision covering diverse dynamical mechanisms, offering powerful dynamics priors. To exploit this advantage, we adopt a two-stage training strategy. In Stage I, the operator is pretrained on known constitutive relations to learn a structured representation space of constitutive functions. In Stage II, video observations are aligned with this pretrained space, enabling the operator to infer intrinsic dynamics consistent with the observed motions. This two-stage design effectively transfers knowledge from strongly supervised constitutive data to weakly supervised visual observations, facilitating stable learning of the mapping from visual observations to intrinsic dynamics.

Stage I: Dynamics Operator Pretraining. This stage aims to learn a structured *representation space of constitutive functions* under strong physical supervision, so that diverse intrinsic dynamics can be expressed. This representation space is parameterized by a shared trunk basis and indexed by a learnable dynamics latent. To construct the training data, we collect a large set of known constitutive relations and use the MPM simulator to generate material trajectories for each constitutive function. For each trajectory i, we record deformation gradients and their corresponding constitutive responses over particles and timesteps. A total of M trajectories are collected to form the training dataset:

$$\mathcal{D} = \left\{ \left(\mathbf{F}_{p,t}^{(i)},\, \mathbf{S}_{p,t}^{(i)},\, \widehat{\mathbf{F}}_{p,t}^{(i)} \right) \right\}, \quad p = 1, \ldots, P,\ t = 1, \ldots, T, \tag{5}$$

where $\mathbf{F}$ is the deformation gradient, $\mathbf{S}$ is the elastic stress, and $\hat{\mathbf{F}}$ denotes the plasticity-induced corrected deformation gradient. As visual observations are not yet introduced in this stage, we bypass the dynamics encoder $\mathbf{E}$ and instead assign each trajectory a learnable dynamics latent $\mathbf{z}^{(i)}$. This latent serves as an index to instantiate a specific constitutive function within the shared trunk basis. Concretely, we map $\mathbf{z}^{(i)}$ to coefficient vectors $\boldsymbol{\lambda}^{E,(i)}$ and $\boldsymbol{\lambda}^{P,(i)}$ via lightweight elastic and plastic MLP heads, and evaluate the trunk networks on $\mathbf{F}$ to form

latent-conditioned constitutive operators. The operator parameters (excluding the dynamics encoder) and the dynamics latents are jointly optimized by minimizing reconstruction errors of elastic and plastic responses:

$$\min_{\theta^E,\theta^P,\{\mathbf{z}^{(i)}\}} \sum_{i=1}^{M} \sum_{p,t} \Big(\mathcal{L}_r\big(\widehat{\mathcal{G}}^E(\mathbf{z}^{(i)})(\mathbf{F}_{p,t}^{(i)}),\, \mathbf{S}_{p,t}^{(i)}\big) + \mathcal{L}_r\big(\widehat{\mathcal{G}}^P(\mathbf{z}^{(i)})(\mathbf{F}_{p,t}^{(i)}),\, \widehat{\mathbf{F}}_{p,t}^{(i)}\big) \Big), \quad (6)$$

where $\mathcal{L}_r$ denotes the relative loss [23], and θ^E and θ^P represent the parameters of the elastic and plastic trunk networks, respectively. After Stage I, the model captures shared structures across constitutive functions and learns a physically grounded latent space for material dynamics. This pretrained space serves as a strong prior for Stage II, where videos are aligned to it by learning the mapping from visual observations to dynamics latents via the branch video encoder $\mathbf{E}$.

Stage II: Visual-to-Dynamics Alignment. Stage II aligns visual observations with the constitutive representation space learned in Stage I, establishing a connection between videos and intrinsic dynamics. Specifically, we freeze the trunk networks and optimize only the branch-side dynamics encoder $\mathbf{E}$, which maps a video observation $\mathbf{u}$ to a dynamics latent $\mathbf{z} = \mathbf{E}(\mathbf{u})$. Given a training set of N dynamic scenes, where each scene provides multi-view videos capturing object motions governed by intrinsic dynamics, each scene i is treated as one training instance. For each scene, the predicted dynamics latent $\mathbf{z}^{(i)}$ instantiates elastic constitutive function $\widehat{\mathcal{G}}^E(\mathbf{z}^{(i)})$ and plastic constitutive function $\widehat{\mathcal{G}}^P(\mathbf{z}^{(i)})$ within the frozen operator space. These functions are then used to drive a differentiable physics-integrated 4D simulation pipeline $\mathcal{R}$ (see Appendix for details) to generate rendered videos $\widetilde{V}^{(i)} = \mathcal{R}(\widehat{\mathcal{G}}^E(\mathbf{z}^{(i)}), \widehat{\mathcal{G}}^P(\mathbf{z}^{(i)}))$. We supervise the dynamics encoder $\mathbf{E}$ by minimizing the reconstruction error between rendered and observed videos over all scenes:

$$\min_{\omega} \sum_{i=1}^{N} \mathcal{L}_{\text{mse}}\Big(\mathcal{R}\big(\widehat{\mathcal{G}}^E(\mathbf{z}^{(i)}), \widehat{\mathcal{G}}^P(\mathbf{z}^{(i)})\big), \mathbf{u}^{(i)}\Big), \quad (7)$$

where ω denotes the parameters of the dynamics encoder $\mathbf{E}$. By optimizing this objective over diverse scenes, the dynamics encoder learns a generalizable mapping from visual motion patterns to the intrinsic-dynamics latents that index the pretrained constitutive representation space.

