Latent Knowledge-Guided Video Diffusion for Scientific Phenomena Generation from a Single Initial Frame

Qinglong Cao[1,2], Xirui Li[1], Ding Wang[2], Chao Ma[1], Yuntian Chen[*], Xiaokang Yang[1]

[1] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China
[2] Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, China

Abstract

Video diffusion models have achieved impressive results in natural scene generation, yet they struggle to generalize to scientific phenomena such as fluid simulations and meteorological processes, where underlying dynamics are governed by scientific laws. These tasks pose unique challenges, including severe domain gaps, limited training data, and the lack of descriptive language annotations. To handle this dilemma, we extracted the latent scientific phenomena knowledge and further proposed a fresh framework that teaches video diffusion models to generate scientific phenomena from a single initial frame. Particularly, static knowledge is extracted via pre-trained masked autoencoders, while dynamic knowledge is derived from pre-trained optical flow prediction. Subsequently, based on the aligned spatial relations between the CLIP vision and language encoders, the visual embeddings of scientific phenomena, guided by latent scientific phenomena knowledge, are projected to generate the pseudo-language prompt embeddings in both spatial and frequency domains. By incorporating these prompts and fine-tuning the video diffusion model, we enable the generation of videos that better adhere to scientific laws. Extensive experiments on both computational fluid dynamics simulations and real-world typhoon observations demonstrate the effectiveness of our approach, achieving superior fidelity and consistency across diverse scientific scenarios.

Introduction

Diffusion models, originally inspired by nonequilibrium thermodynamics (Sohl-Dickstein et al. 2015), have achieved remarkable success in various video generation tasks (Saharia et al. 2022a; Kazerouni et al. 2022; He et al. 2025; Zhang et al. 2025; He et al. 2024). The Video Diffusion Model (VDM) (Ho et al. 2022b) pioneered the extension of diffusion models to video generation, yielding promising results. Subsequent works, such as Make-a-Video (Singer et al. 2022), extended Text-to-Image (T2I) models to Text-to-Video (T2V) generation through super-resolution strategies. Further advancements, including Video LDM (Blattmann et al. 2023b) and Imagen Video (Ho et al. 2022a), incorporated temporal consistency layers and cascaded generation architectures. Efficient training paradigms such as TAV (Wu et al. 2023) have also been explored to reduce training overhead. Most recently, Stable Video Diffusion (SVD) (Blattmann et al. 2023a) and CogVideoX (Yang et al. 2024), inspired by the principles of stable diffusion, has demonstrated superior performance through large-scale training.

Building upon these advancements, video diffusion models have begun to attract attention as general-purpose world simulators (Yang et al. 2023), which must internalize fundamental world laws to extrapolate beyond training data. However, recent studies (Kang et al. 2024) have shown that these models struggle to capture generalizable scientific principles. When applied to the scientific phenomena or systems governed by real-world scientific laws—such as dynamic fluids and typhoons—the performance of current video diffusion models degrades significantly. This is largely due to a pronounced domain gap and practical challenges like data scarcity and the absence of language annotations, which collectively hinder their ability to generate consistent dynamics.

To address these challenges, there is a growing need for video generation models that not only retain the generative power of diffusion architectures but also incorporate latent scientific phenomena knowledge to ensure alignment with real-world dynamics. To this end, we propose a novel framework that moves beyond traditional natural vision-based synthesis and toward latent knowledge-grounded scientific phenomena generation. Specifically, in scientific domains such as fluid mechanics and meteorology, realistic video generation requires models capable of generalizing from limited data availability, scientific consistency, and the absence of natural language prompts. Motivated by this need, our approach integrates latent scientific phenomenon knowledge into video diffusion models via parameter-efficient fine-tuning (Figure 1), enabling more consistent and plausible generation under data-constrained scenarios.

In real data-scarce scientific environments, we assume access to only the first frame of a video. This raises a critical question: how can one extract meaningful latent scientific phenomenon knowledge from a single frame? Self-supervised learning approaches (He et al. 2022; Zhang et al. 2022; Cao et al. 2025) have demonstrated strong capabilities in capturing generalized visual representations, suggesting that their learned embeddings encode generalized latent knowledge. However, typical self-supervised augmentations such as noise injection, rotation, and color jitter are inappropriate for scientific data, where each pixel corresponds

[*]Corresponding Author. Email: ychen@eitech.edu.cn

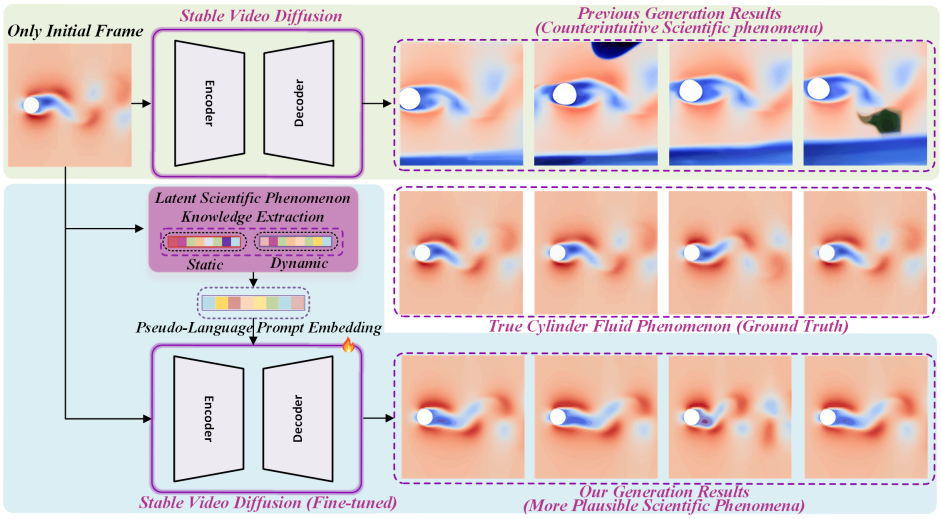

Figure 1: Our approach integrates latent scientific phenomenon knowledge into video diffusion models via parameter-efficient fine-tuning, enabling more consistent and plausible generation under data-constrained scenarios.

to a field variable with scientific meaning and directional force. These operations would violate the underlying scientific laws. To avoid such violations, we could only employ a Masked Autoencoder (MAE) to extract static latent scientific phenomenon knowledge. By reconstructing masked regions from unaltered observations, the MAE is implicitly encouraged to learn the governing scientific laws. While this provides a static understanding, scientific systems also involve dynamic behaviors. To capture dynamic knowledge, we leverage pre-trained optical flow predictors, which model the apparent motion patterns in scientific dynamic processes.

Although scientific knowledge can improve generative quality, existing video diffusion models often rely on language prompt embeddings to control generation. However, scientific phenomena are hard to articulate with descriptive natural language, making prompt formulation infeasible. Thus, we generate *pseudo-language prompt embeddings* from the available visual data and scientific knowledge. This is enabled by the CLIP architecture, which offers well-aligned visual and language embedding spaces. We project the visual embeddings—enriched by latent scientific knowledge—into the language embedding space to produce pseudo-prompts. To capture complex multimodal dependencies, we adopt quaternion networks (Shi et al. 2024; Cao et al. 2024) to perform the projection, which has shown effectiveness in cross-modal modeling of CLIP. Moreover, since frequency-domain information is key in representing scientific phenomena (Zhang, Chen, and Chen 2024; Chen et al. 2025), we further inject spectral information to enrich the prompts.

Finally, we incorporate the pseudo-language prompt embeddings into video diffusion models through parameter-efficient fine-tuning, generating more scientifically realistic videos without relying on large-scale annotations or textual descriptions. Contributions are summarized as follows:

- We present the first framework for teaching video diffusion models to generate more scientifically plausible phenomena under practical constraints, including limited data and the absence of language annotations, by embedding latent scientific phenomenon knowledge.
- We introduce a novel static-dynamic decomposition strategy that separates scientific phenomenon knowledge into static and dynamic components: static properties are preserved through masked autoencoding (MAE) to encourage scientifically consistent spatial representation, while dynamic properties are extracted via optical flow to capture temporal evolution patterns.
- To overcome the challenge of lacking explicit language supervision, we design a quaternion-based projection module that transforms static and dynamic cues into pseudo-language prompt embeddings across spatial and frequency domains, enabling semantic-guided diffusion without natural language input.
- Extensive experiments are performed on both numerical simulations and real-world observations of the scientific phenomena to validate the proposed method, demonstrating its promising performance across diverse scenarios.

