Physics Attention: A Physics-Learned Generative World Model with Disentangled Latent Physical Factors for Fluid

Jinghao Cui*
DCST, Tsinghua University
Beijing, China
cuijh22@mails.tsinghua.edu.cn

Xin Wang†
DCST, BNRist, Tsinghua University
Beijing, China
xin_wang@tsinghua.edu.cn

Tongtong Feng†
DCST, BNRist, Tsinghua University
Beijing, China
fengtongtong@tsinghua.edu.cn

Yong Rui
Tsinghua University
Beijing, China
yrui@air.tsinghua.edu.cn

Wenwu Zhu
DCST, BNRist, Tsinghua University
Beijing, China
wwzhu@tsinghua.edu.cn

Abstract

Generative world models have demonstrated strong capability in modeling complex environment dynamics. However, existing approaches mainly rely on data pattern matching or physics-informed generation with externally given physical priors, failing to learn the intrinsic and coupled physical laws. Those challenges are particularly severe in fluid environments, where governing physical laws are implicit and jointly determined by multiple related physical factors. To solve these challenges, we propose the Physics Attention World Model (**PAWM**), a physics-learned generative world model that autonomously learns and decorrelated informs latent physics from raw fluid observations. Specifically, PAWM comprises two core modules: i) the Physics Attention module, which learns disentangled latent physics and achieves physics-based consistent constraints, and ii) the Decorrelated Physics-informed Prediction module, which first conducts decorrelation reweighting to quantify the contribution of different physical factors in video generation with theoretical support, then performs state transition and prediction conditioned on these learned and decorrelated latent physics. Extensive experiments on three fluid benchmarks demonstrate that PAWM significantly outperforms existing generative world models with respect to accurate and physically consistent prediction across both simulated and real-world fluid scenarios, as well as maintaining strong generalization across diverse fluid properties.

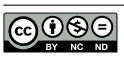
CCS Concepts

• **Computing methodologies** → *Learning latent representations.*

Keywords

Generative World Model, Physics Attention, Physical Law

*DCST is the abbreviation for Department of Computer Science and Technology.
†Corresponding author. BNRist is the abbreviation for Beijing National Research Center for Information Science and Technology.



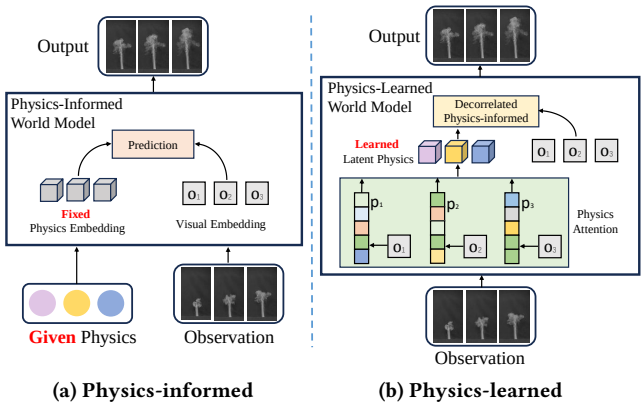
This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
MM '26, Rio de Janeiro, Brazil
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3835812

ACM Reference Format:
Jinghao Cui, Xin Wang, Tongtong Feng, Yong Rui, and Wenwu Zhu. 2026. Physics Attention: A Physics-Learned Generative World Model with Disentangled Latent Physical Factors for Fluid. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3835812

1 Introduction

Recently, generative world models [1, 20] have achieved impressive success in modeling complex visual dynamics, enabling realistic video generation and future prediction by learning high-dimensional spatiotemporal patterns from large-scale datasets. However, recent physical law evaluation experiments [22, 50, 55] reveal that these models frequently violate fundamental physical principles, indicating that existing world models fail to learn the underlying physical laws and instead rely on data pattern matching [22], a critical limitation for complex world-modeling environments [12, 13, 44, 49]. To address this issue, physics-informed world models [8, 17, 36, 39] incorporate explicit physical parameters or external constraints to guide generation and enforce physical consistency. Despite their effectiveness, these approaches largely depend on externally provided physical knowledge, underscoring the need for world models that can learn and adhere to physical laws from data internally, as explored in this work.

This challenge becomes particularly pronounced in modeling fluid phenomena, which present two fundamental difficulties. First, governing physical laws are implicit and not directly observable from visual data. Unlike rigid-body systems with explicit state factors, fluid dynamics involves latent fields (e.g., velocity, pressure, density) whose evolution is only indirectly reflected in observations, making the underlying structure difficult to infer [10, 33]. Second, fluid behavior arises from the coupled interaction of multiple physical mechanisms. Factors such as velocity fields, pressure gradients, density variations, and dimensionless quantities (Reynolds number) jointly govern highly nonlinear and interdependent dynamics, leading to turbulence and multiscale structures that are tightly coupled [2, 31]. Consequently, the high dimensionality and strong nonlinearity of fluid motion render accurate prediction from visual observations alone exceptionally challenging, especially over long horizons or under distribution shifts [9], leading purely data-driven



(a) Physics-informed **(b) Physics-learned**

Figure 1: (a) Physics-informed World Models; (b) Physics-learned World Models (Ours). In contrast to Physics-informed World Models requiring external physics inputs, our Physics-learned World Model autonomously learns and attends to physical information from raw videos.

world models to suffer from poor generalization, error accumulation, and physically inconsistent behaviors. These challenges highlight a fundamental gap between visual realism and physical understanding, underscoring the need for world models that can infer implicit physical structure and disentangle interacting physical factors, rather than relying solely on external supervision or appearance-level correlations.

In this paper, we introduce the Physics Attention World Model (PAWM), a novel framework that aims to bridge the gap between visual prediction and physical understanding by enabling world models to autonomously extract and utilize underlying physical laws from data. Inspired by the observation that physical dynamics manifest as structured and persistent patterns across time, PAWM formulates physical law discovery as an attention-driven representation learning process within a structured latent space, thereby enabling the model to infer implicit physical laws that are not directly observable from visual data. The Physics Attention module is designed not only to selectively attend to physically meaningful representations in the observations, but also to disentangle and decouple these representations into distinct latent channels, thereby isolating interpretable physical components [7, 26, 40–42, 45]. The Decorrelated Physics-informed Prediction module integrates the learned representations into state-transition and prediction functions while explicitly constraining the dynamics to be physically consistent; it further implements a decorrelation reweighting pipeline to quantify the predictive importance of individual physical factors. Together, the disentanglement and decorrelation mechanisms mitigate the challenge of multiple intertwined physical processes by promoting independent and interpretable representations. By joining data-driven perception with physics-learned policy in a unified framework, PAWM provides a general and scalable approach for learning physically grounded world models, where external supervision of physical parameters is optional and can further improve model performance when available.

We evaluate PAWM on three representative fluid benchmarks (FlenBench, ScalarFlow, and FluidNexus), covering diverse flow regimes and physical settings. Experimental results demonstrate that PAWM significantly outperforms existing generative world models in fluid prediction accuracy. Ablation studies further validate the effectiveness of physics-learned framework, which contributes to performance gains. In addition, quantitative analysis of decorrelation reveals the relative importance of different physical factors, providing an interpretable perspective on the model's underlying physical representations. These results indicate that learning and attending to intrinsic latent physics enables more accurate and physically consistent modeling of complex fluid dynamics.