3.4 Inference with Unified Intrinsic Dynamics Representation

Zero-Shot Dynamics Inference. After the two-stage training, *NeuIDO* has learned (i) a pretrained constitutive representation space and (ii) a shared dynamics encoder that maps visual observations to the dynamics latents indexing this space. Therefore, *NeuIDO* can infer intrinsic dynamics for an unseen scene in a purely feed-forward manner. Given an input video observation $\mathbf{u}$ from a new scene, we first extract its dynamics latent using the branch encoder $\mathbf{z} = \mathbf{E}(\mathbf{u})$.

The latent $\mathbf{z}$ then instantiates a scene-specific constitutive function within the pretrained operator space, denoted as $\widehat{\mathcal{G}}^E(\mathbf{z})$ and $\widehat{\mathcal{G}}^P(\mathbf{z})$. We plug these constitutive functions into the physics-driven simulation pipeline to roll out the 4D dynamics. Importantly, zero-shot inference requires *no additional optimization*. This enables efficient constitutive inference and interaction simulation in previously unseen scenes, reflecting the learned cross-scene generalization of the visual-to-dynamics mapping.

Few-Shot Dynamics Adaptation. While the zero-shot paradigm enables immediate dynamics identification in unseen scenes, it may encounter challenges in complex real-world scenarios where the target dynamics significantly deviate from the training distribution. To improve alignment in such cases, we perform a few-shot adaptation mechanism. Instead of updating the entire model, we insert a lightweight adaptor into the trunk network and jointly refine the scene-specific dynamics latent $\mathbf{z}$ predicted by the dynamics encoder $\mathbf{E}$. This design enables efficient adaptation while preventing overfitting. Given visual observations $\mathbf{u}$ from a scene, we jointly optimize the adaptor parameters θ^E_{adp}, θ^P_{adp} and latent $\mathbf{z}$ by minimizing the reconstruction error:

$$\min_{\theta^E_{\text{adp}},\theta^P_{\text{adp}},\mathbf{z}} \mathcal{L}_{\text{mse}}\Big(\mathcal{R}\big(\widehat{\mathcal{G}}^E(\mathbf{z};\theta^E_{\text{adp}}),\widehat{\mathcal{G}}^P(\mathbf{z};\theta^P_{\text{adp}})\big),\mathbf{u}\Big), \tag{8}$$

where $\widehat{\mathcal{G}}(\mathbf{z};\theta_{\text{adp}})$ denotes the constitutive operator instantiated by latent z with the trunk adaptor applied. This lightweight adaptation enables the model to rapidly specialize to out-of-distribution dynamical environments.

4 Experiments

4.1 Experimental Setup

Implementation Details. We follow a DeepONet-style operator [33] parameterization with a branch–trunk architecture, and set the dynamics latent dimension to 128 (see the Appendix for architectural details). During Training Stage I, each material trajectory is assigned a learnable dynamics latent initialized from $\mathcal{N}(0,1)$. These latents are jointly optimized together with the trunk network for 10 epochs using a batch size of 10^4. In Training Stage II, each visual scene is treated as one training instance. We freeze the operator and optimize only the projector in the branch network. Training alternates over all N scenes for a total of $N \times 100$ epochs. For few-shot adaptation, we insert LoRA modules [17] into the trunk ($r = 16$, $\alpha = 16$) and jointly optimize the adaptor parameters and the scene-specific latent for 100 epochs. All dynamics rollouts are performed with an MPM simulator in a $[0,1]^3$ domain under gravity ($9.8m/s^2$). We use grid resolutions of 32^3 for synthetic data and 70^3 for real-world data. For each scene, we initialize the simulation by reconstructing a 3DGS representation from the first frame following NeuMA [6], and supervise alignment using 2-view videos