Related Work

Diffusion Models for Image Generation. Diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020), rooted in nonequilibrium thermodynamic theory, have demonstrated remarkable success across a range of image synthesis tasks (Croitoru et al. 2023; Zhou et al. 2024; Gao, Cao, and Chen 2025). Ho *et al.* (Ho and Salimans 2022) proposed classifier-free guidance to improve conditional generation quality, while Karras *et al.* (Karras et al. 2022) investigated architectural refinements to advance generative fidelity. To accelerate inference, Salimans *et al.* (Salimans and Ho 2022) introduced progressive distillation and alternative parameterizations. Beyond synthesis, diffusion-based

methods have been applied to related domains including denoising (Kawar et al. 2022), super-resolution (Li et al. 2022), and inpainting (Lugmayr et al. 2022). Leveraging the capabilities of large-scale vision-language models (Radford et al. 2021), text-guided diffusion approaches (Nichol et al. 2021; Ramesh et al. 2022) have improved image controllability.

Diffusion Models for Video Generation. Compared to static images, video generation presents additional challenges due to temporal dynamics and motion continuity. Recent work has explored both adapting image diffusion models and training specialized architectures (Xing et al. 2024a; Jiang et al. 2024). Specifically, VideoLDM (Blattmann et al. 2023b) extends latent diffusion to video through temporal tuning, while TAV (Wu et al. 2023) adopts parameter-efficient strategies for text-to-video synthesis. Ho *et al.* (Ho et al. 2022b) introduced a 3D extension of image diffusion for low-resolution video, and Stable Video Diffusion (SVD) (Blattmann et al. 2023a) later proposed a 3D U-Net to generate higher-resolution sequences. CogVideoX (Yang et al. 2024) utilizes expert transformer blocks to jointly process image and text tokens and reaches state-of-the-art video generation performance. Building upon SVD, Text2Video-Zero (Khachatryan et al. 2023) introduces cross-frame attention to enable zero-shot generation, while VidToMe (Li et al. 2024) refines inter-frame coherence via token-based fusion. Additionally, ControlVideo (Zhang et al. 2023), inspired by ControlNet (Zhang, Rao, and Agrawala 2023), allows controllable video generation through structural priors.

Preliminaries

Latent Diffusion Models. Diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Song et al. 2020) generate data by reversing a forward noising process:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and $\{\alpha_t\}$ is a noise schedule. Latent diffusion models (LDMs) (Rombach et al. 2022; Saharia et al. 2022b) operate in a compressed latent space via an autoencoder $z = \mathcal{E}(x)$, improving efficiency and quality. A UNet-based denoiser ϵ_θ is trained to predict noise:

$$\min_\theta \mathbb{E}_{z,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|^2\right], \tag{2}$$

where c is optional conditioning (e.g., text). Our method builds upon video LDMs such as Stable Video Diffusion (Blattmann et al. 2023a) and CogVideoX (Yang et al. 2024).

Quaternion Neural Networks. A quaternion is $Q = r + x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ with $r, x, y, z \in \mathbb{R}$. Quaternion multiplication (Hamilton product) preserves 3D rotational and spatial structure. In quaternion neural networks (QNNs) (Parcollet et al. 2018), layers compute:

$$Q_{\text{out}} = \alpha(W \otimes Q), \tag{3}$$

where W is a quaternion weight and α applies an activation (e.g., ReLU) component-wise:

$$\alpha(Q) = f(r) + f(x)\mathbf{i} + f(y)\mathbf{j} + f(z)\mathbf{k}. \tag{4}$$

QNNs inherently model multi-dimensional correlations, making them suitable for spatially aligned multimodal fusion—unlike real-valued networks that often lose geometric structure.

Method

Video Generation Pipeline

As illustrated in Figure 2, our framework generates scientific phenomenon videos from a single initial frame via two stages: (1) latent knowledge extraction and (2) knowledge-guided generation. From raw training videos (without annotations), we extract static knowledge using a Masked Autoencoder (MAE) and dynamic motion via an Optical Flow Predictor (OFP) trained on pseudo ground-truth flows (Farnebäck 2003a). Both modules are frozen after training. Given an initial frame f_0, MAE and OFP produce static and dynamic embeddings E_M and E_O. Meanwhile, a pre-trained VAE and CLIP encoder yield conditional features E_c and F_o. F_o is concatenated with E_M and E_O, then projected into quaternion space to form pseudo-language prompt embeddings E_L, which are injected into the UNet's cross-attention layers. The UNet is fine-tuned using LoRA (Hu et al. 2021). During training, all frames of the target video $\mathcal{V}_p$ are encoded into T latents, noised to form N_{input}, and the UNet learns denoising via iterative reconstruction. At inference, frozen MAE and OFP re-extract E_M and E_O from f_0; E_c is combined with sampled noise N_{input} to form conditioned input N'_{input}, enabling video synthesis guided by both visual cues and E_L.

Latent Knowledge Extraction

Given K training scientific phenomena $[\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_K]$, we decompose each video into individual frames to construct an image set $[I_1^{V_1}, I_2^{V_1}, \ldots, I_N^{V_K}]$. To extract static latent scientific phenomenon knowledge, each image is partitioned into non-overlapping, regular patches to facilitate processing. Then, since self-supervised learning methods like SimCLR or DINO rely on augmentations (e.g., rotation, noise addition) that would violate the underlying scientific law, we adopt a Masked Autoencoder (MAE), which operates on non-overlapping image patches. A subset of patches P_{vis} is randomly retained as visible tokens, while the remaining P_{masked} are withheld to promote spatial prediction. A Vision Transformer (ViT) (Dosovitskiy 2020) serves as both encoder and decoder of the MAE, trained to reconstruct the masked regions from visible context:

$$P_{\text{masked}} = \text{MAE}(P_{\text{vis}}). \tag{5}$$

The reconstruction is optimized via mean squared error (MSE). By reconstructing masked regions from unaltered observations, the MAE is implicitly encouraged to learn the governing scientific laws. For dynamic latent knowledge, we utilize optical flow to capture inter-frame motion. Pseudo ground-truth flows are computed via the Farnebäck algorithm (Farnebäck 2003a), and used to supervise an Optical Flow Predictor (OFP), which also adopts a ViT backbone. Given an input frame I_{input}, the OFP is trained to estimate the flow f_{op}:

$$f_{\text{op}} = \text{OFP}(I_{\text{input}}). \tag{6}$$

The MSE loss again guides the training, enabling the OFP to encode dynamic evolution patterns. Upon convergence, we extract static and dynamic knowledge embeddings E_M and E_O of the only acquirable initial frame from the decoder outputs of the frozen MAE and OFP, respectively.