In summary, the main contributions of the paper are as follows:

- We propose **PAWM**, the first physics-learned generative world model that autonomously learns and informs latent physics from raw fluid observations.
- We propose the Physics Attention module and the Decorrelated Physics-informed Prediction module, which enables the learned latent physics from fluid environments to be capable of achieving physically consistent prediction.
- We propose the decorrelation reweighting policy, which enhances interpretability by quantifying the relative contribution of different physical factors in video generation with theoretical support.
- We conduct extensive experiments to demonstrate that the proposed PAWM can accurately model and predict fluid environments, outperforming existing generative world models in both simulated and real-world fluid scenarios.

2 Related Work

Generative World Models. Generative world models aim to learn predictive representations of environment dynamics from large-scale sensory data, enabling video generation, future prediction, and planning. Early approaches model latent dynamics from visual observations via variational inference or autoregressive formulations [9, 14, 18, 19, 23], while recent methods leverage diffusion and transformer-based architectures to capture high-dimensional spatiotemporal structure with improved visual fidelity and long-horizon consistency [1, 5, 6, 15, 20, 47, 53, 56]. Despite these advances, purely data-driven models often prioritize appearance-level correlations over underlying physical mechanisms [3, 24], leading to violations of fundamental physical laws and degraded performance under distribution shifts or long-horizon rollouts [22, 29, 50, 55]. This limitation becomes especially evident under distribution shifts or long-horizon rollouts, where small prediction errors accumulate and lead to physically implausible behaviors. These findings suggest that high visual realism alone does not guarantee physical understanding, motivating research beyond pattern-based generative modeling.

Physics-informed World Models. To improve physical consistency, physics-informed world models incorporate explicit physical knowledge into the learning or generation process. A common strategy is to inject known physical parameters, equations, or constraints, such as rigid-body dynamics, force models, or conservation laws, either as additional inputs or as regularization terms during training [17, 24, 32, 39]. Other approaches integrate differentiable physics simulators or structured inductive biases, such as graph-based interaction models or equivariant architectures, to guide predictions and improve extrapolation and stability [3, 4, 35, 36].

More recently, hybrid frameworks combine neural world models with symbolic reasoning or language-based physical priors, where external solvers or large language models provide corrective feedback or high-level guidance [8, 30]. Although these methods improve physical plausibility, they rely on external knowledge, simulators, or supervision, limiting scalability in complex domains like fluid dynamics, where governing laws are highly nonlinear and context-dependent. In contrast, our approach enables world models to autonomously discover and attend to intrinsic physical structure from data, without explicit annotations or hand-crafted constraints, bridging visual prediction and physical understanding.

Fluid Reconstruction Models. A line of research in fluid dynamics, complementary to general-purpose world modeling, focuses on reconstructing or inferring flow fields from sparse, partial, or indirect observations. Early and representative approaches combine neural networks with numerical solvers or physical constraints to recover physically consistent states, such as hybrid neural frameworks [32, 38, 48]. Building on this direction, graph-based and structured models capture spatial coherence and long-range interactions for more faithful reconstruction of complex flows [3, 27, 34]. More recently, unified frameworks and neural operator methods enable generalization across diverse physical parameters, boundary conditions, and forcing regimes [16, 25, 28]. Despite these advances, most methods focus on fluid state estimation or inverse problems under predefined assumptions, relying on explicit supervision, solvers, or handcrafted objectives, which limits scalability to diverse phenomena and long-horizon prediction. In contrast, our approach models fluid dynamics as an emergent structure within a unified world model, leveraging internal representation learning and physics-grounded attention to capture intrinsic patterns directly from data.

3 Physics Attention World Model

We propose the Physics Attention World Model (PAWM), a unified framework for discovering and incorporating abstract physical laws into world model. As illustrated in Figure 2, PAWM consists of three components: (i) a VAE-based encoder–decoder module for compact stochastic latent representations; (ii) a Physics Attention module that extracts semantically meaningful and physically grounded representations; and (iii) a transformer-based dynamics core, Decorrelated Physics-informed Prediction module, which conditions state transitions on these representations. In this section, we formalize PAWM in physically governed environments, describe its core modules, and define training objectives.

3.1 Framework Formulation

Given a sequence of consecutive visual observations $\{o_t\}_{t=1}^{T}$ from a dynamical system (e.g., fluid flow videos), the goal of generative world models is to learn a latent dynamics model that can predict future states o_{t+1}. Formally, the model aims to maximize the likelihood $p(o_{t+1}|o_{1:t})$, typically by learning a latent state representation z_t and a transition function p_θ. In the context of physics systems, the underlying dynamics are governed by physical laws P, such as conservation laws, Navier-Stokes equations. Based on generative world models, our objective is to learn a model that not only learns a disentangled latent representation p_t that encodes physically meaningful representations but also leverages both p and z to accurately predict future observations, thereby enabling predictions consistent with the intrinsic physical laws P. By decorrelating z_t and p_t, w_t denotes decorrelation-based reweighting coefficients, reducing interference from intertwined physical processes and enabling the analysis of the importance of individual physical factors. We formalize the overall PAWM workflow in a world model setting with physics-learned latent representations. The complete PAWM process can be expressed as:

$$\begin{aligned} \text{Encoder:} \quad & z_t \sim q_\phi(z_t \mid o_t), \\ \text{Physics Attention:} \quad & p_t = f_{\text{phyAttn}}(z_t), \\ \text{Decorrelation:} \quad & w_t = f_{\text{deco}}(z_t, p_t), \\ \text{Physics-informed Prediction:} \quad & \hat{Z}_{t+1} = g_\theta(\hat{z}_{t+1} \mid z_{1:t}, p_{1:t}, w_t), \\ \text{Decoder:} \quad & \hat{o}_t = p_\phi(z_t). \end{aligned} \tag{1}$$

3.2 Physics Attention

To obtain structured and disentangled physics representations, we introduce the Physics Attention module that maps the latent state z_t to physics representation p_t. This module consists of two complementary components: a Physics Encoder, which learns disentangled and semantically meaningful physics representations from the latent state, and a Physics Decoder, which reconstructs physical parameters from these representations while enforcing physics consistency constraints. Together, these components enable the model to capture underlying physical laws in a latent space.

Physics Encoder. Physics Encoder is implemented as an attention mechanism between a sequence of latent state derived from z_t and a small set of learnable latent physics: $p_t = f_{\text{phyAttn}}(z_t)$. Intuitively, each latent physics is designed to capture a different latent physical factor, such as a global control parameter or a localized flow characteristic, as illustrated in Figure 2. By using multi-head attention, the module adaptively routes physics information from the latent state to different latent physics, thereby producing disentangled and semantically structured physics representations.