	Method	Ball	Cat	Bottle	Duck	Pawn	Fish	Average
CD↓	PAC-NeRF [25]	516.300	15.380	2.210	137.730	15.470	1.71	114.80
	NCLaw [34]	56.690	2.350	0.920	11.970	3.910	1.300	12.860
	NeuMA [6]	1.271	0.679	0.680	3.506	0.882	0.953	1.329
	VisionLaw [27]	**1.085**	0.768	**0.639**	5.205	0.942	1.078	1.620
	Ours(w/o finetune)	1.159	0.844	0.714	5.113	0.973	1.141	1.643
	Ours(w/ finetune)	1.356	**0.617**	0.687	**2.850**	**0.848**	**0.937**	**1.216**

Table 1: Comparison of chamfer distance on the synthetic dataset. Lower chamfer distance indicates better consistency with ground-truth dynamics.

for synthetic scenes and 3-view videos for real scenes. All optimization is performed using the AdamW optimizer with a cosine learning rate scheduler. All experiments are conducted on a single NVIDIA H100 (80GB) GPU.

Datasets and Metrics. For Stage I operator pretraining, we construct a constitutive function dataset covering three classical material families: elastic, plasticine, and sand. For visual dynamics modeling, we adopt six dynamic scenes from NeuMA [6] as synthetic data. Each scene provides RGB videos from 10 viewpoints with 400 frames per view, along with ground-truth particle trajectories. For real-world evaluation, we use three scenes from Spring-Gaus [56]. Each scene provides RGB videos from three viewpoints with 19 frames per view. To further evaluate zero-shot generalization, we additionally collect 10 scenes with distinct elastic dynamics from PAC-NeRF [25], using eight for training and two for testing. Additional dataset details are reported in the Appendix. To evaluate the consistency between the modeled dynamics and the ground-truth dynamics, we (1) use chamfer distance to measure the distribution discrepancy between simulated and ground-truth particle trajectories, and (2) report PSNR, SSIM, and LPIPS to assess visual reconstruction quality.

4.2 Modeling Observed Intrinsic Dynamics

Comparison on Synthetic Datasets. To evaluate intrinsic dynamics modeling within the training distribution, we conduct quantitative comparisons on the synthetic dataset. The chamfer distance (CD) results are reported in Tab. 1. Under the zero-shot setting, *NeuIDO* achieves performance comparable to state-of-the-art methods. Meanwhile, unlike PAC-NeRF, which relies on predefined constitutive models, NCLaw, which requires particle trajectory supervision, and NeuMA and VisionLaw, which perform scene-specific modeling, *NeuIDO* learns a unified dynamics operator from multi-scene visual observations. We observe that zero-shot inference may exhibit larger errors in scenes with significant distribution shifts from the pretraining data (e.g., JellyDuck). With few-shot adaptation, the average CD is reduced from 1.643 to 1.216, outperforming all baselines. Notably, *NeuIDO* reaches this performance within only 100 adaptation

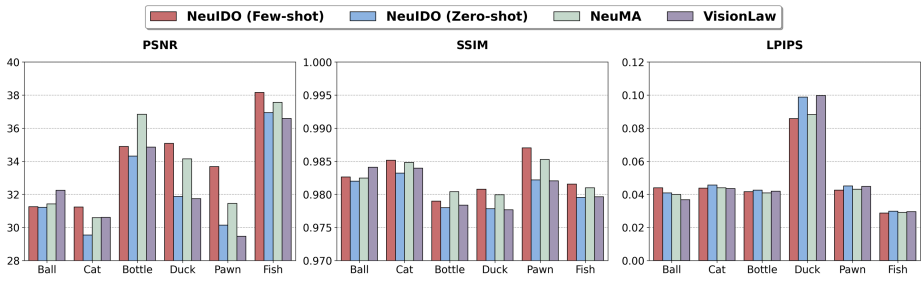

Fig. 3: Comparison of visual fidelity on the synthetic dataset. We report average PSNR, SSIM and LPIPS between simulated and ground-truth videos on novel views. Higher PSNR/SSIM and lower LPIPS indicate better reconstruction quality.

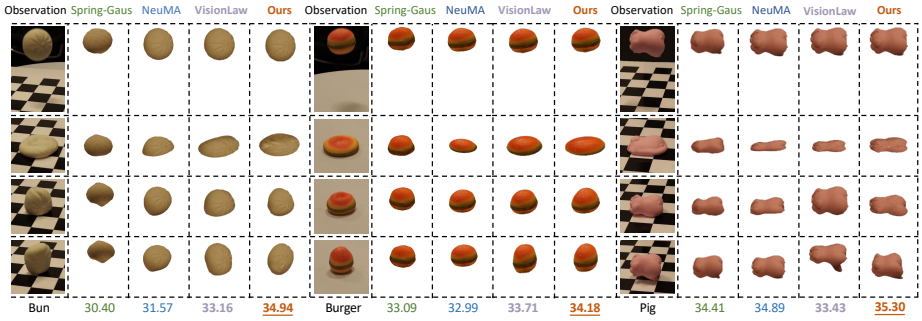

Fig. 4: Comparison on real-world datasets. Quantitative metrics (i.e., PSNR) between the predicted and observed frames are reported in the bottom row. Our method achieves the best visual dynamics reconstruction.

iterations, while NeuMA requires around 1000 iterations even when initialized from an NCLaw-pretrained model, demonstrating its ability to efficiently adapt to complex scenes. Fig. 3 further reports PSNR, SSIM, and LPIPS on unseen viewpoints. These trends align with the CD results in Tab. 1. Notably, although VisionLaw achieves the lowest CD on the Bottle scene, it does not yield the best visual quality, suggesting that accurate dynamics modeling does not necessarily translate to superior visual reconstruction.

