

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

Qinglong Cao<sup>1,2</sup>, Xirui Li<sup>1</sup>, Ding Wang<sup>1,2</sup>, Chao Ma<sup>1</sup>, Yuntian Chen<sup>2\*</sup>, Xiaokang Yang<sup>1</sup>

<sup>1</sup> Shanghai Jiao Tong University, Shanghai, China

<sup>2</sup> 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 video generation (Saharia et al. 2022a; Kazerooni et al. 2022; He et al. 2025; Zhang et al. 2025; He et al. 2024). Video Diffusion Models (VDM) (Ho et al. 2022b) first extended diffusion processes to videos, followed by works such as Make-a-Video (Singer et al. 2022), Video LDM (Blattmann et al. 2023b), and Imagen Video (Ho et al. 2022a), which improved temporal consistency and generative quality via cascaded architectures. Efficient training strategies like TAV (Wu et al. 2023) further reduce computational cost, while recent large-scale systems, including Stable Video Diffusion (SVD) (Blattmann et al. 2023a) and CogVideoX (Yang et al. 2024) demonstrate strong performance through stable diffusion principles.

Beyond natural video synthesis, diffusion models have been explored as general-purpose world simulators (Yang et al. 2023). However, their ability to generalize real scientific laws remains limited. Recent studies (Kang et al. 2024) show that when applied to fluid dynamics or meteorological systems, current video diffusion models degrade significantly due to domain gaps, data scarcity, and the absence of language annotations.

To overcome these challenges, we aim to build video generators that preserve the strengths of diffusion models while incorporating latent scientific phenomenon knowledge. Scientific domains such as fluid mechanics and meteorology demand models capable of generalizing under limited data, enforcing physical consistency, and operating without natural-language prompts. We therefore integrate latent scientific phenomenon knowledge into video diffusion models via parameter-efficient tuning (Fig. 1), enabling more realistic generation under data constraints.

In scientific datasets, we often have access only to the first frame of each video, raising the challenge of extracting meaningful knowledge from a single observation. Self-supervised methods (He et al. 2022; Zhang et al. 2022; Cao et al. 2025) indicate that learned embeddings can capture generalized structure, but typical augmentations violate scientific semantics. Thus, we use a Masked Autoencoder (MAE) to extract *static* latent scientific knowledge while preserving physical validity. To capture *dynamic* information, we employ pre-trained optical flow predictors that model motion patterns underlying scientific processes.

Although scientific knowledge enhances representations, video diffusion models typically rely on language prompt embeddings for controllability. Since scientific phenomena cannot be easily described with natural language, we construct *pseudo-language prompt embeddings* by projecting scientific visual embeddings into the CLIP language space. The quaternion-based projection network (Cao et al. 2024a; Shi et al. 2024; Cao et al. 2024b) is able to capture complex multimodal dependencies, and frequency-domain cues (Zhang, Chen, and Chen 2024; Chen et al. 2025) are further adopted to enhance scientific representational capacity.

Finally, we inject the pseudo-language prompts into video diffusion models through parameter-efficient fine-tuning, enabling scientifically plausible video generation without textual supervision. Contributions are summarized as follows:

\*Corresponding Author. Email: ychen@eitech.edu.cn  
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