To promote disentangled representations across different physics attention heads, we introduce an independence disentangle, which encourages each latent physics to specialize in one meaningful physical factor while discouraging highly correlated encodings across features. Let $p_{t,k} \in \mathbb{R}^d$ denote the latten representation produced by the k-th physical head at time step t, and let n be the total number of heads. To associate each representation with its corresponding factor, we estimate the probability that $p_{t,k}$ belongs to the k-th factor via a softmax classifier:

$$P(k \mid p_{t,k}) = \text{softmax}(W p_{t,k} + b), \tag{2}$$

where $W \in \mathbb{R}^{n \times d}$ and b are learnable parameters. For the representation generated by head k, we assign the auxiliary label $y = k$ [54]. The independence loss is then defined as

$$\mathcal{L}_{\text{ind}} = \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{n} \text{CrossEntropy}\big(P(k \mid p_{t,k}),\, k\big). \tag{3}$$

This objective encourages different heads to specialize in different physical factors, thereby improving the disentanglement and explainability of the learned physics representations.

Physics Decoder. Subsequently, p_t is processed by the Physics Decoder to produce explicit estimates of the underlying physical parameters: $\hat{P}_t = \text{RegHead}(p_t)$, which are then used to compute

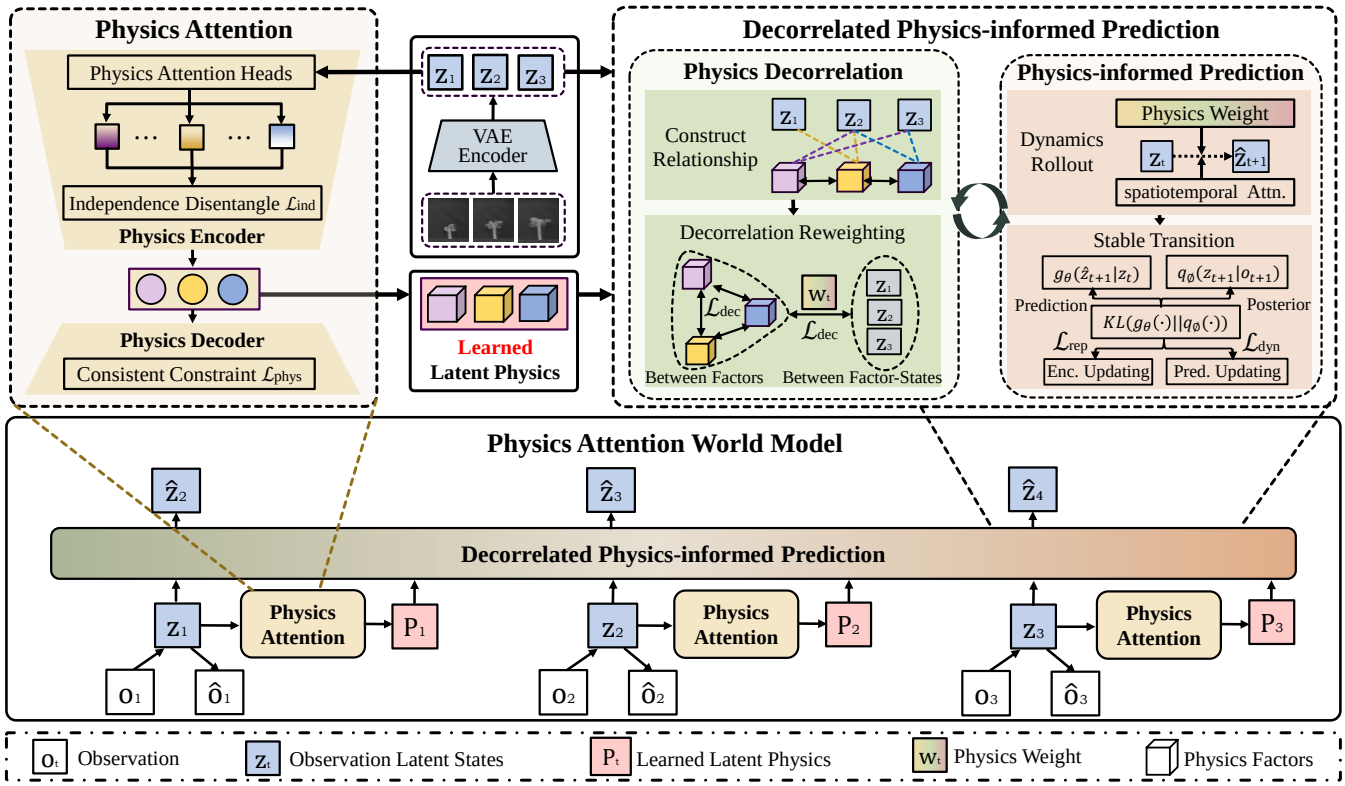


Figure 2: Architecture of the proposed Physics Attention World Model. The model encodes observed states into latent state and latent physics, which interact through physics attention and spatiotemporal attention mechanisms. A Physical Prediction Transformer iteratively updates latent representations across time, while physical parameters are inferred via a regression head from learned physics representations. Cross-attention between latent state and latent physics enables the model to capture underlying physical dynamics and enforce physically consistent future prediction.

the physics loss $\mathcal{L}_{\text{phys}}$. This loss consists of two complementary components: (i) a regression term that supervises the predicted physical parameters against available ground-truth quantities (e.g., velocity (u, v), pressure p, and Reynolds number Re), and (ii) a physics-based constraint that enforces consistency with known physical laws or governing equations. Together, these components provide direct guidance for learning physically meaningful representations, ensuring that the Physics Attention module captures physical factors relevant to the true fluid dynamics rather than spurious correlations.

When ground-truth physical parameters are available, we further apply a supervised regression loss on the predicted physical outputs, including velocity components (u, v), pressure p, and Reynolds number Re, to reinforce physical consistency in the learned representations.

$$\mathcal{L}_{\text{sup}} = \frac{1}{T}\sum_{t=1}^{T}\left(\|\hat{u}_t - u_t\|_2^2 + \|\hat{v}_t - v_t\|_2^2 + \|\hat{p}_t - p_t\|_2^2 + \kappa(\widehat{\text{Re}}_t - \text{Re}_t)^2\right). \tag{4}$$

In addition, we impose physics-based constraints on the decoded velocity and pressure fields derived from the physics representation p_t. Let $\hat{V}$ and $\hat{P}$ denote the corresponding velocity and pressure fields, respectively. The physics constraint loss is defined as

$$\mathcal{L}_{\text{con}} = \mathcal{L}_{\text{cont}} + \mathcal{L}_{\text{NS}} + \mathcal{L}_{\text{bern}} + \mathcal{L}_{\text{noslip}}, \tag{5}$$

where the continuity term enforces incompressibility:

$$\mathcal{L}_{\text{cont}} = \frac{1}{T}\sum_{t=1}^{T}\|\nabla \cdot \hat{V}\|_2^2, \tag{6}$$

the Navier–Stokes residual penalizes violations of the momentum equation:

$$\mathcal{L}_{\text{NS}} = \frac{1}{T}\sum_{t}\|\rho(\partial_t \hat{V} + (\hat{V}\cdot\nabla)\hat{V}) + \nabla\hat{P} - \rho\nu\Delta\hat{V}\|_2^2, \tag{7}$$

the Bernoulli term encourages energy consistency:

$$\mathcal{L}_{\text{bern}} = \frac{1}{T}\sum_{t=1}^{T}\left\|\nabla\left(\hat{P} + \frac{1}{2}\rho\|\hat{V}\|^2\right)\right\|_2^2, \tag{8}$$

and the no-slip term enforces zero velocity on solid boundaries:

$$\mathcal{L}_{\text{noslip}} = \frac{1}{T}\sum_{t=1}^{T}\|B \odot \hat{V}\|_2^2. \tag{9}$$