Comparison on Real-World Datasets. To evaluate real-world dynamics modeling, Fig. 4 compares *NeuIDO* (with few-shot adaptation) with Spring-Gaus, NeuMA, and VisionLaw on real-world datasets. Spring-Gaus embeds a spring–mass system into 3DGS for elastic modeling, but its limited dynamic expressivity struggles with complex real-world dynamics. NeuMA integrates MPM simulation with neural networks and outperforms Spring-Gaus, though its performance heavily relies on an NCLaw-pretrained model. VisionLaw introduces physical inductive biases via LLMs to guide constitutive evolution, yet its op-

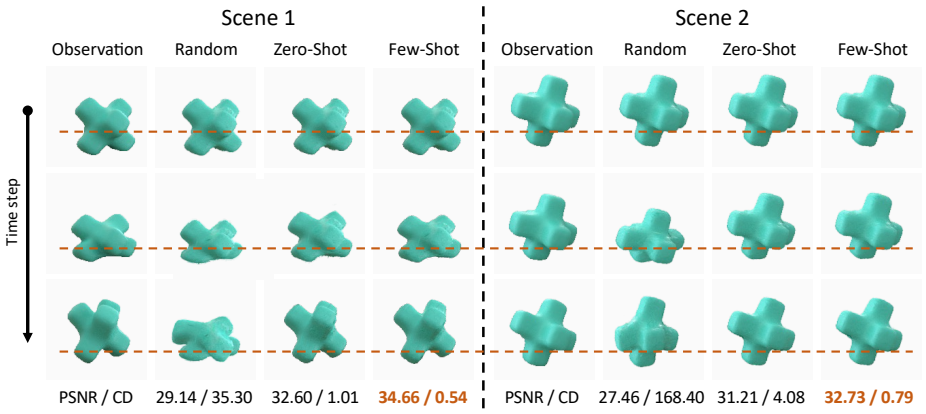

Fig. 5: Generalization on unseen scenes. Quantitative metrics (PSNR ↑ / CD ↓) and qualitative results for two test scenes. The random baseline uses initialized dynamics latents randomly. *NeuIDO* produces dynamics close to the ground-truth observation in the zero-shot setting and further improves with few-shot adaptation.

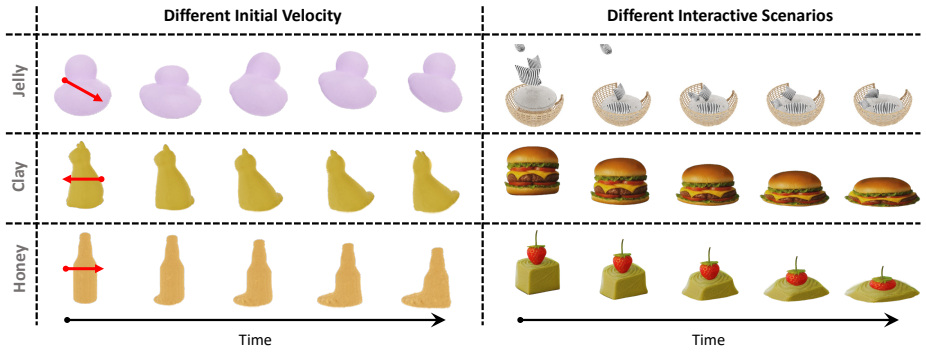

Fig. 6: Transfer of modeled intrinsic dynamics to new simulation conditions. The left text indicates the intrinsic dynamics learned by *NeuIDO*. Left: Examples with different initial velocities (red arrows denote velocity directions). Right: Interactions with new simulation objects.

timization is prone to local minima in complex real-world scenes. As shown in Fig. 4, *NeuIDO* yields reconstructions closer to real observations and achieves the highest PSNR across all real scenes. We attribute this improvement to the unified dynamics representation learned by *NeuIDO*, which provides strong physical priors for few-shot adaptation, thus enabling more effective alignment to complex real-world dynamics. Together, these results demonstrate that *NeuIDO* effectively captures intrinsic dynamics in real-world scenarios, providing a practical pathway toward physics-informed world models and a stronger foundation for dynamics understanding in embodied agents interacting with real environments.