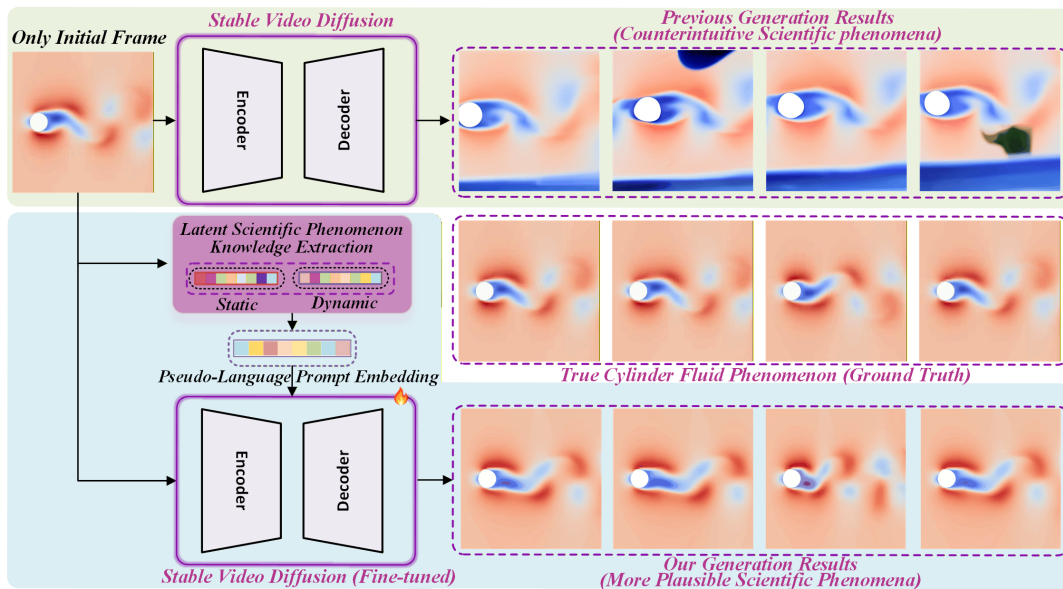


Figure 1: Our approach integrates latent scientific phenomenon knowledge into video diffusion models via parameter-efficient fine-tuning, enabling more consistent and plausible generation under data-constrained scenarios.

- We propose the first framework that embeds latent scientific phenomenon knowledge into video diffusion models to generate scientifically plausible dynamics under limited data and absent language annotations.
- We introduce a static–dynamic decomposition strategy: MAE-based static knowledge for spatial consistency and optical-flow-based dynamic knowledge for temporal evolution.
- We design a quaternion-based projection module that transforms scientific cues into pseudo-language prompt embeddings across spatial and frequency domains, enabling semantic-guided diffusion without textual input.
- Extensive experiments on numerical simulations and real-world observations demonstrate strong improvements in scientific realism and generalization.

## Related Work

**Diffusion Models for Image Generation.** Diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020), grounded in nonequilibrium thermodynamics, have achieved strong performance in image synthesis (Croitoru et al. 2023; Zhou et al. 2024; Gao, Cao, and Chen 2025). Classifier-free guidance (Ho and Salimans 2022) and architectural improvements (Karras et al. 2022) further boost fidelity, while progressive distillation (Salimans and Ho 2022) accelerates sampling. Diffusion models have also been applied to denoising (Kawar et al. 2022), super-resolution (Li et al. 2022), and inpainting (Lugmayr et al. 2022). With the rise of large-scale vision–language models (Radford et al. 2021), text-conditioned diffusion methods (Nichol et al. 2021; Ramesh et al. 2022) provide improved controllability.

**Diffusion Models for Video Generation.** Video generation introduces temporal dependencies beyond static images. Early approaches adapt image diffusion models or design video-specific architectures (Xing et al. 2024a; Jiang et al. 2024). VideoLDM (Blattmann et al. 2023b) incorporates temporal tuning in latent diffusion, and TAV (Wu et al. 2023) enables parameter-efficient text-to-video adaptation. Ho et al. (Ho et al. 2022b) proposed a 3D extension for low-resolution synthesis, later advanced by SVD (Blattmann et al. 2023a) with a 3D U-Net for high-resolution sequences. CogVideoX (Yang et al. 2024) employs expert transformer blocks to jointly process visual and textual tokens, achieving state-of-the-art quality. Building on SVD, Text2Video-Zero (Khachatryan et al. 2023) introduces cross-frame attention for zero-shot video generation, while VidToMe (Li et al. 2024) enhances temporal coherence via token fusion. ControlVideo (Zhang et al. 2023), inspired by ControlNet (Zhang, Rao, and Agrawala 2023), enables controllable 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:

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1 - \alpha_t}\epsilon, \quad (1)$$

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  $\epsilon_\theta$  is trained to predict noise:

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

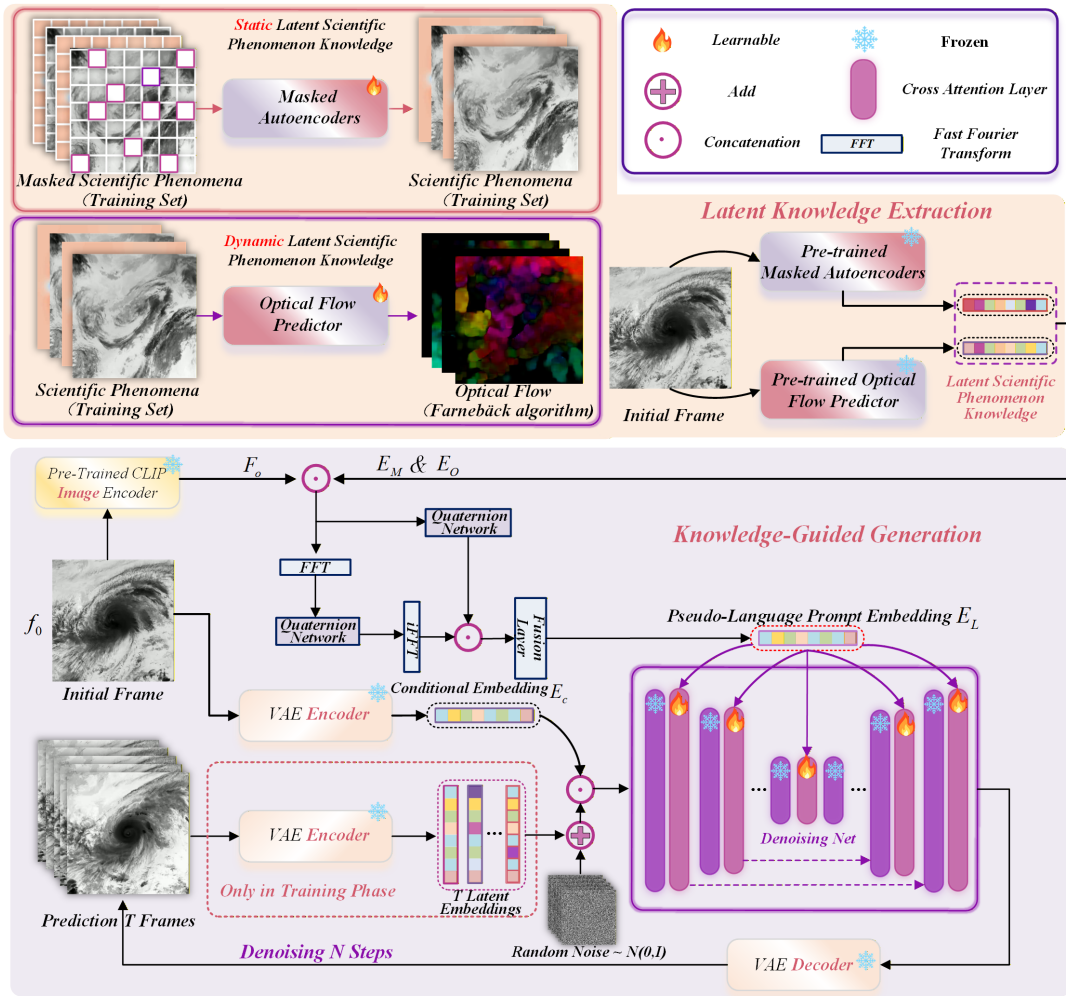


Figure 2: Overview of our proposed method. Using the MAE and optical flow prediction to extract latent physical phenomenon knowledge. Projecting CLIP vision features guided by latent physical phenomenon knowledge to obtain pseudo-language prompt embeddings. Incorporating these embeddings to generate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

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:

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

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

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

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 (Farneback 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  $N_{\text{input}}$ , 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  $N_{\text{input}}$  to form conditioned input  $N'_{\text{input}}$ , enabling video synthesis guided by both visual cues and  $E_L$ .

### Latent Knowledge Extraction

Given  $K$  training scientific phenomena  $[\mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_K]$ , we decompose each video into individual frames to construct an image set  $[I_1^{V_1}, I_2^{V_1}, \dots, I_N^{V_K}]$ . 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  $P_{\text{vis}}$  is randomly retained as visible tokens, while the remaining  $P_{\text{masked}}$  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:

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

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  $I_{\text{input}}$ , the OFP is trained to estimate the flow  $f_{\text{op}}$ :

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

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.