Overall, the physics loss is defined as

$$\mathcal{L}_{\text{phy}} = \mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{con}}. \tag{10}$$

3.3 Decorrelated Physics-informed Prediction

Physics-informed Prediction. To model the temporal evolution of the system, we employ a transformer-based dynamics core that conditions state transitions on both the latent states and the inferred physics representations. Given the past latent states $z_{1:t}$ and their corresponding physics representations $p_{1:t}$, the dynamics core produces the predictive distribution $\hat{Z}_{t+1} = p_\theta(\hat{z}_{t+1} \mid z_{1:t}, p_{1:t}, w_t)$, where w_t denotes decorrelation-based reweighting coefficients that adjust the contributions of different samples. A latent state for the next time step can then be obtained by sampling from this predictive distribution: $z_{t+1} \sim \hat{Z}_{t+1}$. This formulation allows the latent dynamics to be modulated by the extracted physics representations, ensuring prediction consistent with underlying physical laws.

To learn stable latent state transitions while preventing representation collapse, we adopt Kullback–Leibler objectives with stop-gradient operations, following the design of latent world models [19]. Let $q_\phi(z_t \mid o_t)$ denote the posterior inferred from the observation, and $\hat{Z}_t$ denote the predictive distribution obtained via the latent state transition, as formulated in Equation 1. The dynamics loss updates the transition model while keeping the encoder fixed:

$$\mathcal{L}_{\text{dyn}} = \frac{1}{T}\sum_{t=1}^{T} \max\left[1,\ \mathrm{KL}\Big(\mathrm{sg}\big(q_\phi(z_t \mid o_t)\big) \parallel \hat{Z}_t\Big)\right], \tag{11}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator. Conversely, the representation loss updates the encoder while freezing the dynamics predictor:

$$\mathcal{L}_{\text{rep}} = \frac{1}{T}\sum_{t=1}^{T} \max\left[1,\ \mathrm{KL}\Big(q_\phi(z_t \mid o_t) \parallel \mathrm{sg}\big(\hat{Z}_t\big)\Big)\right]. \tag{12}$$

These two complementary objectives decouple representation learning from transition modeling, thereby enabling stable and efficient training.

Physics Decorrelation. We introduce a decorrelation-based analysis algorithm to improve the explainability of the physics representations learned by PAWM. Our goal is to quantify the individual contribution of each inferred physical factor to next latent state prediction. Inspired by decorrelated reweighting techniques for feature importance analysis [21, 43], we adapt this strategy to the PAWM framework, where physical factors are extracted by the Physics Attention module. Specifically, we reduce statistical dependencies among different physical factors to disentangle their individual effects, while further mitigating the influence of irrelevant visual correlations between physical representations and latent states. This enables a more faithful attribution of how each physical factor contributes to future state prediction.

To analyze the contribution of the physics representation p_t to next state prediction, we consider n observations. Let $P_i \in \mathbb{R}^{l_1}$ denote an inferred physics representation, and $Z_i \in \mathbb{R}^{l_2}$ denote a corresponding latent state. The prediction target is the next latent state Y_i. Stacking rows yields $P \in \mathbb{R}^{n\times l_1}$, $Z \in \mathbb{R}^{n\times l_2}$, and $X = [P, Z] \in \mathbb{R}^{n\times l}$ with $l = l_1 + l_2$. To decorrelate physical factors and reduce the influence of irrelevant state features, we introduce nonnegative weights $w \in \mathbb{R}^n$ to construct a weighted empirical distribution. We then enforce decorrelation between physical factors and other features by minimizing

$$\mathcal{L}_{\text{deco}} = \sum_{j=1}^{l_1} \Big\| \mathbb{E}_w[P_{\cdot j}\Sigma_w X_{\cdot -j}] - \mathbb{E}_w[P_{\cdot j}w]\mathbb{E}_w[X_{\cdot -j}w] \Big\|_2^2, \tag{13}$$

which corresponds to the sum of squared weighted covariances between each physical factor and the remaining features. Here, $X_{\cdot -j}$ denotes the feature vector with the j-th component removed, and Σ_w denotes the weighting operator associated with w, inducing the weighted empirical expectation $\mathbb{E}_w[\cdot]$. Minimizing $\mathcal{L}_{\text{dec}}$ thus constructs a reweighted empirical distribution in which the physical factors are approximately uncorrelated in first-order with other factors.

Under fixed-dimensional asymptotics where l_1 and l_2 are fixed, the proposed decorrelation objective admits a reweighting scheme that asymptotically removes statistical dependence across factors.

Theorem 3.1 (Existence of Decorrelating Weights). *There exists a non-negative weighting vector $w \succeq 0$ such that*

$$\lim_{n\to\infty} \mathcal{L}_{\text{deco}}(w) = 0, \tag{14}$$

and one such solution is given by the density-ratio form $w_i^* = \frac{\prod_{j=1}^{l} \hat{f}(X_{i,j})}{\hat{f}(X_{i,1},\ldots,X_{i,l})}$. Consequently, the weighted empirical cross-covariances asymptotically vanish.

To obtain a well-posed estimator, we consider the regularized optimization

$$\hat{w} = \arg\min_{w\in C} \mathcal{L}_{\text{deco}}(w) + \frac{\alpha}{n}\sum_{i=1}^{n} w_i^2 + \beta\Big(\frac{1}{n}\sum_{i=1}^{n} w_i - 1\Big)^2, \tag{15}$$

where $C = \{w : |w_i| \le c\}$, and c is a fixed constant.

Theorem 3.2 (Uniqueness of Regularized Solution). *Under mild conditions $\alpha n \gg l^2 + \beta, l^2 \gg \max(\alpha, \beta), |X_{i,j}| \le c$, the Equal 15 admits a unique solution $\hat{w}$.*

By combining Theorems 3.1 and 3.2, the learned weighting $\hat{w}$ asymptotically recovers a decorrelating reweighting scheme. In particular, when $n \to \infty$, it approximately enforces $\mathrm{Cov}_{\hat{w}}(P_{\cdot,j}, X_{-j}) \approx 0, \forall j$, thereby reducing spurious correlations among covariates in the reweighted training distribution and improving the reliability of physics representation importance estimation. Proofs of Theorems 3.1 and 3.2 are provided in the supplementary material.

With $\hat{w}$ fixed, we fit a weighted nonlinear surrogate $\phi(\cdot;\pi)$ by

$$\hat{\pi} = \arg\min_{\pi} \sum_{i=1}^{n} \hat{w}_i\big(Y_i - \phi(X_i;\pi)\big)^2. \tag{16}$$

Because decorrelation suppresses spurious dependencies among physical factors, their contributions become identifiable in the surrogate model. We quantify the importance of the j-th physical factor by aggregating the magnitudes of its incoming weights in the first hidden layer,

$$\text{Importance}(j) = \sum_{h=1}^{H} \big|\pi_{j,h}^{(1)}\big|, \tag{17}$$

which measures how strongly the corresponding physical factor influences the prediction of the next-step state.