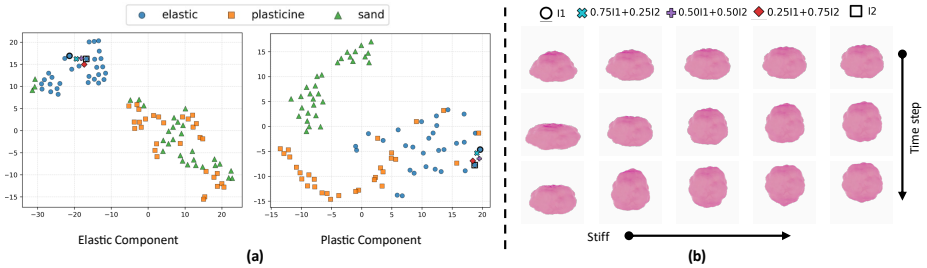

Fig. 7: Representation analysis. (a) t-SNE visualization of Stage-I pretrained dynamics latents of 90 material trajectories. (b) 4D rollouts generated by interpolating between two dynamics latents. Top markers indicate the interpolation settings, showing a smooth transition from soft to stiff behavior.

4.3 Generalization Analysis

Generalization to Unseen Scenarios. While the previous experiments evaluate intrinsic dynamics modeling within the training distribution, we further examine whether *NeuIDO* can infer dynamics in completely unseen scenes. Following PAC-NeRF, we select 10 scenes with diverse elastic behaviors, using eight for training and two as unseen test scenes. As shown in Fig. 5, *NeuIDO* achieves strong zero-shot performance on the unseen scenes, significantly outperforming the random baseline in both qualitative results and quantitative metrics (PSNR / CD). Importantly, *NeuIDO* correctly distinguishes different physical behaviors purely from visual observations without any optimization. For example, Scene 1 exhibits softer deformation patterns, while Scene 2 shows a noticeably more rigid response. These results suggest that the model does not simply memorize the appearance of training scenes, but instead learns a transferable mapping from visual observations to intrinsic dynamics. Overall, *NeuIDO* generalizes beyond the training distribution and enables immediate dynamics inference in unseen environments, providing a promising foundation for physics-informed world modeling to support embodied decision-making in novel environments.

Generalization to Different Simulation Conditions. We further evaluate whether the learned intrinsic dynamics can generalize across different simulation conditions. To this end, we transfer the modeled dynamics to 4D interaction tasks following the interaction paradigms of PhysGaussian and Phy124, while modifying the simulation setup, including the initial velocity and interacting objects. As shown in Fig. 6, the resulting dynamics remain consistent with the original observations despite changes in simulation conditions. For example, when an initial velocity toward the lower-right direction is applied, JellyDuck still exhibits clear elastic deformation and recovery behavior. Similarly, when clay dynamics are applied to Hamburger, the object reproduces the slow settling behavior characteristic of clay materials. These results suggest that *NeuIDO* learns scene-independent intrinsic dynamics that generalize across simulation conditions.

Dim	Pretraining		Alignment						
	Elasticity	Plasticity	Ball	Cat	Bottle	Duck	Pawn	Fish	Average
32	1.023e3	3.598e-8	**3.721%**	-33.850%	-12.497%	-0.052%	-2.108%	-13.289%	-9.679%
64	8.392e2	**3.211e-8**	-7.361%	-29.942%	-9.340%	**-0.046%**	**3.240%**	-14.247%	-9.616%
256	**6.885e2**	3.341e-8	-3.229%	**4.631%**	**6.397%**	-1.573%	-0.890%	**3.418%**	**1.459%**

Table 2: Analysis of impact of dynamics latent dimensionality on operator expressiveness. We report the MSE loss in Stage I (pretraining) and the relative change in chamfer distance with respect to the 128-dimensional in Stage II (alignment).

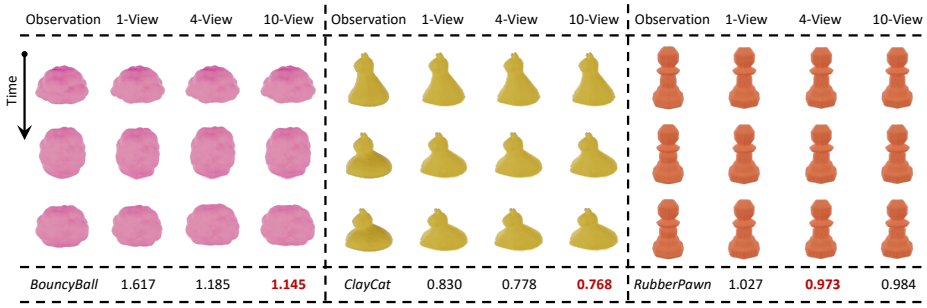

Fig. 8: Effect of multi-view supervision. Comparisons using 1, 4, and 10 input views during Stage II alignment. Chamfer distances are shown in the bottom row.