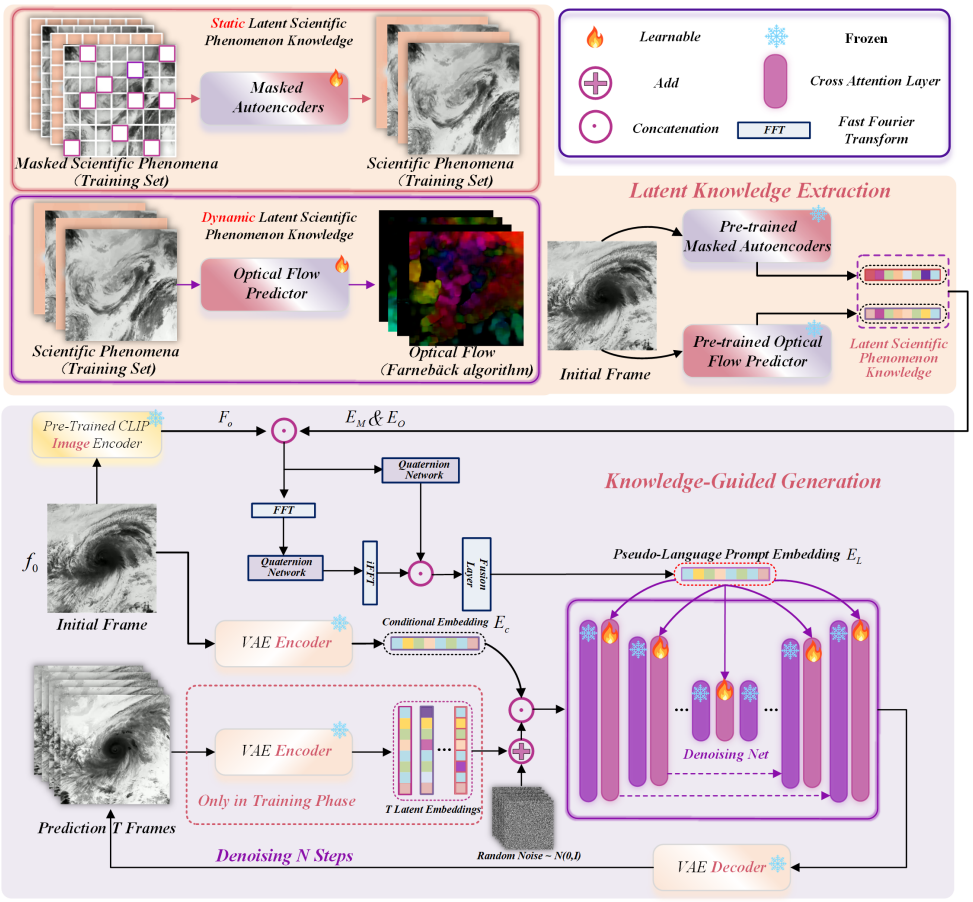

Figure 2: Overview of our proposed method. Using the MAE and optical flow prediction to extract latent physical phenomenon knowledge. Projecting CLIP vision features guided by latent physical phenomenon knowledge to obtain pseudo-language prompt embeddings. Incorporating these embeddings to generate more physically plausible physical phenomena.

Knowledge-Guided Generation

We first encode the initial frame f_0 using a pre-trained VAE and CLIP image encoder to obtain conditional features E_c and visual features F_o. Unlike traditional video diffusion models that predominantly rely on natural language prompts, our method targets more realistic scenarios where describing scientific phenomena in natural language is often intractable. While prior models commonly use CLIP's text encoder to generate language embeddings, we instead leverage CLIP's robust vision-language alignment to derive pseudo-language prompt embeddings directly from visual features and extracted latent knowledge embeddings. To this end, we incorporate static and dynamic latent physical knowledge, denoted as E_M and E_O, respectively. Inspired by the success of prior works (Shi et al. 2024; Cao et al. 2024) that utilize quaternion networks (Parcollet et al. 2018) for modeling multimodal CLIP's relations, we adopt quaternion networks to derive pseudo-language prompt embeddings. We first apply linear layers $[L_{d1}, L_{d2}, L_{d3}]$ to obtain projected features:

$$\hat{F}_o = L_{d1}(F_o), \hat{E}_M = L_{d2}(E_M), \hat{E}_O = L_{d3}(E_O). \quad (7)$$

We randomly initialize learnable text embeddings T_L and organize the four components—text, vision, static knowledge, and dynamic knowledge—along the orthogonal axes of the quaternion latent space:

$$Q_l = T_L + \hat{F}_o\mathbf{i} + \hat{E}_M\mathbf{j} + \hat{E}_O\mathbf{k}. \quad (8)$$

As frequency-domain information are critical for modeling scientific dynamics (Zhang, Chen, and Chen 2024; Chen et al. 2025), we extend this formulation using the Fast Fourier Transform (FFT):

$$Q_l^{FFT} = T_L^{FFT} + FFT(\hat{F}_o)\mathbf{i} + FFT(\hat{E}_M)\mathbf{j} + FFT(\hat{E}_O)\mathbf{k}, \quad (9)$$

where T_L^{FFT} denotes learnable embeddings in the frequency domain. These quaternion latent vectors are processed by different quaternion networks Q_t and Q_t^{FFT} to generate pseudo-language prompt embeddings:

$$E_L^S = Q_t([T_L, \hat{F}_o, \hat{E}_M, \hat{E}_O]),$$
$$E_L^F = Q_t^{FFT}([T_L^{FFT}, FFT(\hat{F}_o), FFT(\hat{E}_M), FFT(\hat{E}_O)]). \quad (10)$$

A fusion network L_{fuse}, consisting of two linear layers, combines spatial and frequency-domain features to form the final

prompt embeddings:

$$E_L = L_{\text{fuse}}([E_L^S, E_L^F]). \tag{11}$$

These embeddings E_L are then injected into each cross-attention layer of the stable video diffusion UNet via parameter-efficient fine-tuning. Let $\varphi_i(z_t) \in \mathbb{R}^{N \times d_\epsilon^i}$ be the projected UNet hidden state at layer i and $\tau_\theta(E_L) \in \mathbb{R}^{M \times d_\tau}$ be the prompt representation. The attention mechanism is:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V. \tag{12}$$

We apply LoRA (Hu et al. 2021) to the projection matrices:

$$\begin{aligned} Q &= \text{LoRA}(W_Q^{(i)}) \cdot \varphi_i(z_t), \\ K &= \text{LoRA}(W_K^{(i)}) \cdot \tau_\theta(E_L), \\ V &= \text{LoRA}(W_V^{(i)}) \cdot \tau_\theta(E_L), \end{aligned} \tag{13}$$

where $W_Q^{(i)}, W_K^{(i)}, W_V^{(i)}$ are frozen pretrained weights with trainable LoRA adapters. In summary, the video diffusion model D_t, combined with the CLIP image encoder E_{img} and the quaternion projection process Q applied in both spatial and frequency domains, enables robust video generation:

$$\begin{aligned} E_L &= Q(E_{img}(f_0), \text{MAE}(f_0), \text{OFP}(f_0)), \\ \mathcal{V}_p &= D_t([\text{DDIM}_{\text{sample}}, f_0], E_L), \end{aligned} \tag{14}$$

where f_0 is the initial frame and $\mathcal{V}_p$ is the generated video.

Experiments

Experiment Setting

We adopt Stable Video Diffusion (SVD) (Blattmann et al. 2023a) and CogVideoX (Yang et al. 2024) as the baseline and evaluate methods on both simulated fluid dynamics and real-world typhoon datasets. We employ a mix of numerical accuracy metrics and physically grounded evaluation criteria to ensure a comprehensive assessment.

Datasets and Evaluation Metrics. We generate four simulated fluid scenarios using a computational fluid dynamics (CFD) toolkit: Rayleigh-Bénard Convection (RBC), Cylinder flow, DamBreak, and DepthCharge. To validate performance on real-world data, we further randomly select four typhoon events from true Typhoon dataset (Kitamoto et al. 2023), identified by observation timestamps: 202001, 202009, 202102, and 202204. Formally, in each phenomenon test, we construct 10 videos, using 9 for fine-tuning and 1 for testing. We employ eight metrics to evaluate both numerical accuracy and scientific fidelity: RMSE, SSIM (Wang et al. 2004), Stream Function Error (SFE) (Kundu, Cohen, and Dowling 2015; Farnebäck 2003b), Smoothness Error (SE) (Jeong et al. 2022), Gradient Smoothness (GS), Continuity Score (CS) (Kundu, Cohen, and Dowling 2015), Q-Criterion Error (QCE) (Hunt, Wray, and Moin 1988), and Vorticity Error (VE) (Jeong and Hussain 1995). RMSE and SSIM measure per-pixel error and structural similarity, respectively. The remaining metrics are tailored for physical dynamics: SFE tests streamline accuracy, SE quantifies propagation smoothness, GS and CS (without ground truth) assess spatial-temporal consistency, and QCE and VE measure vortex preservation and vorticity accuracy. Notably, QCE and VE apply only to incompressible 2D flow fields from simulation data. Higher SSIM and lower scores on the other metrics indicate better performance.