The Algorithm 1 summarizes the proposed Physics Representation Decorrelation Weighting (PRD) procedure to jointly estimate

Algorithm 1 Physics Representation Decorrelation (PRD)

Require: Observations $X = [P, Z] \in \mathbb{R}^{n \times l}$, next latent state $Y \in \mathbb{R}^{n \times l_2}$.

Ensure: Sample weights $w \in \mathbb{R}^n$ and prediction parameters π.

Initialize $w^{(0)} \geq 0$ and $\pi^{(0)}$.
Compute initial loss $\mathcal{L}(w^{(0)}, \pi^{(0)})$.
Set iteration counter $t \leftarrow 0$.
repeat
  $t \leftarrow t + 1$.
  Update $w^{(t)}$ with gradient descent by fixing π.
  Update $\pi^{(t)}$ with gradient descent by fixing w.
  Evaluate $\mathcal{L}(w^{(t)}, \pi^{(t)})$.
until convergence or maximum iterations reached
return $w = w^{(t)},\ \pi = \pi^{(t)}$.

the weights w and the prediction parameters π. By reweighting samples to remove statistical dependencies among physical factors, PRD yields interpretable importance scores that quantify how different physical components contribute to predicting the future state.

3.4 Training Objectives

Reconstruction Loss. The reconstruction objective ensures that the latent state preserves sufficient visual information for high-fidelity observation prediction. Let $\hat{o}_t$ denote the decoded reconstruction from latent state z_t. The reconstruction loss is defined as

$$\mathcal{L}_{\text{rec}} = \frac{1}{T} \sum_{t=1}^{T} \|o_t - \hat{o}_t\|_2^2 . \tag{18}$$

This loss encourages the latent representation to retain fine-grained information necessary for reconstructing the input observations.

Total Objective. The final training objective aggregates all aforementioned loss terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_1 \mathcal{L}_{\text{dyn}} + \lambda_2 \mathcal{L}_{\text{rep}} + \lambda_3 \mathcal{L}_{\text{phy}} + \lambda_4 \mathcal{L}_{\text{ind}} + \lambda_5 \mathcal{L}_{\text{deco}}, \tag{19}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, and λ_5 are scalar hyperparameters that balance the contributions of different objectives. Here, $\mathcal{L}_{\text{phy}}$, $\mathcal{L}_{\text{ind}}$, and $\mathcal{L}_{\text{deco}}$ denote the physics, independence, and decorrelation losses defined in Eqs. 10, 3, and 13, respectively.

Together, these objectives form a coherent multi-objective training framework: $\mathcal{L}_{\text{rec}}$, $\mathcal{L}_{\text{dyn}}$, and $\mathcal{L}_{\text{rep}}$ ensure accurate and stable latent dynamics, while $\mathcal{L}_{\text{phy}}$, $\mathcal{L}_{\text{ind}}$, and $\mathcal{L}_{\text{deco}}$ regularize the learned physics representations toward interpretability, independence, and physical consistency. This unified formulation enables PAWM to learn dynamics that are both visually faithful and physically grounded.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate PAWM on three complementary fluid dynamics benchmarks that capture real-world complexity and controlled physical variation. Flowbench [37] offers a physics-based simulation benchmark with diverse parameters and boundary conditions for systematic evaluation of physical consistency and generalization. ScalarFlow [11] provides real-world smoke data featuring complex turbulence, partial observability, and long-range dependencies, suitable for realistic long-horizon prediction. FluidNexus [16] supplies real-world multi-view fluid datasets with object interactions and textured backgrounds, facilitating 3D-aware reconstruction and prediction under complex visual conditions.

Baselines. We compare PAWM against four representative baselines that span both fluid-specific modeling approaches and general-purpose visual world models. HyFluid [48], STG [27], and FluidNexus [16] are specialized methods designed explicitly for fluid synthesis and flow prediction. These approaches incorporate domain-specific architectures and inductive biases inspired by fluid dynamics. In contrast, Storm [52] is a generic world model that emphasizes high-fidelity visual generation and short- to medium-horizon prediction across diverse visual domains.

Metrics. To comprehensively evaluate predictive performance, we adopt a suite of image-space metrics: PSNR for pixel-wise fidelity, SSIM [46] for structural preservation, and LPIPS [51] for perceptual similarity. Additionally, we introduce a physics-based metric ΔV, defined as the mean absolute velocity error, to directly assess dynamical fidelity.

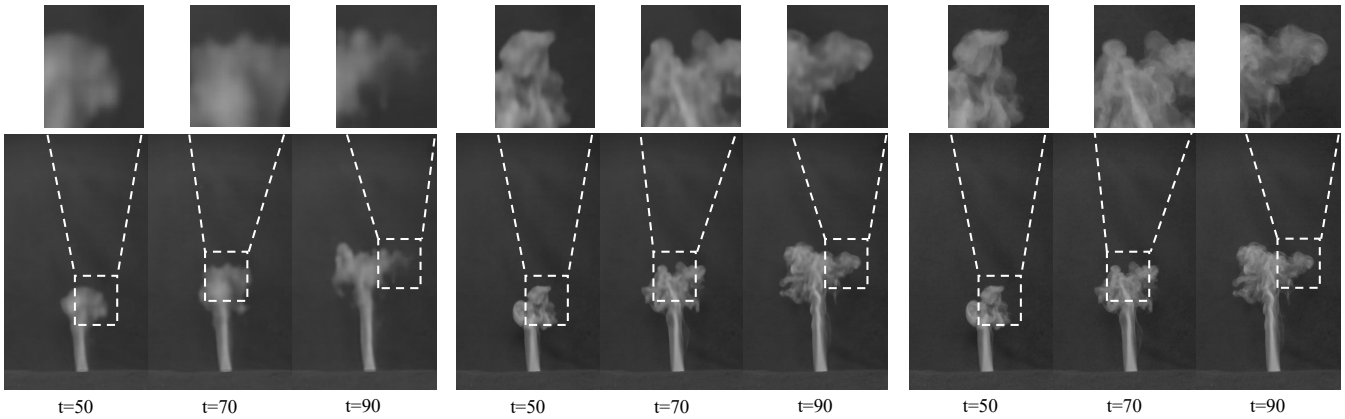
Implementation Details. PAWM is trained from scratch rather than fine-tuned from an existing model. The model uses 32.42M parameters and requires approximately 4.4 GB peak GPU memory during training on a single NVIDIA A100 GPU. On FlowBench, training for 200 epochs with a batch size of 16 takes approximately 6 hours. For inputs of resolution 32×128, the average inference time is 767.5 ms/sample (1.30 FPS).

4.2 Main Results

Table 1 reports quantitative comparisons on ScalarFlow, FluidNexus-Smoke, and FluidNexus-Ball. PAWM achieves consistently strong performance across all datasets, obtaining the best SSIM and LPIPS scores on every benchmark and the highest PSNR on ScalarFlow. Compared with existing methods, PAWM achieves more accurate and visually coherent predictions, particularly in capturing complex deformation patterns and fine-scale dynamics. Although some baselines obtain comparable PSNR values on specific datasets, PAWM provides a better balance between pixel-level fidelity and perceptual quality. Overall, these results demonstrate the effectiveness of integrating physical factors into world modeling and the generalization capability of PAWM across diverse fluid dynamics scenarios.