4.4 Representation Analysis

To examine whether *NeuIDO* learns a structured intrinsic dynamics representation, we visualize the Stage-I pretrained dynamics latents of 90 material trajectories using t-SNE and analyze their interpolation behavior. As shown in Fig. 7 (a), the latents from the same constitutive family are clustered together, while different constitutive families are clearly separated. We further evaluate the continuity of this representation by interpolating between two dynamics latents and rolling out the corresponding 4D dynamics. Fig. 7 (b) reveals a smooth transition from soft to stiff motion along the interpolation. These results show that the learned latent space is discriminative across constitutive families and continuous within the dynamics manifold, supporting *NeuIDO*'s zero-shot inference.

4.5 Ablation Study

Effect of Latent Dimensionality. We study the impact of dynamics latent dimensionality on operator expressiveness by evaluating different latent sizes. Specifically, we report the MSE loss in Stage I (pretraining) and relative change in chamfer distance with respect to the 128-dimensional default in Stage II (alignment). As shown in Tab. 2, in the pretraining phase, smaller latent dimensions (e.g., 32) result in significantly higher fitting errors, indicating insufficient capacity to index diverse constitutive functions. Increasing the latent dimension

to 64 and 256 substantially reduces the error. A similar trend is observed in the alignment phase. The 32-dimensional setting degrades performance across several scenes, whereas the 256-dimensional latent achieves the best or near-best results in most cases. These results suggest that latent dimensionality directly affects the expressive capacity of the operator: higher-dimensional latents allow the model to capture more complex dynamics and improve fitting accuracy.

Effect of Multi-View Supervision. We analyze the impact of multi-view supervision by training the operator with 1, 4, and 10 input views during Stage II alignment (Fig. 8). Increasing supervision from a single view to multiple views significantly improves dynamics reconstruction, as additional viewpoints help reduce geometric ambiguity and incomplete observations. However, increasing the number of views from 4 to 10 provides only marginal improvement, indicating diminishing returns from additional visual supervision. A similar trend is observed in the few-shot multi-view ablation reported in the Appendix. These results suggest that simply increasing the number of views does not continuously improve intrinsic dynamics modeling. Since intrinsic dynamics are inferred indirectly through image-space reconstruction, rendering errors may accumulate. Therefore, further improvements in dynamics reconstruction likely require stronger physical constraints rather than additional visual observations.

5 Conclusion

In this paper, we introduce *NeuIDO*, a world dynamics modeling framework that learns a unified representation of intrinsic dynamics from visual observations across scenes. By formulating world modeling as a neural operator learning problem, *NeuIDO* establishes a mapping from the observation space to the constitutive function space, enabling diverse scene dynamics to be captured within a shared representation. To make this feasible, we introduce a two-stage training strategy: (i) pretraining a structured representation space of constitutive functions from a collection of known constitutive relations, and (ii) aligning visual observations with this learned space to establish the observation–dynamics link. Experiments on both synthetic and real-world datasets demonstrate that *NeuIDO* effectively models cross-scene dynamics, enabling real-time zero-shot dynamics inference and efficient few-shot adaptation in unseen scenes. These results highlight the potential of *NeuIDO* as a step toward physics-informed world models for embodied planning and decision-making.

Acknowledgements

This work was supported by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China, the Shanghai Committee of Science and Technology, China (Grant No.24511103200), Shanghai Artificial Intelligence Laboratory, XPLORER PRIZE grants. We sincerely thank Jikuang Zhang for his valuable contributions to this work.