### 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. 2024b) 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:

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

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):

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

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:

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

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

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

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:

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

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

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

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  $E_{\text{img}}$  and the quaternion projection process  $Q$  applied in both spatial and frequency domains, enables robust video generation:

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

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

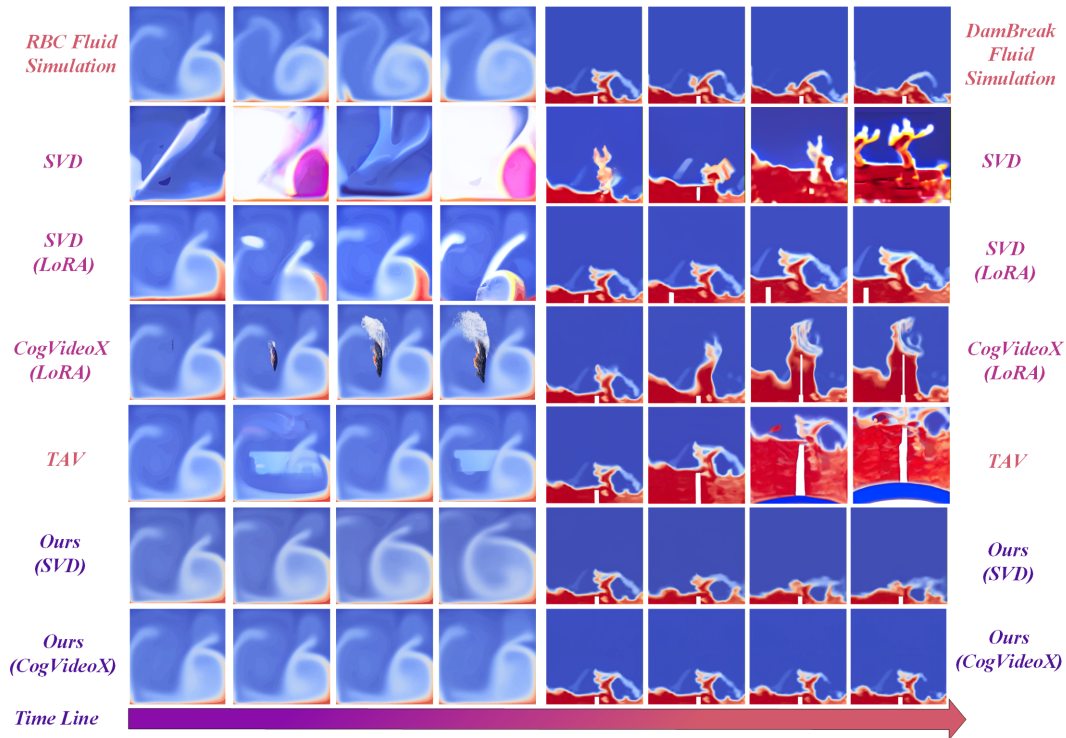


Figure 3: Qualitative results in fluid simulation datasets. Our method, guided by latent physical knowledge, produces phenomena more consistent with physical laws.

## 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 CFD toolkit: Rayleigh–Bénard Convection (RBC), Cylinder Flow, DamBreak, and DepthCharge. To assess real-world generalization, we further select four typhoon events from the Digital Typhoon dataset (Kitamoto et al. 2023): 202001, 202009, 202102, and 202204. For each phenomenon, we construct 10 videos, using 9 for fine-tuning and 1 for testing.

We evaluate performance with eight metrics covering numerical accuracy and physical fidelity: RMSE, SSIM (Wang et al. 2004), Stream Function Error (SFE) (Kundu, Cohen, and Dowling 2015; Farneback 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 capture per-pixel and structural accuracy; the remaining metrics measure scientific consistency—SFE evaluates streamline correctness, SE motion smoothness, GS and CS spatial–temporal coherence (without ground truth), and QCE/VE

assess vortex and vorticity preservation (applicable to 2D incompressible flows). Higher SSIM and lower values on other metrics indicate better results.

**Implementation Details.** We build on pre-trained SVD, CogVideoX, and CLIP ViT-B/16 for visual conditioning. Fine-tuning uses SGD with a learning rate of  $2 \times 10^{-4}$  for 15 epochs on NVIDIA A100 GPUs. Inference employs a DDIM sampler (Song, Meng, and Ermon 2020) with 50 steps. LoRA rank is set to 4, and videos are generated at  $512 \times 512$  resolution. The MAE and optical flow predictor (OFP) are trained separately with a learning rate of  $1 \times 10^{-3}$ , batch size 60, 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.