The qualitative comparisons in Figure 3 further corroborate these findings. PAWM preserves coherent smoke structures, sharp interfaces, and realistic motion patterns that closely match the ground truth, whereas baseline models tend to lose fine details or produce visually implausible artifacts. The consistent advantage in perceptual quality, as reflected by LPIPS, indicates that PAWM generates predictions that are not only quantitatively accurate but also visually faithful.

Learning Settings. PAWM learns physical factors from raw videos in an observation-only setting, and also supporting external supervision from physical parameters when available. In Equation 10, if physical parameters are unavailable (available), $\mathcal{L}_{\text{phy}} = \mathcal{L}_{\text{con}}$ ($\mathcal{L}_{\text{phy}} = \mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{con}}$). In this subsection, all PAWM results in Tables 1 are obtained under the observation-only setting (marked with *).

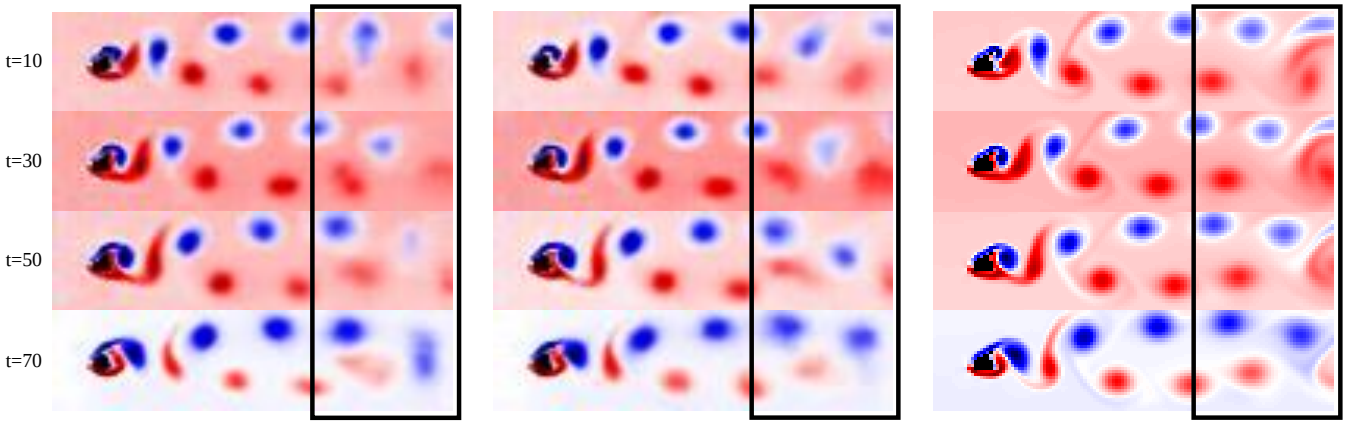


(a) Storm (b) PAWM (Ours) (c) Ground Truth

Figure 3: Qualitative comparison of smoke results on the ScalarFlow dataset under different models. Compared with the appearance-driven baseline Storm, PAWM better preserves coherent smoke structures, sharp boundaries, and fine-scale turbulent details. These results demonstrate PAWM's superior ability to capture the underlying physical evolution of smoke.

Table 1: Quantitative evaluation results on three datasets. Higher PSNR and SSIM indicate better performance, while lower LPIPS is desirable. Models with * were trained without external physical parameters.

Model	ScalarFlow			FluidNexus-Smoke			FluidNexus-Ball		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
HyFluid	26.84	0.9072	0.1776	21.14	0.7044	0.5417	20.93	0.7005	0.5019
STG	21.79	0.8759	0.2142	23.94	0.8408	0.2639	25.80	0.8629	0.2251
FluidNexus	28.51	0.9159	0.1754	**27.79**	0.8747	0.2337	27.70	0.8733	0.2019
Storm*	26.12	0.8456	0.1984	25.26	0.9060	0.2496	26.18	0.9001	0.2711
PAWM(Ours)*	**29.94**	**0.9649**	**0.1445**	26.38	**0.9214**	**0.2164**	**27.91**	**0.9255**	**0.1846**



(1) w/o Physical Attn (2) Whole Model (3) Ground Truth

Figure 4: Comparison of different model variants on the FlowBench fluid simulation dataset at multiple time steps, illustrating the impact of each component on simulation accuracy.

4.3 Ablation Studies

We conduct ablation studies on the FlowBench dataset to assess each component of PAWM, with quantitative results reported in Table 2. The full model achieves the best performance across all metrics, attaining the highest PSNR and SSIM and the lowest LPIPS. Removing the Physics Attention module causes the most pronounced degradation in reconstruction quality, as reflected by drops in PSNR and SSIM and an increase in LPIPS, which aligns with the visual comparisons in Fig. 4, where the variant without Physics Attention exhibits noticeable blurring and distorted structures. This underscores the critical role of Physics Attention in capturing meaningful interactions and preserving coherent flow patterns.

Table 2: Ablation study results on the FlowBench dataset.

Model	PSNR↑	SSIM↑	LPIPS↓	ΔV ↓
w/o Physics Attn	21.06	0.8044	0.0277	–
w/o Independence Loss	20.86	0.7492	0.0265	0.892
w/o Decorrelation Loss	20.70	0.7245	0.0280	1.376
w/o Physics Constraint	21.44	0.7466	0.0274	15.809
w/o Physics Loss	21.51	0.8457	0.0276	18.529
Whole Model	**22.50**	**0.7732**	**0.0241**	**0.682**

Table 3: OOD evaluation on unseen FlowBench datasets.

Test Dataset	PSNR↑	SSIM↑	LPIPS↓	ΔV ↓
FlowBench-nurbs (ID)	22.50	0.7732	0.0240	0.682
FlowBench-harmonics (OOD)	21.31	0.7981	0.0223	0.649
FlowBench-skeleton (OOD)	19.74	0.6609	0.0382	0.721

Table 4: Long-horizon prediction performance.

Horizon	w/o Physics Attn			Whole Model		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
50	18.43	0.6352	0.0429	19.09	0.6696	0.0402
100	17.62	0.6198	0.0471	18.73	0.6686	0.0410
150	16.29	0.5555	0.0557	17.39	0.6153	0.0467

The physics-based loss proves equally essential for dynamical fidelity. As shown in Table 2, excluding this loss drastically elevates ΔV from 0.682 to over 15, indicating severe accumulation of dynamic errors during long-term rollout, despite visually plausible outputs. In contrast, the full model maintains stable and accurate temporal evolution, closely matching the ground truth (Fig. 4). These results demonstrate that Physics Attention and physical regularization are complementary: the former enhances representation learning, while the latter enforces physical consistency, and their joint integration is crucial for accurate and reliable fluid prediction.