Implementation Details. Our model leverages the pre-trained SVD, the pre-trained CogVideoX, and the CLIP ViT-B/16 image encoder for visual conditioning. Fine-tuning is performed using SGD with a learning rate of 2×10^{-4} for 15 epochs on NVIDIA A100 GPUs. A DDIM sampler (Song, Meng, and Ermon 2020) with 50 steps is used during inference. The LoRA adaptation rank is set to 4, and frames are generated at a resolution of 512×512. The MAE and OFP are trained separately with a learning rate of 1×10^{-3}, batch size 60, and for 50 epochs each.

Experimental Results

Qualitative Evaluation. To visually assess the performance, we compare it against recent advanced approaches. Qualitative results are shown in Figure 3 (simulation fluid dynamic) and Figure 4 (real-world typhoon). In line with prior work (Wu et al. 2023; Xing et al. 2024b), we use the initial frame as a conditioning signal by concatenating it with the input noise. On simulated fluid data (Figure 3), baseline models such as SVD without latent knowledge often generate unrealistic artifacts and hallucinations, even after fine-tuning. For example, in the DamBreak scenario, generated frames depict incorrect fluid behavior, including unnatural stillness or upward motion. By contrast, our approach, guided by latent scientific knowledge, produces phenomena that align better with underlying scientific principles. A similar trend is observed in the real-world typhoon dataset (Figure 4). SVD frequently produces hallucinated content that violates scientific plausibility. While methods like LoRA and TAV partially improve stability, they still introduce implausible distortions—such as transforming typhoon clouds into textures resembling snow-covered mountains or leather. In contrast, our model demonstrates a more accurate interpretation of input frame and generates temporally coherent sequences that reflect physically meaningful progression, highlighting the benefit of embedding latent scientific priors.

Quantitative Evaluation. We further evaluate our method using both numerical and physics-aware scientific metrics, as summarized in Table 1. Overall, our method outperforms across most scenarios. On simulation datasets, our method achieves the best and the second-best performance in most metrics, while our method based on CogVideoX does not achieve top results in two scientific metrics, the results on other metrics still demonstrate its superiority. More importantly, the substantial improvements in scientific metrics achieved by our method, often by an order of magnitude, confirm its ability to generate sequences that better conform to fluid dynamics. On the more complex typhoon dataset, our model continues to deliver state-of-the-art results across nearly all metrics. The only exception is the CS metric, which measures internal consistency without access to ground truth. Here, our performance is slightly weaker, likely due to the residual bias from pre-trained diffusion models toward nature-image coherence. When fine-tuned with limited data and physically grounded priors, this bias may not be fully cor-

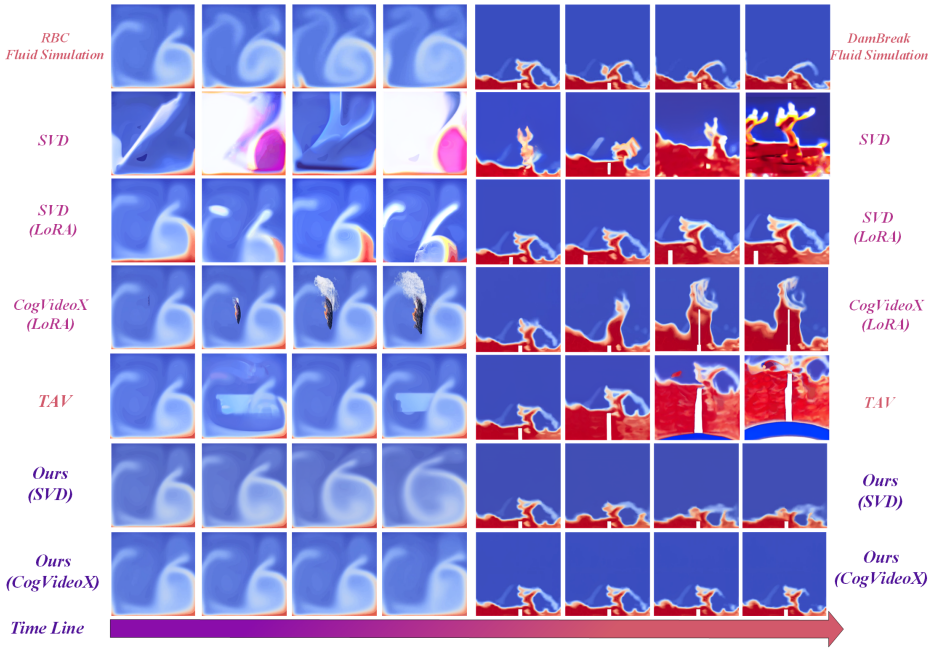

Figure 3: Qualitative results in fluid simulation datasets. Our method, guided by latent physical knowledge, produces phenomena more consistent with physical laws.

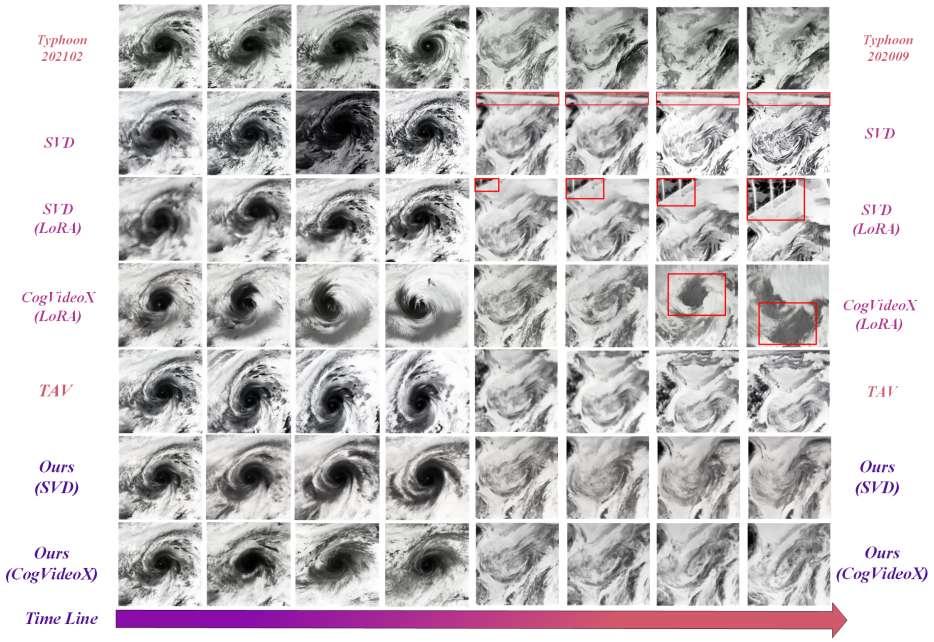

Figure 4: Qualitative comparisons in true typhoon dataset. The red box denotes some hallucinations.

rected—pointing to a promising direction for future research.