	RMSE↓	SSIM↑	SFE↓	SE↓	GS↓	CS↓	QCE↓	VE↓	RMSE↓	SSIM↑	SFE↓	SE↓	GS↓	CS↓
<i>RBC fluid simulation</i>									<i>Typhoon 202001</i>					
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	<b>16.66</b>
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)	<b>8.490</b>	<b>0.902</b>	<b>4.4569</b>	<b>0.099</b>	<b>2.081</b>	<b>0.819</b>	<b>0.001</b>	<b>2e-4</b>	<b>9.351</b>	<b>0.446</b>	<b>606.36</b>	<b>1.346</b>	<b>10.90</b>	<b>19.19</b>
Ours(CogX)	<b>7.933</b>	<b>0.911</b>	<b>4.2683</b>	<b>0.064</b>	<b>1.893</b>	<b>0.459</b>	<b>7e-4</b>	<b>2e-4</b>	<b>9.314</b>	<b>0.502</b>	<b>1063.5</b>	<b>2.531</b>	<b>9.167</b>	28.27
<i>Cylinder fluid simulation</i>									<i>Typhoon 202009</i>					
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	<b>0.027</b>	0.007	10.17	0.360	2812.0	<b>2.455</b>	9.752	21.09
SimDA	7.682	0.883	180.45	0.667	4.012	<b>6.721</b>	0.031	<b>0.006</b>	10.15	0.357	2965.1	2.832	9.212	25.61
Ours(SVD)	<b>5.852</b>	<b>0.948</b>	<b>62.456</b>	<b>0.303</b>	<b>2.631</b>	<b>4.681</b>	<b>0.013</b>	<b>0.003</b>	<b>9.890</b>	<b>0.411</b>	<b>1742.5</b>	<b>2.486</b>	<b>8.160</b>	<b>20.25</b>
Ours(CogX)	<b>5.946</b>	<b>0.944</b>	<b>124.81</b>	<b>0.608</b>	<b>2.931</b>	8.508	0.040	<b>0.006</b>	<b>9.954</b>	<b>0.381</b>	<b>1992.8</b>	2.553	<b>7.058</b>	<b>13.06</b>
<i>DamBreak fluid simulation</i>									<i>Typhoon 202102</i>					
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	<b>20.35</b>
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	<b>3.429</b>	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	<b>3.509</b>	14.01	26.07
Ours(SVD)	<b>4.921</b>	<b>0.862</b>	<b>158.36</b>	<b>0.701</b>	<b>3.271</b>	<b>12.32</b>	<b>0.032</b>	<b>0.012</b>	<b>9.931</b>	<b>0.371</b>	<b>1111.4</b>	<b>2.514</b>	<b>9.262</b>	<b>23.87</b>
Ours(CogX)	<b>4.492</b>	<b>0.884</b>	<b>88.493</b>	<b>0.231</b>	<b>1.219</b>	<b>1.198</b>	<b>0.009</b>	<b>0.003</b>	<b>9.903</b>	<b>0.374</b>	<b>1149.4</b>	4.391	<b>10.25</b>	33.72
<i>DepthCharge fluid simulation</i>									<i>Typhoon 202204</i>					
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	<b>22.66</b>
LoRA(CogX)	4.775	0.812	212.26	1.366	5.492	<b>12.04</b>	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)	<b>4.672</b>	<b>0.832</b>	<b>418.18</b>	<b>0.875</b>	<b>8.026</b>	12.83	<b>0.050</b>	<b>0.025</b>	<b>9.610</b>	<b>0.406</b>	<b>1462.8</b>	<b>2.830</b>	<b>10.71</b>	<b>29.48</b>
Ours(CogX)	<b>3.990</b>	<b>0.872</b>	<b>94.474</b>	<b>0.238</b>	<b>1.179</b>	<b>1.454</b>	<b>0.011</b>	<b>0.003</b>	<b>9.614</b>	<b>0.396</b>	<b>1608.8</b>	<b>3.702</b>	<b>11.39</b>	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).

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.

	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	<b>5.852</b>	<b>0.948</b>	<b>62.456</b>	<b>0.303</b>	<b>2.631</b>	<b>4.681</b>	<b>0.013</b>	<b>0.003</b>

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.