References

1. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15619–15629 (2023)
2. Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting feature prediction for learning visual representations from video. arXiv preprint arXiv:2404.08471 (2024)
3. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. OpenAI Blog **1**(8), 1 (2024)
4. Bruce, J., Dennis, M.D., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., et al.: Genie: Generative interactive environments. In: Forty-first International Conference on Machine Learning (2024)
5. Cai, J., Yang, Y., Yuan, W., He, Y., Dong, Z., Bo, L., Cheng, H., Chen, Q.: Gic: Gaussian-informed continuum for physical property identification and simulation. arXiv preprint arXiv:2406.14927 (2024)
6. Cao, J., Guan, S., Ge, Y., Li, W., Yang, X., Ma, C.: Neuma: Neural material adaptor for visual grounding of intrinsic dynamics. vol. 37, pp. 65643–65669 (2024)
7. Chaves, E.W.: Notes on continuum mechanics. Springer Science & Business Media (2013)
8. Chen, C., Wu, Y.F., Yoon, J., Ahn, S.: Transdreamer: Reinforcement learning with transformer world models. arXiv preprint arXiv:2202.09481 (2022)
9. Feng, T., Wang, X., Jiang, Y.G., Zhu, W.: Embodied ai: From llms to world models. arXiv preprint arXiv:2509.20021 (2025)
10. Feng, Y., Shang, Y., Li, X., Shao, T., Jiang, C., Yang, Y.: Pie-nerf: Physics-based interactive elastodynamics with nerf. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4450–4461 (2024)
11. Ha, D., Schmidhuber, J.: Recurrent world models facilitate policy evolution. Advances in neural information processing systems **31** (2018)
12. Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to control: Learning behaviors by latent imagination. arXiv preprint arXiv:1912.01603 (2019)
13. Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., Davidson, J.: Learning latent dynamics for planning from pixels. In: International conference on machine learning. pp. 2555–2565. PMLR (2019)
14. Hafner, D., Lillicrap, T., Norouzi, M., Ba, J.: Mastering atari with discrete world models. arXiv preprint arXiv:2010.02193 (2020)
15. Hafner, D., Pasukonis, J., Ba, J., Lillicrap, T.: Mastering diverse domains through world models. arXiv preprint arXiv:2301.04104 (2023)
16. Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., Corrado, G.: Gaia-1: A generative world model for autonomous driving. arXiv preprint arXiv:2309.17080 (2023)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
18. Huang, T., Zeng, Y., Li, H., Zuo, W., Lau, R.W.: Dreamphysics: Learning physical properties of dynamic 3d gaussians with video diffusion priors (2025)
19. Jiang, Y., Yu, C., Xie, T., Li, X., Feng, Y., Wang, H., Li, M., Lau, H., Gao, F., Yang, Y., et al.: Vr-gs: A physical dynamics-aware interactive gaussian splatting system in virtual reality. In: Proceedings of the ACM SIGGRAPH. pp. 1–1 (2024)

20. Kang, B., Yue, Y., Lu, R., Lin, Z., Zhao, Y., Wang, K., Huang, G., Feng, J.: How far is video generation from world model: A physical law perspective. arXiv preprint arXiv:2411.02385 (2024)
21. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (TOG) **42**(4), 139–1 (2023)
22. Kong, L., Yang, W., Mei, J., Liu, Y., Liang, A., Zhu, D., Lu, D., Yin, W., Hu, X., Jia, M., et al.: 3d and 4d world modeling: A survey. arXiv preprint arXiv:2509.07996 (2025)
23. Kossaifi, J., Kovachki, N., Li, Z., Pitt, D., Liu-Schiaffini, M., Duruisseaux, V., George, R.J., Bonev, B., Azizzadenesheli, K., Berner, J., et al.: A library for learning neural operators. arXiv preprint arXiv:2412.10354 (2024)
24. LeCun, Y., et al.: A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. Open Review **62**(1), 1–62 (2022)
25. Li, X., Qiao, Y.L., Chen, P.Y., Jatavallabhula, K.M., Lin, M., Jiang, C., Gan, C.: Pac-nerf: Physics augmented continuum neural radiance fields for geometry-agnostic system identification (2023)
26. Li, Z., Yu, H.X., Liu, W., Yang, Y., Herrmann, C., Wetzstein, G., Wu, J.: Wonderplay: Dynamic 3d scene generation from a single image and actions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9080–9090 (2025)
27. Lin, J., Jiang, S., Zeng, Q., Wang, Z., Jiang, M.: Visionlaw: Inferring interpretable intrinsic dynamics from visual observations via bilevel optimization. arXiv preprint arXiv:2508.13792 (2025)
28. Lin, J., Wang, Z., Hou, Y., Tang, Y., Jiang, M.: Phy124: Fast physics-driven 4d content generation from a single image. arXiv preprint arXiv:2409.07179 (2024)
29. Lin, J., Wang, Z., Jiang, S., Hou, Y., Jiang, M.: Phys4dgen: A physics-driven framework for controllable and efficient 4d content generation from a single image. arXiv e-prints pp. arXiv–2411 (2024)
30. Lin, Y., Lin, C., Xu, J., Mu, Y.: Omniphysgs: 3d constitutive gaussians for general physics-based dynamics generation. arXiv preprint arXiv:2501.18982 (2025)
31. Liu, F., Wang, H., Yao, S., Zhang, S., Zhou, J., Duan, Y.: Physics3d: Learning physical properties of 3d gaussians via video diffusion. arXiv preprint arXiv:2406.04338 (2024)
32. Liu, Z., Ye, W., Luximon, Y., Wan, P., Zhang, D.: Unleashing the potential of multi-modal foundation models and video diffusion for 4d dynamic physical scene simulation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11016–11025 (2025)
33. Lu, L., Jin, P., Karniadakis, G.E.: Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. arXiv preprint arXiv:1910.03193 (2019)
34. Ma, P., Chen, P.Y., Deng, B., Tenenbaum, J.B., Du, T., Gan, C., Matusik, W.: Learning neural constitutive laws from motion observations for generalizable pde dynamics. In: Proceedings of the International Conference on Machine Learning. pp. 23279–23300. PMLR (2023)
35. Macklin, M., Müller, M., Chentanez, N.: Xpbd: position-based simulation of compliant constrained dynamics. In: Proceedings of the International Conference on Motion in Games. pp. 49–54 (2016)
36. Micheli, V., Alonso, E., Fleuret, F.: Transformers are sample-efficient world models. arXiv preprint arXiv:2209.00588 (2022)

37. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
38. Müller, M., Heidelberger, B., Hennix, M., Ratcliff, J.: Position based dynamics. Journal of Visual Communication and Image Representationv (JVCI) **18**(2), 109–118 (2007)
39. Robine, J., Höftmann, M., Uelwer, T., Harmeling, S.: Transformer-based world models are happy with 100k interactions. arXiv preprint arXiv:2303.07109 (2023)
40. Russell, L., Hu, A., Bertoni, L., Fedoseev, G., Shotton, J., Arani, E., Corrado, G.: Gaia-2: A controllable multi-view generative world model for autonomous driving. arXiv preprint arXiv:2503.20523 (2025)
41. Stomakhin, A., Schroeder, C., Chai, L., Teran, J., Selle, A.: A material point method for snow simulation. ACM Transactions on Graphics (TOG) **32**(4), 1–10 (2013)
42. Sun, J., Li, Y., Wei, J., Xu, L., Wang, N., Zhang, Y., Lu, C.: Arti-pg: A toolbox for procedurally synthesizing large-scale and diverse articulated objects with rich annotations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6396–6405 (2025)
43. Sun, J., Li, Y., Xu, L., Wei, J., Chai, L., Lu, C.: Discovering conceptual knowledge with analytic ontology templates for articulated objects. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 14681–14689 (2025)
44. Sun, J., Lu, C.: Digital gene: Learning about the physical world through analytic concepts. arXiv preprint arXiv:2504.04170 (2025)
45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
46. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: European conference on computer vision. pp. 55–72. Springer (2024)
47. Wei, J., Li, Y., Lu, C., Sun, J.: Physically ground commonsense knowledge for articulated object manipulation with analytic concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13519–13528 (2026)
48. Wu, P., Escontrela, A., Hafner, D., Abbeel, P., Goldberg, K.: Daydreamer: World models for physical robot learning. In: Conference on robot learning. pp. 2226–2240. PMLR (2023)
49. Wu, Z., Xu, H., Xu, G., Nie, P., Yan, Z., Zheng, J., Qu, L., Li, M., Nie, L.: Textsplat: Text-guided semantic fusion for generalizable gaussian splatting. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 8478–8487 (2025)
50. Xie, T., Zong, Z., Qiu, Y., Li, X., Feng, Y., Yang, Y., Jiang, C.: Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4389–4398 (2024)
51. Yang, M., Du, Y., Ghasemipour, K., Tompson, J., Schuurmans, D., Abbeel, P.: Learning interactive real-world simulators. arXiv preprint arXiv:2310.06114 **1**(2), 6 (2023)
52. Zhang, T., Yu, H.X., Wu, R., Feng, B.Y., Zheng, C., Snavely, N., Wu, J., Freeman, W.T.: Physdreamer: Physics-based interaction with 3d objects via video generation (2024)

53. Zhang, Z., Liao, J., Li, M., Dai, Z., Qiu, B., Zhu, S., Qin, L., Wang, W.: Tora: Trajectory-oriented diffusion transformer for video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2063–2073 (2025)
54. Zhao, G., Wang, X., Zhu, Z., Chen, X., Huang, G., Bao, X., Wang, X.: Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10412–10420 (2025)
55. Zhao, Y., Chen, H., Liu, C., Li, Z., Herrmann, C., Hur, J., Li, Y., Yang, M.H., Raj, B., Xu, M.: Toward material-agnostic system identification from videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5944–5956 (2025)
56. Zhong, L., Yu, H.X., Wu, J., Li, Y.: Reconstruction and simulation of elastic objects with spring-mass 3d gaussians (2024)
57. Zhou, S., Du, Y., Chen, J., Li, Y., Yeung, D.Y., Gan, C.: Robodreamer: Learning compositional world models for robot imagination. arXiv preprint arXiv:2404.12377 (2024)
58. Zhu, Z., Wang, X., Zhao, W., Min, C., Li, B., Deng, N., Dou, M., Wang, Y., Shi, B., Wang, K., et al.: Is sora a world simulator? a comprehensive survey on general world models and beyond. arXiv preprint arXiv:2405.03520 (2024)