Ablation Study

To further assess the effectiveness of individual components in our framework, we conduct a series of ablation studies on the Cylinder fluid simulation dataset. The results are summarized in Table 2. We systematically remove key modules—such as static or dynamic scientific phenomenon knowledge and quaternion-based modeling—to evaluate their respective contributions. The ablation results demonstrate that each component plays a critical role. In particular, incorporating both static and dynamic latent scientific priors,

	RMSE↓	SSIM↑	SFE↓	SE↓	GS↓	CS↓	QCE↓	VE↓	RMSE↓	SSIM↑	SFE↓	SE↓	GS↓	CS↓
	RBC fluid simulation								***Typhoon 202001***					
SVD	9.323	0.773	170.48	1.893	4.826	12.46	0.075	0.029	9.702	0.365	1575.8	2.942	20.26	**16.66**
LoRA(SVD)	8.876	0.831	82.159	0.609	3.670	5.773	0.018	0.004	9.642	0.395	2733.6	7.240	17.32	31.21
LoRA(CogX)	8.885	0.871	222.46	0.971	5.2284	10.29	0.068	0.010	9.909	0.411	2966.7	14.39	10.73	51.50
TAV	8.457	0.863	15.470	0.229	2.731	1.706	0.003	6e-4	9.696	0.367	3297.7	15.53	18.94	37.20
SimDA	8.786	0.857	14.895	0.256	3.012	1.716	0.003	8e-4	9.710	0.375	2846.3	14.67	18.37	34.61
Ours(SVD)	**8.490**	**0.902**	**4.4569**	**0.099**	**2.081**	**0.819**	**0.001**	**2e-4**	**9.351**	**0.446**	**606.36**	**1.346**	**10.90**	**19.19**
Ours(CogX)	**7.933**	**0.911**	**4.2683**	**0.064**	**1.893**	**0.459**	**7e-4**	**2e-4**	**9.314**	**0.502**	**1063.5**	**2.531**	**9.167**	28.27
	Cylinder fluid simulation								***Typhoon 202009***					
SVD	8.081	0.871	399.88	1.453	4.832	15.83	0.211	0.022	10.29	0.308	3443.9	7.828	18.94	32.44
LoRA(SVD)	7.206	0.906	185.33	0.760	3.921	9.221	0.046	0.014	10.16	0.367	4113.3	2.793	12.02	29.51
LoRA(CogX)	6.151	0.928	126.25	0.654	3.712	6.405	0.043	0.008	9.983	0.371	2415.8	4.881	12.58	42.51
TAV	7.710	0.891	178.10	0.633	4.420	6.693	**0.027**	0.007	10.17	0.360	2812.0	**2.455**	9.752	21.09
SimDA	7.682	0.883	180.45	0.667	4.012	**6.721**	0.031	**0.006**	10.15	0.357	2965.1	2.832	9.212	25.61
Ours(SVD)	**5.852**	**0.948**	**62.456**	**0.303**	**2.631**	**4.681**	**0.013**	**0.003**	**9.890**	**0.411**	**1742.5**	**2.486**	**8.160**	**20.25**
Ours(CogX)	**5.946**	**0.944**	**124.81**	**0.608**	**2.931**	8.508	0.040	**0.006**	**9.954**	**0.381**	**1992.8**	2.553	**7.058**	**13.06**
	DamBreak fluid simulation								***Typhoon 202102***					
SVD	8.189	0.743	1388.6	7.686	9.483	51.69	0.539	0.205	10.08	0.250	3290.9	8.953	17.99	**20.35**
LoRA(SVD)	5.812	0.786	392.37	1.233	6.471	15.97	0.257	0.021	9.981	0.319	6264.3	17.94	16.48	49.02
LoRA(CogX)	5.901	0.791	311.89	1.649	4.357	17.65	0.122	0.036	9.965	0.357	2179.5	4.353	12.36	42.72
TAV	7.792	0.686	1543.6	2.572	10.08	34.33	0.205	0.140	10.07	0.298	3440.6	**3.429**	14.46	25.95
SimDA	7.679	0.691	1243.5	2.667	9.892	36.55	0.281	0.161	10.06	0.308	2898.5	3.509	14.01	26.07
Ours(SVD)	**4.921**	**0.862**	**158.36**	**0.701**	**3.271**	**12.32**	**0.032**	**0.012**	**9.931**	**0.371**	**1111.4**	**2.514**	**9.262**	**23.87**
Ours(CogX)	**4.492**	**0.884**	**88.493**	**0.231**	**1.219**	**1.198**	**0.009**	**0.003**	**9.903**	**0.374**	**1149.4**	4.391	**10.25**	33.72
	DepthCharge fluid simulation								***Typhoon 202204***					
SVD	9.745	0.690	899.84	6.716	11.04	48.52	0.488	0.236	9.760	0.358	2624.2	7.883	15.78	39.09
LoRA(SVD)	7.916	0.733	2213.9	7.405	13.19	75.91	1.300	0.482	9.677	0.338	1860.4	4.660	14.52	**22.66**
LoRA(CogX)	4.775	0.812	212.26	1.366	5.492	**12.04**	0.113	0.050	9.828	0.357	2937.1	4.785	14.86	39.46
TAV	7.679	0.734	1307.7	5.777	13.04	61.66	0.587	0.256	9.649	0.359	2357.3	7.702	14.67	32.52
SimDA	7.682	0.745	1298.2	5.867	14.01	67.21	0.576	0.301	9.686	0.347	2657.1	7.919	15.12	31.89
Ours(SVD)	**4.672**	**0.832**	**418.18**	**0.875**	**8.026**	12.83	**0.050**	**0.025**	**9.610**	**0.406**	**1462.8**	**2.830**	**10.71**	**29.48**
Ours(CogX)	**3.990**	**0.872**	**94.474**	**0.238**	**1.179**	**1.454**	**0.011**	**0.003**	**9.614**	**0.396**	**1608.8**	**3.702**	**11.39**	31.29

Table 1: Quantitative evaluation on fluid simulation dataset (Left) and true typhoon dataset (Right). SVD: Stable Video Diffusion (Blattmann et al. 2023a). CogX: CogVideoX (Yang et al. 2024) LoRA(SVD): SVD+LoRA (Hu et al. 2021). LoRA(CogX): SVD+CogX (Yang et al. 2024). TAV: Tune-A-video (Wu et al. 2023). SimDA: Simple Diffusion Adapter (Xing et al. 2024b).

	RMSE↓	SSIM↑	SFE↓	SE↓	GS↓	CS↓	QCE↓	VE↓
w/o SPK	6.021	0.938	73.368	0.373	2.821	4.966	0.028	0.004
w/o DPK	5.983	0.937	68.446	0.401	2.791	4.786	0.021	0.003
w/o SDM	6.423	0.928	76.563	0.411	3.012	5.128	0.043	0.004
w/o FDM	5.923	0.941	67.463	0.343	2.721	4.923	0.016	0.003
w/o QN	5.913	0.940	73.161	0.398	2.823	4.823	0.018	0.004
Ours	**5.852**	**0.948**	**62.456**	**0.303**	**2.631**	**4.681**	**0.013**	**0.003**

Table 2: Ablation study on different components. w/o: without, SPK: static physical phenomenon knowledge, DPK: dynamic physical phenomenon knowledge, SDM: spatial domain modeling, FDM: frequency domain modeling, QN: quaternion network. w/o QN denotes replacing quaternion network with linear network. The best result is bolded.

Rank	RMSE↓	SSIM↑	SFE↓	SE↓	GS↓	CS↓	QCE↓	VE↓
2	6.132	0.928	78.213	0.376	2.798	4.945	0.034	0.004
4	**5.852**	**0.948**	**62.456**	**0.303**	**2.631**	**4.681**	**0.013**	**0.003**
8	6.018	0.936	68.356	0.351	2.756	5.084	0.027	0.003

Table 3: Ablation study of LoRA rank.

Projection Methods	RMSE↓	SSIM↑	SFE↓	SE↓	GS↓	CS↓	QCE↓	VE↓
Linear Network	5.913	0.940	73.161	0.398	2.823	4.823	0.018	0.004
Cross Attention	5.921	0.941	70.012	0.358	2.705	4.754	0.016	0.004
Quaternion Network	**5.852**	**0.948**	**62.456**	**0.303**	**2.631**	**4.681**	**0.013**	**0.003**

Table 4: Ablation study of projection method.

as well as jointly modeling spatial and frequency representations via quaternion networks, significantly enhances the model's ability to generate scientifically plausible sequences consistent with fundamental laws of motion. We also explore the effect of varying the LoRA rank, with results reported in Table 3. The performance is sensitive to this hyperparameter, indicating that careful selection is essential. For all main experiments, we use the rank that achieves the most stable and consistent results. Finally, we evaluate different projection strategies for generating pseudo-language prompt embeddings. As shown in Table 4, quaternion network-based projection yields better performance than alternative designs. This highlights the advantage of quaternion representations in capturing cross-domain semantics more effectively.