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 corrected—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 stud-

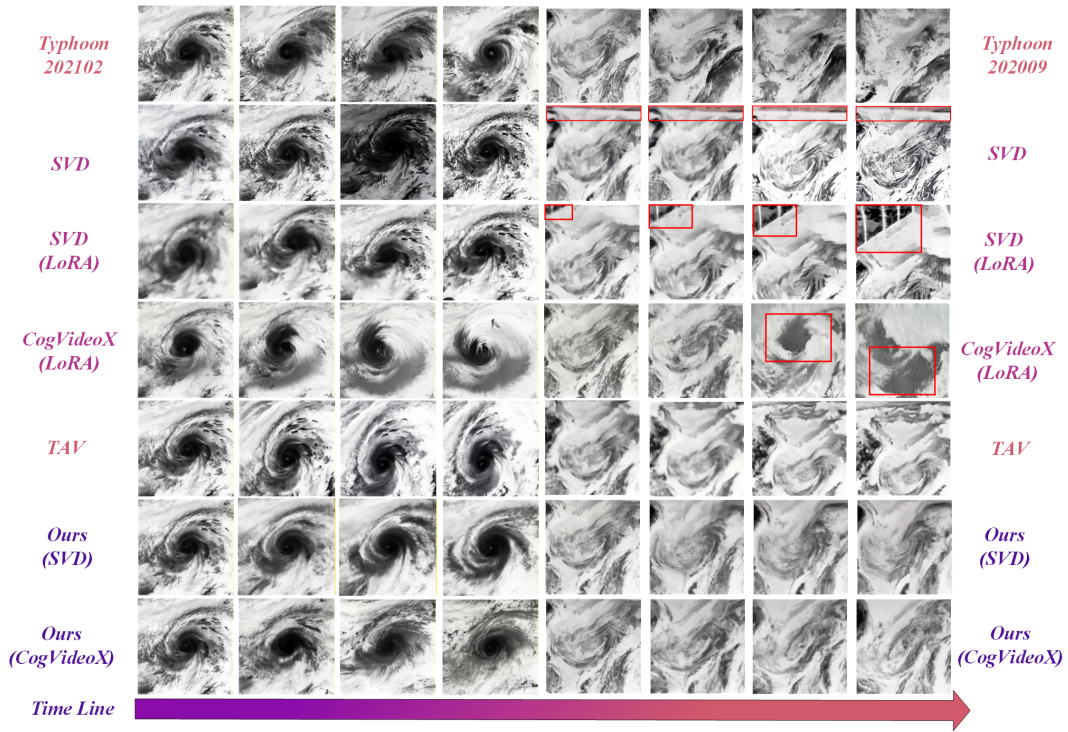


Figure 4: Qualitative comparisons in true typhoon dataset. The red box denotes some hallucinations.

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	<b>5.852</b>	<b>0.948</b>	<b>62.456</b>	<b>0.303</b>	<b>2.631</b>	<b>4.681</b>	<b>0.013</b>	<b>0.003</b>
8	6.018	0.936	68.356	0.351	2.756	5.084	0.027	0.003

Table 3: Ablation study of LoRA rank.

Methods	RMSE↓	SSIM↑	SFE↓	SE↓	GS↓	CS↓	QCE↓	VE↓
Linear	5.913	0.940	73.161	0.398	2.823	4.823	0.018	0.004
Cross Att.	5.921	0.941	70.012	0.358	2.705	4.754	0.016	0.004
Quate.	<b>5.852</b>	<b>0.948</b>	<b>62.456</b>	<b>0.303</b>	<b>2.631</b>	<b>4.681</b>	<b>0.013</b>	<b>0.003</b>

Table 4: Ablation study of projection method.

ies 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, 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.; Mendeleevitch, 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.; Chen, Y.; Wang, D.; Xu, Z.; Ma, C.; Yang, X.; and Chen, S. 2024a. Vision-informed flow field super-resolution with quaternion spatial modeling and dynamic fluid convolution. *Physics of Fluids*, 36(9).
- Cao, Q.; Xu, Z.; Chen, Y.; Ma, C.; and Yang, X. 2024b. 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*.
- Farneback, 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.
- Farneback, 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.; Traubelsi, 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*.