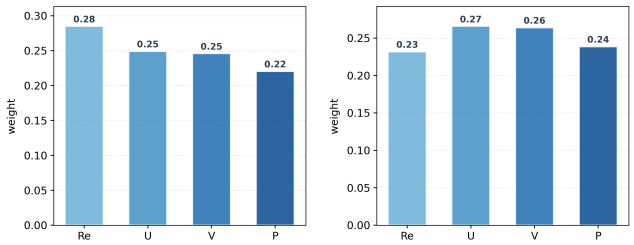
4.4 Generalization and Long-Horizon Analysis

OOD Generalization. To evaluate the generalization capability of PAWM beyond the training distribution, we conduct OOD experiments on unseen flow patterns from FlowBench. As shown in Table 3, PAWM maintains competitive prediction accuracy and physical consistency across different flow types. Although the performance slightly decreases on more challenging OOD scenarios, the model still achieves comparable visual quality and low physical errors, demonstrating its ability to capture transferable physical dynamics rather than overfitting to specific flow patterns.

Long-Horizon Prediction. We further evaluate the long-horizon prediction stability of PAWM under extended rollout horizons. As reported in Table 4, the full model consistently outperforms the variant without Physics Attention at 50, 100, and 150 prediction steps. In particular, PAWM exhibits slower degradation in performance as the prediction horizon increases, indicating that PAWM effectively preserves coherent flow structures and mitigates error accumulation during long-horizon dynamics prediction.

4.5 Explainability of Physics Representations

To evaluate whether the physics representations learned by PAWM are explainable and aligned with underlying physical principles, we perform a quantitative importance analysis to measure the contribution of each inferred physical factor to prediction performance. Specifically, given physical factors including the Reynolds number (Re), velocity components (U, V), and pressure (P), we first apply a reweighting strategy to reduce statistical dependencies and mitigate spurious correlations for more faithful attribution. We then estimate each factor's contribution by aggregating the magnitudes of network weights connecting the corresponding physics representations to the prediction module, and normalize the scores to obtain relative importance weights.



FlowBench ScalarFlow

Figure 5: Importance of learned physical factors across datasets. The bar charts show the relative importance weights of different physical factors.

The results are shown in Figure 5, revealing two key observations. First, PAWM consistently assigns high importance to velocity components across both datasets, which are closely related to fundamental fluid dynamics such as advection and turbulence. This suggests that the model captures physically meaningful factors rather than spurious statistical correlations. Second, the importance distribution adapts to different data characteristics. For FlowBench, where physical parameters are explicitly defined in a controlled simulator, PAWM places greater emphasis on the global parameter Re. In contrast, for ScalarFlow with real-world observations and noise, the model relies more on local velocity components (U, V). This adaptive behavior indicates that PAWM effectively exploits available physical information under different scenarios. Overall, these results demonstrate that the learned physical factors encode interpretable and meaningful dynamics, enabling both physically grounded representation learning and accurate prediction.

5 Conclusions

In this work, we present PAWM, a generative world model that learns structured physics representations from visual observations and integrates them into future prediction. By leveraging the Physics Attention module, PAWM learns structured physics representations, enabling the extraction of implicit physical laws underlying fluid dynamics. By disentangling and decorrelating physical factors, the model overcomes the challenge of their complex interactions, producing independent and interpretable representations that enable accurate prediction. Experiments on challenging fluid benchmarks show that PAWM consistently outperforms both physics-informed and generative world models for fluid dynamics. These results validate that learning disentangled and interpretable physics representations is essential for modeling complex dynamical systems, and that physics-learned mechanisms provide an effective solution for bridging visual prediction with physical understanding.

6 Acknowledgments

This work is supported by the National Key Research and Development Program of China under Grant No.2022ZD0115903.

References

[1] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. 2025. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985* (2025).

[2] George Keith Batchelor. 2000. *An introduction to fluid dynamics*. Cambridge university press.

[3] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261* (2018).

[4] Johannes Brandstetter, Daniel Worrall, and Max Welling. 2022. Message passing neural PDE solvers. *arXiv preprint arXiv:2202.03376* (2022).

[5] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. 2024. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*.

[6] Hong Chen, Xin Wang, Guanning Zeng, Yipeng Zhang, Yuwei Zhou, Feilin Han, Yaofei Wu, and Wenwu Zhu. 2025. Videodreamer: Customized multi-subject text-to-video generation with disen-mix finetuning on language-video foundation models. *IEEE Transactions on Multimedia* 27 (2025), 2875–2885.

[7] Hong Chen, Xin Wang, Yipeng Zhang, Yuwei Zhou, Zeyang Zhang, Siao Tang, and Wenwu Zhu. 2024. Disenstudio: Customized multi-subject text-to-video generation with disentangled spatial control. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 3637–3646.

[8] Anoop Cherian, Radu Corcodel, Siddarth Jain, and Diego Romeres. 2024. Llmphy: Complex physical reasoning using large language models and world models. *arXiv preprint arXiv:2411.08027* (2024).

[9] Silvia Chiappa, Sébastien Racaniere, Daan Wierstra, and Shakir Mohamed. 2017. Recurrent environment simulators. *arXiv preprint arXiv:1704.02254* (2017).

[10] Yitong Deng, Hong-Xing Yu, Jiajun Wu, and Bo Zhu. 2023. Learning vortex dynamics for fluid inference and prediction. *arXiv preprint arXiv:2301.11494* (2023).

[11] Marie-Lena Eckert, Kiwon Um, and Nils Thuerey. 2019. ScalarFlow: a large-scale volumetric data set of real-world scalar transport flows for computer animation and machine learning. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–16.

[12] Tongtong Feng, Xin Wang, Feilin Han, Leping Zhang, and Wenwu Zhu. 2024. U2udata: A large-scale cooperative perception dataset for swarm uavs autonomous flight. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 7600–7608.

[13] Tongtong Feng, Xin Wang, Feilin Han, Leping Zhang, and Wenwu Zhu. 2026. U2UData+: A scalable swarm UAVs autonomous flight dataset for embodied long-horizon tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 1792–1800.

[14] Tongtong Feng, Xin Wang, Yu-Gang Jiang, and Wenwu Zhu. 2025. Embodied ai: From llms to world models [feature]. *IEEE Circuits and Systems Magazine* 25, 4 (2025), 14–37.

[15] Wei Feng, Xin Wang, Yu-Wei Zhan, Yuwei Zhou, and Wenwu Zhu. 2026. Modularized Dynamic-Granularity Video LLM for Multi-Event Long Video Understanding. *arXiv preprint arXiv:2607.15778* (2026).

[16] Yue Gao, Hong-Xing Yu, Bo Zhu, and Jiajun Wu. 2025. FluidNexus: 3D fluid reconstruction and prediction from a single video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 26091–26101.

[17] Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. 2025. Force Prompting: Video Generation Models Can Learn and Generalize Physics-based Control Signals. *arXiv preprint arXiv:2505.19386* (2025).

[18] David Ha and Jürgen Schmidhuber. 2018. World models. *arXiv preprint arXiv:1803.10122* 2, 3 (2018).

[19] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. 2019. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603* (2019).

[20] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. 2023. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080* (2023).

[21] Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. 2014. An efficient approach for assessing hyperparameter importance. In *International conference on machine learning*. PMLR, 754–762.