Conclusion

We present a novel framework that teaches video diffusion models with latent scientific phenomenon knowledge to enable more plausible generation from a single initial frame. Unlike existing approaches that rely heavily on natural vision priors and language prompts, our method leverages self-supervised learning to extract static knowledge and dynamic knowledge via optical flow prediction. We further propose a quaternion-based projection mechanism to convert latent knowledge into pseudo-language prompt embeddings, capturing both spatial and frequency-domain semantics. These embeddings are seamlessly integrated into video diffusion models through parameter-efficient fine-tuning, allowing the model to internalize latent knowledge and generalize to scientific phenomena. Extensive experiments on both simulated and real-world datasets demonstrate that our method potentially improves the scientific plausibility of generated videos, marking a step forward in bridging the gap between generative video models and scientific phenomena.

References

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023a. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv:2311.15127*.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023b. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Cao, Q.; Chen, Y.; Lu, L.; Sun, H.; Zeng, Z.; Yang, X.; and Zhang, D. 2025. Generalized domain prompt learning for accessible scientific vision-language models. *Nexus*, 2(2).

Cao, Q.; Xu, Z.; Chen, Y.; Ma, C.; and Yang, X. 2024. Domain prompt learning with quaternion networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26637–26646.

Chen, Y.; Wang, D.; Feng, D.; Tian, G.; Gupta, V.; Cao, R.; Wan, M.; and Chen, S. 2025. Three-dimensional spatiotemporal wind field reconstruction based on LiDAR and multi-scale PINN. *Applied Energy*, 377: 124577.

Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*.

Farnebäck, G. 2003a. Two-frame motion estimation based on polynomial expansion. In *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*, 363–370. Springer.

Farnebäck, G. 2003b. Two-Frame Motion Estimation Based on Polynomial Expansion. In *Scandinavian Conference on Image Analysis*.

Gao, J.; Cao, Q.; and Chen, Y. 2025. Auto-regressive moving diffusion models for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 16727–16735.

He, H.; Xu, Y.; Guo, Y.; Wetzstein, G.; Dai, B.; Li, H.; and Yang, C. 2024. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*.

He, H.; Yang, C.; Lin, S.; Xu, Y.; and Wei, M. 2025. Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. *arXiv preprint arXiv:2503.10592*, 19.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022a. Imagen video: High definition video generation with diffusion models. *arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv:2207.12598*.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022b. Video diffusion models. *Advances in Neural Information Processing Systems*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv:2106.09685*.

Hunt, J. C. R.; Wray, A. A.; and Moin, P. 1988. Eddies, streams, and convergence zones in turbulent flows. In *Studying Turbulence Using Numerical Simulation Databases, 2*.

Jeong, J.; and Hussain, F. 1995. On the identification of a vortex. *Journal of Fluid Mechanics*.

Jeong, J.; Lin, J. M.; Porikli, F. M.; and Kwak, N. 2022. Imposing Consistency for Optical Flow Estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jiang, Y.; Wu, T.; Yang, S.; Si, C.; Lin, D.; Qiao, Y.; Loy, C. C.; and Liu, Z. 2024. Videobooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kang, B.; Yue, Y.; Lu, R.; Lin, Z.; Zhao, Y.; Wang, K.; Huang, G.; and Feng, J. 2024. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*.

Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*.

Kawar, B.; Elad, M.; Ermon, S.; and Song, J. 2022. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*.

Kazerouni, A.; Aghdam, E. K.; Heidari, M.; Azad, R.; Fayyaz, M.; Hacihaliloglu, I.; and Merhof, D. 2022. Diffusion models for medical image analysis: A comprehensive survey. *arXiv:2211.07804*.

Khachatryan, L.; Movsisyan, A.; Tadevosyan, V.; Henschel, R.; Wang, Z.; Navasardyan, S.; and Shi, H. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Kitamoto, A.; Hwang, J.; Vuillod, B.; Gautier, L.; Tian, Y.; and Clanuwat, T. 2023. Digital Typhoon: Long-term Satellite Image Dataset for the Spatio-Temporal Modeling of Tropical Cyclones. In *NeurIPS 2023 Datasets and Benchmarks*.

Kundu, P. K.; Cohen, I. M.; and Dowling, D. R. 2015. *Fluid Mechanics*. Elsevier, 6 edition.

Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q.; and Chen, Y. 2022. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*.

Li, X.; Ma, C.; Yang, X.; and Yang, M.-H. 2024. Vidtome: Video token merging for zero-shot video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv:2112.10741*.

Parcollet, T.; Ravanelli, M.; Morchid, M.; Linarès, G.; Trabelsi, C.; De Mori, R.; and Bengio, Y. 2018. Quaternion recurrent neural networks. *arXiv:1806.04418*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022a. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022b. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*.

Salimans, T.; and Ho, J. 2022. Progressive distillation for fast sampling of diffusion models. *arXiv:2202.00512*.

Shi, B.; Xu, Z.; Jia, S.; and Ma, C. 2024. Prompt Learning with Quaternion Networks. In *The Twelfth International Conference on Learning Representations*.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv:2209.14792*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv:2010.02502*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv:2011.13456*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Xing, Z.; Dai, Q.; Hu, H.; Wu, Z.; and Jiang, Y.-G. 2024a. Simda: Simple diffusion adapter for efficient video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Xing, Z.; Dai, Q.; Hu, H.; Wu, Z.; and Jiang, Y.-G. 2024b. SimDA: Simple Diffusion Adapter for Efficient Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yang, M.; Du, Y.; Ghasemipour, K.; Tompson, J.; Schuurmans, D.; and Abbeel, P. 2023. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 1(2): 6.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*.

Zhang, D.; Chen, Y.; and Chen, S. 2024. Filtered partial differential equations: a robust surrogate constraint in physics-informed deep learning framework. *Journal of Fluid Mechanics*, 999: A40.

Zhang, F.; Zhou, T.; Yao, J.; Zhang, Y.; Tsang, I. W.; and Wang, Y. 2025. Decouple before align: Visual disentanglement enhances prompt tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Zhang, Y.; Wei, Y.; Jiang, D.; Zhang, X.; Zuo, W.; and Tian, Q. 2023. Controlvideo: Training-free controllable text-to-video generation. *arXiv:2305.13077*.

Zhou, T.; Xia, W.; Zhang, F.; Chang, B.; Wang, W.; Yuan, Y.; Konukoglu, E.; and Cremers, D. 2024. Image segmentation in foundation model era: A survey. *arXiv preprint arXiv:2408.12957*.

Supplementary Material for Latent Knowledge-Guided Video Diffusion for Scientific Phenomena Generation from a Single Initial Frame

Criteria

Root Mean Squared Error (RMSE). RMSE provides a comprehensive measure of pixel-level deviations between the generated and real videos. A lower RMSE indicates that the generated video is closer to the real video in terms of pixel values. The formula is as follows:

$$\mathrm{RMSE} = \frac{1}{N}\sum_{t=1}^{N}\sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(X_i^{\mathrm{real}} - X_i^{\mathrm{gen}}\right)^2}, \tag{15}$$

where N is the number of frames, and M is the total number of pixels per frame. $< \cdot >^{\mathrm{real}}$ and $< \cdot >^{\mathrm{gen}}$ are true values and generated values respectively, as well as X_i are pixel values.

Structural Similarity Index Measure (SSIM). SSIM (Wang et al. 2004) assesses structural similarity between generated and real video frames, focusing on aspects like brightness, contrast, and texture. This metric helps ensure that the generated video maintains the structural integrity of the real video. Higher SSIM values indicate greater similarity. The formula is:

$$\mathrm{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{16}$$

where μ_x and μ_y are the mean values, σ_x and σ_y are the variances of frames x and y, σ_{xy} is the covariance, and C_1 and C_2 are constants to stabilize the division.

Additionally, to evaluate the generated video's quality, we analyze the velocity fields derived from optical flow based on physical quantities such as velocity, temperature, and volume fraction in the original video. This allows us to apply metrics that assess the alignment of the generated video with the original's physical characteristics. These metrics, grounded in convective dynamics, are suitable for analyzing the flow consistency across different physical quantities, determining if they exhibit smooth and realistic motion over time.

We compute the optical flow fields using the Farneback method (Farnebäck 2003b), a dense optical flow estimation technique that approximates each pixel's local neighborhood as a polynomial function. This approach facilitates precise, pixel-wise motion estimates, making it ideal for capturing fine-grained movements in fluid-related fields.

In Farneback's method, each pixel neighborhood is represented by a quadratic polynomial:

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c, \tag{17}$$

where $\mathbf{x}$ is the pixel location, $\mathbf{A}$ represents second-order terms, $\mathbf{b}$ denotes first-order terms, and c is a constant. The displacement vector $\mathbf{d}$, representing pixel-wise motion, is derived from changes in these terms across frames:

$$\mathbf{d} = -\left(\mathbf{A} + \mathbf{A}^\top\right)^{-1}(\mathbf{b} - \mathbf{b}'). \tag{18}$$

Using a pyramidal approach, the Farneback algorithm captures motion at multiple scales, refining the motion field at each level for accurate displacement estimation. The resulting optical flow vectors $\mathbf{u} = (u, v)$ provide the basis for calculating metrics that evaluate the generated video's fidelity to the physical properties of the original.

Stream Function Error (SFE). The stream function, ψ, represents a scalar field from which the velocity components of a two-dimensional incompressible flow can be derived, where $\partial\psi/\partial y = u$ and $\partial\psi/\partial x = -v$ (Kundu, Cohen, and Dowling 2015). For optical flow fields, ψ is computed via numerical integration:

$$\psi(x, y) = \int u\,\mathrm{d}y - \int v\,\mathrm{d}x. \tag{19}$$

The Stream Function Error (SFE) between generated and real data is calculated as:

$$\mathrm{SFE} = \frac{1}{N}\sum_{t=1}^{N}\sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\psi_i^{\mathrm{real}} - \psi_i^{\mathrm{gen}}\right)^2}. \tag{20}$$

SFE assesses the dynamic consistency of the generated flow field with the real flow by comparing streamline characteristics. Lower SFE values indicate that the generated flow better replicates advective properties, providing insights into the physical accuracy and quality of the generated video.

Smoothness Error (SE). Smooth changes in velocity generally reflect the asymptotic behavior of physical phenomena, while abrupt fluctuations may be unrealistic (Jeong et al. 2022). Temporal smoothness in optical flow velocity can capture the steady propagation characteristics of the underlying physical quantities. Smoothness Error (SE) measures the timewise smoothness of both the generated and real flow fields, providing insight into physical continuity over time. SE is defined as:

$$\mathrm{SE} = \frac{1}{N-1}\sum_{t=1}^{N-1}\sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\Delta u_i^{\mathrm{gen}} - \Delta u_i^{\mathrm{real}}\right)^2}, \tag{21}$$

where $\Delta u_i = u_{i,t+1} - u_{i,t}$ represents the velocity change across consecutive time intervals. A lower SE value indicates greater temporal smoothness in the generated flow, reflecting the essential physical continuity of the quantity.

Gradient Smoothness (GS). Gradient Smoothness evaluates the temporal smoothness of the gradient field in the generated frames, capturing the physical continuity of spatial features across time steps. The formula is:

$$\mathrm{GS} = \frac{1}{N-1}\sum_{t=1}^{N-1} \sqrt{\frac{\sum_{i=1}^{M}\left(\left(\frac{\partial X_i^{t+1}}{\partial x} - \frac{\partial X_i^{t}}{\partial x}\right)^2 + \left(\frac{\partial X_i^{t+1}}{\partial y} - \frac{\partial X_i^{t}}{\partial y}\right)^2\right)}{2M}}. \tag{22}$$

This metric is computed solely from the generated frames and reflects the smoothness of changes in the gradient field over time.

Continuity Score (CS). Continuity Score measures the spatial continuity of the generated optimal flow field by calculating the divergence of each frame in the generated sequence. The formula is:

$$\mathrm{CS} = \frac{1}{N}\sum_{t=1}^{N}\sqrt{\frac{1}{M}\sum_{i=1}^{M}(\nabla\cdot\mathbf{u}_i)^2}, \tag{23}$$

where $\nabla\cdot\mathbf{u} = \partial u/\partial x + \partial v/\partial y$ represents the divergence of the generated velocity field $\mathbf{u} = (u, v)$ at each pixel i and time step t. Lower CS values indicate better preservation of spatial continuity in the generated field, reflecting adherence to mass conservation principles.

Q-Criterion Error (QCE). The Q-Criterion identifies vortices within a flow field by balancing rotational and strain rates (Hunt, Wray, and Moin 1988). It's calculated as:

$$Q = \frac{1}{2}\left(\|\mathbf{\Omega}\|^2 - \|\mathbf{S}\|^2\right), \tag{24}$$

where $\mathbf{\Omega} = 1/2\left(\nabla\mathbf{u} - (\nabla\mathbf{u})^T\right)$ is rotation tensor and $\mathbf{S} = 1/2\left(\nabla\mathbf{u} + (\nabla\mathbf{u})^T\right)$ is strain tensor. In the two-dimensional plane, the calculation can be simplified as:

$$Q = \frac{1}{2}\left(-\left(\frac{\partial u}{\partial x}\right)^2 - \left(\frac{\partial v}{\partial y}\right)^2 - 2\frac{\partial u}{\partial y}\frac{\partial v}{\partial x}\right), \tag{25}$$

with error:

$$\mathrm{QCE} = \frac{1}{N}\sum_{t=1}^{N}\sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(Q_i^{\mathrm{gen}} - Q_i^{\mathrm{real}}\right)^2}. \tag{26}$$

Lower QCE suggests that the generated flow retains rotational structures similar to the real flow.

Vorticity Error (VE). Vorticity represents rotational effects, such as eddies and vortices, which are vital for processes like mixing, energy transfer, and turbulence development (Jeong and Hussain 1995). In two-dimensional flow, local rotation is quantified by vorticity, defined as:

$$\omega = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}. \tag{27}$$

The Vorticity Error (VE) measures the accuracy of the generated flow's rotational dynamics compared to the real flow and is calculated as:

$$\mathrm{VE} = \frac{1}{N}\sum_{t=1}^{N}\sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\omega_i^{\mathrm{gen}} - \omega_i^{\mathrm{real}}\right)^2}. \tag{28}$$

Lower VE values indicate closer alignment in rotational characteristics, showing that the generated flow accurately replicates the real flow's dynamics.

CFD simulations

Rayleigh-Bénard convection is a thermally-driven natural convection flow caused by temperature differences, which induces ascending and descending currents and is commonly used to study turbulence and heat transfer. Cylinder flow simulates the fluid behavior around an obstacle (e.g., a cylinder), producing wake and vortex structures to examine vortex dynamics in bluff-body flows. DamBreak flow models the propagation of free-surface waves and fluid flow following a sudden dam collapse, illustrating gravity-driven, nonlinear free-surface behavior. DepthCharge flow simulates the high-pressure shockwave and bubble expansion dynamics following an underwater explosion, focusing on the transient changes at the water-air interface.

The numerical simulations for these flows were conducted using the open-source software OpenFOAM, which employs the finite volume method for discretizing partial differential equations (PDEs). The Rayleigh-Bénard convection was modeled with a RANS turbulence approach, while the other flows were treated as laminar. A second-order central difference scheme was applied to the diffusion terms, and appropriate discretization schemes were selected for the convection terms according to the specific physical requirements of each flow. Temporal discretization used a second-order backward implicit method for Rayleigh-Bénard convection, while a first-order explicit Euler scheme was applied for the others. The simulations were conducted over 400 time steps, with results interpolated from nonuniform grids to a uniform spatial resolution of 512 × 512. The processed frames were then compiled into 16 videos.

Qualitative Comparisons.

To provide a more comprehensive analysis of the performance of our proposed method, we present additional qualitative comparisons. The visualization results are shown in Figure 5 and Figure 6, with Figure 5 illustrating the comparisons on the fluid simulation dataset and Figure 6 showing the results on the true typhoon dataset. These examples consistently highlight the superior performance of our method, showcasing its robustness and improved alignment with physical realism across various scenarios.

Provided Videos.