[22] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. 2024. How Far is Video Generation from World Model? – A Physical Law Perspective. *arXiv preprint arXiv:2411.02385* (2024).

[23] Maximilian Karl, Maximilian Soelch, Justin Bayer, and Patrick Van der Smagt. 2016. Deep variational bayes filters: Unsupervised learning of state space models from raw data. *arXiv preprint arXiv:1605.06432* (2016).

[24] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. 2021. Physics-informed machine learning. *Nature Reviews Physics* 3, 6 (2021), 422–440.

[25] Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. 2023. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research* 24, 89 (2023), 1–97.

[26] Haoyang Li, Xin Wang, Xueling Zhu, Weigao Wen, and Wenwu Zhu. 2025. Disentangling invariant subgraph via variance contrastive estimation under distribution shifts. In *Forty-second International Conference on Machine Learning*.

[27] Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. 2024. Spacetime Gaussian Feature Splatting for Real-Time Dynamic View Synthesis. In *CVPR*. 8508–8520.

[28] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. 2020. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895* (2020).

[29] Wang Lin, Liyu Jia, Wentao Hu, Kaihang Pan, Zhongqi Yue, Wei Zhao, Jingyuan Chen, Fei Wu, and Hanwang Zhang. 2025. Reasoning physical video generation with diffusion timestep tokens via reinforcement learning. *arXiv preprint arXiv:2504.15932* (2025).

[30] Donggeun Park, Hyeonbin Moon, and Seunghwa Ryu. 2026. A self-correcting multi-agent LLM framework for language-based physics simulation and explanation. *npj Artificial Intelligence* 2, 1 (2026), 10.

[31] Stephen B Pope. 2001. Turbulent flows. *Measurement Science and Technology* 12, 11 (2001), 2020–2021.

[32] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics* 378 (2019), 686–707.

[33] Maziar Raissi, Alireza Yazdani, and George Em Karniadakis. 2020. Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science* 367, 6481 (2020), 1026–1030.

[34] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. 2020. Learning to simulate complex physics with graph networks. In *International conference on machine learning*. PMLR, 8459–8468.

[35] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. 2021. E (n) equivariant graph neural networks. In *International conference on machine learning*. PMLR, 9323–9332.

[36] Yu Shang, Xin Zhang, Yinzhou Tang, Lei Jin, Chen Gao, Wei Wu, and Yong Li. 2025. RoboScape: Physics-informed Embodied World Model. *arXiv preprint arXiv:2506.23135* (2025).

[37] Ronak Tali, Ali Rabeh, Cheng-Hau Yang, Mehdi Shadkhah, Samundra Karki, Abhisek Upadhyaya, Suriya Dhakshinamoorthy, Marjan Saadati, Soumik Sarkar, Adarsh Krishnamurthy, et al. 2024. Flowbench: A large scale benchmark for flow simulation over complex geometries. *arXiv preprint arXiv:2409.18032* (2024).

[38] Rona L Thompson, Luis Lassaletta, Prabir K Patra, Christopher Wilson, Kelley C Wells, Alicia Gressent, Ernest N Koffi, Martyn P Chipperfield, Wilfried Winiwarter, Eric A Davidson, et al. 2019. Acceleration of global N2O emissions seen from two decades of atmospheric inversion. *Nature Climate Change* 9, 12 (2019), 993–998.

[39] Chen Wang, Chuhao Chen, Yiming Huang, Zhiyang Dou, Yuan Liu, Jiatao Gu, and Lingjie Liu. 2025. Physctrl: Generative physics for controllable and physics-grounded video generation. *arXiv preprint arXiv:2509.20358* (2025).

[40] Ren Wang, Xin Wang, Tongtong Feng, Xinyue Gong, Guangyao Li, Yu-Wei Zhan, Qing Li, and Wenwu Zhu. 2025. Improving compositional generalization in cross-embodiment learning via mixture of disentangled prototypes. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 7162–7171.

[41] Xin Wang, Hong Chen, Zirui Pan, Yuwei Zhou, Chaoyu Guan, Lifeng Sun, and Wenwu Zhu. 2025. Automated disentangled sequential recommendation with large language models. *ACM Transactions on Information Systems* 43, 2 (2025), 1–29.

[42] Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. 2024. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 12 (2024), 9677–9696.

[43] Xin Wang, Shuyi Fan, Kun Kuang, and Wenwu Zhu. 2021. Explainable automated graph representation learning with hyperparameter importance. In *International conference on machine learning*. PMLR, 10727–10737.

[44] Xin Wang, Tongtong Feng, Huaping Liu, and Wenwu Zhu. 2026. *Autonomous Embodied AI: Towards Self-evolving Intelligence*. Springer Nature.

[45] Xin Wang, Zirui Pan, Hong Chen, and Wenwu Zhu. 2025. Divico: Disentangled visual token compression for efficient large vision-language model. *IEEE Transactions on Circuits and Systems for Video Technology* 36, 2 (2025), 1392–1405.

[46] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[47] Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. 2024. ivideogpt: Interactive videogpts are scalable world models. *Advances in Neural Information Processing Systems* 37 (2024), 68082–68119.

[48] Hong-Xing Yu, Yang Zheng, Yuan Gao, Yitong Deng, Bo Zhu, and Jiajun Wu. 2023. Inferring hybrid neural fluid fields from videos. *Advances in Neural Information Processing Systems* 36 (2023), 63595–63608.

[49] Yu-Wei Zhan, Xin Wang, Pengzhe Mao, Tongtong Feng, Ren Wang, and Wenwu Zhu. 2026. Modularagent: A task-aware modular framework for joint optimization of multimodal large language models and world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8087–8096.

[50] Chenyu Zhang, Daniil Cherniavskii, Antonios Tragoudaras, Antonios Vozikis, Thijmen Nijdam, Derck WE Prinzhorn, Mark Bodracska, Nicu Sebe, Andrii Zadaianchuk, and Efstratios Gavves. 2025. Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments. *arXiv preprint arXiv:2504.02918* (2025).

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

[52] Weipu Zhang, Gang Wang, Jian Sun, Yetian Yuan, and Gao Huang. 2023. Storm: Efficient stochastic transformer based world models for reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2023), 27147–27166.

[53] Yipeng Zhang, Xin Wang, Hong Chen, Chenyang Qin, Yibo Hao, Hong Mei, and Wenwu Zhu. 2025. ScenarioDiff: text-to-video generation with dynamic transformations of scene conditions. *International Journal of Computer Vision* 133, 7 (2025), 4909–4922.

[54] Yipeng Zhang, Xin Wang, Hong Chen, and Wenwu Zhu. 2023. Adaptive disentangled transformer for sequential recommendation. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*. 3434–3445.

[55] Pengyu Zhao, Jinan Xu, Ning Cheng, Huiqi Hu, Xue Zhang, Xiuwen Xu, Zijian Jin, Fandong Meng, Jie Zhou, and Wenjuan Han. 2025. Bridging the reality gap: A benchmark for physical reasoning in general world models with various physical phenomena beyond mechanics. *Expert Systems with Applications* 270 (2025), 126548.

[56] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 2024. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631* (2024).