To directly assess the performance of our proposed method, we also provide a comprehensive review video, titled "Comparison Results.mp4", which summarizes the results. The video and additional detailed videos corresponding to specific experiments can be found in the supplementary.

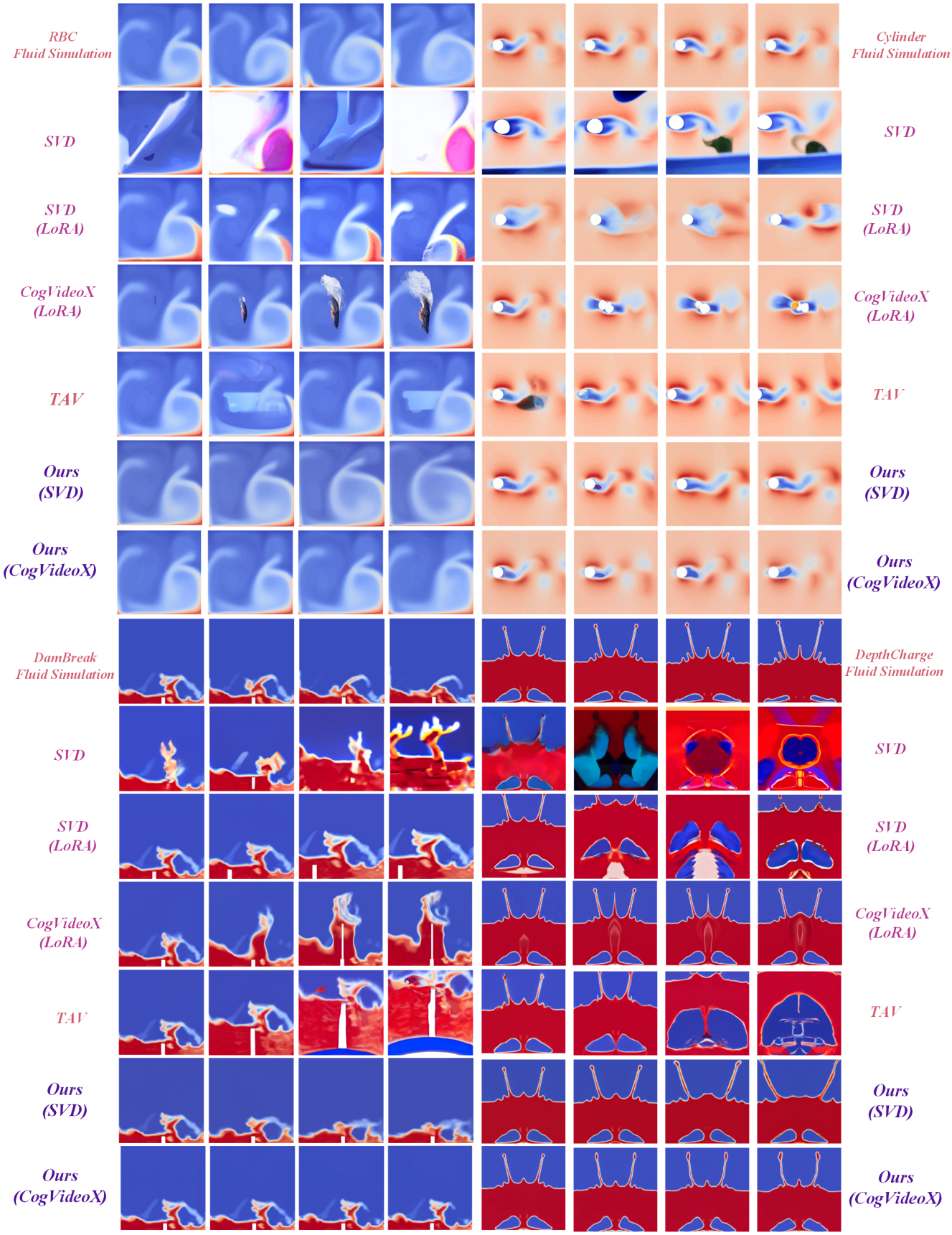

Figure 5: Qualitative comparisons in fluid simulation dataset. Though incorporating physical phenomenon knowledge, our method generates rational phenomena that exhibit better alignment with physical laws.

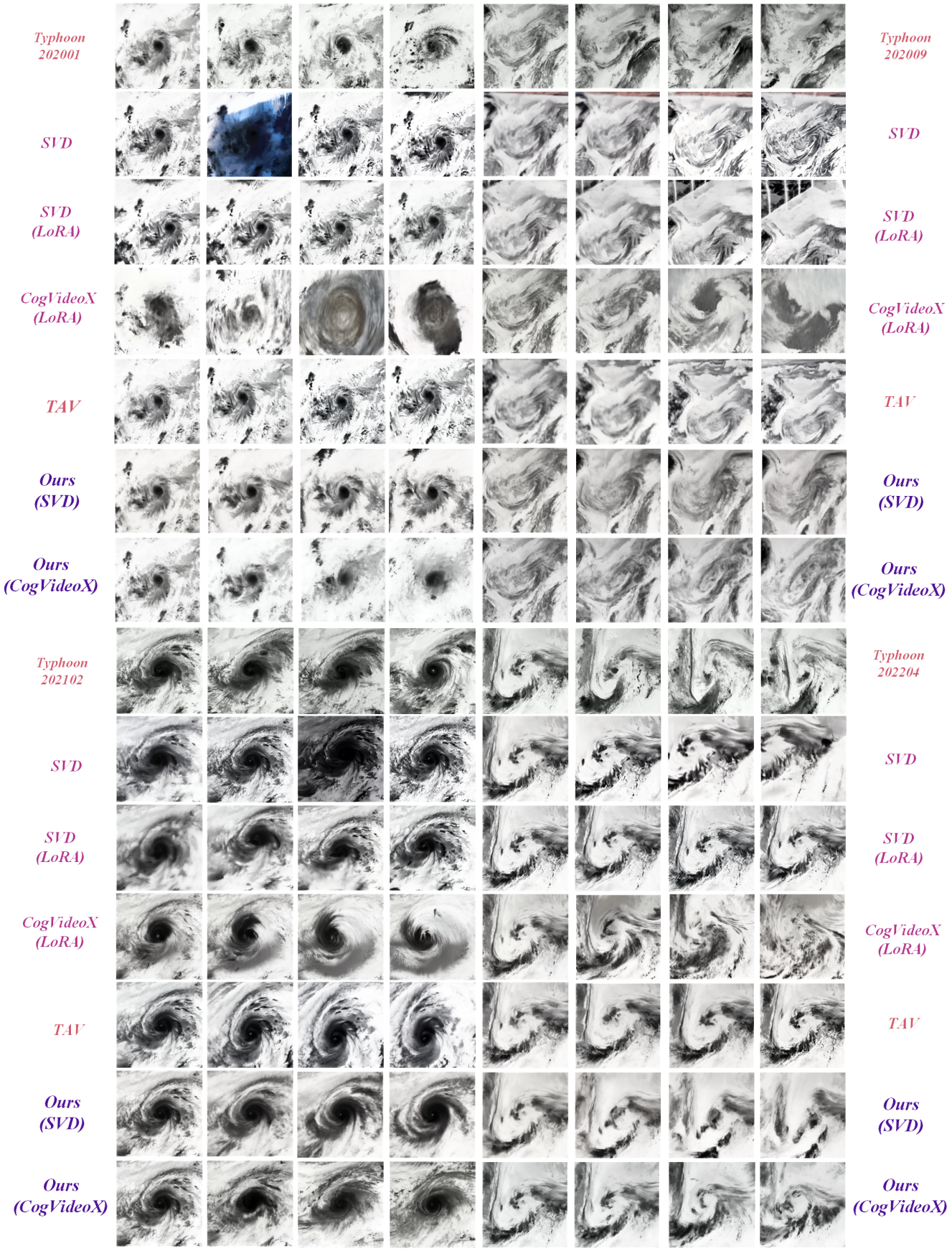

Figure 6: Qualitative comparisons in true typhoon dataset. Though incorporating physical phenomenon knowledge, our method generates rational phenomena that exhibit better alignment with physical laws.