FluidNexus: 3D Fluid Reconstruction and Prediction from a Single Video

Yue Gao*[1,2] Hong-Xing Yu*[1] Bo Zhu[3] Jiajun Wu[1]
[1]Stanford University [2]Microsoft [3]Georgia Institute of Technology

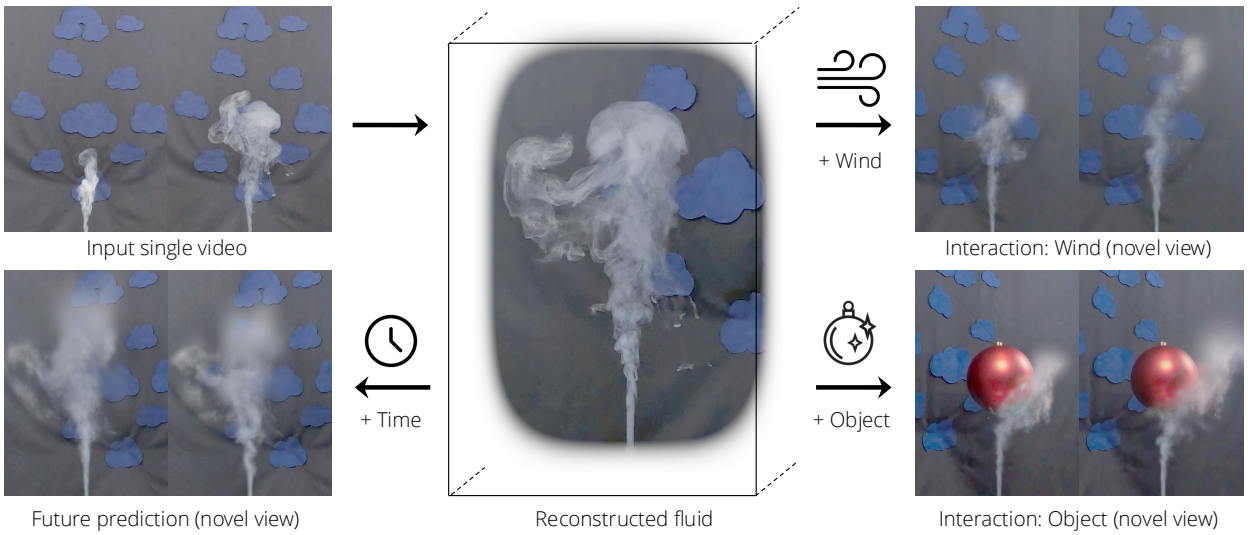

Figure 1. From a single fluid video (top left), we reconstruct the 3D fluid (middle), predict future evolution (bottom left), and simulate wind-fluid interaction (top right) / object-fluid interactions (bottom right).

Abstract

We study reconstructing and predicting 3D fluid appearance and velocity from a single video. Current methods require multi-view videos for fluid reconstruction. We present FluidNexus, a novel framework that bridges video generation and physics simulation to tackle this task. Our key insight is to synthesize multiple novel-view videos as references for reconstruction. FluidNexus consists of two key components: (1) a novel-view video synthesizer that combines frame-wise view synthesis with video diffusion refinement for generating realistic videos, and (2) a physics-integrated particle representation coupling differentiable simulation and rendering to simultaneously facilitate 3D fluid reconstruction and prediction. To evaluate our approach, we collect two new real-world fluid datasets featuring textured backgrounds and object interactions. Our method enables dynamic novel view synthesis, future prediction, and interaction simulation from a single fluid video. Project website: https://yuegao.me/FluidNexus.

*Equal contribution.

1. Introduction

Video-based fluid reconstruction and prediction presents a promising direction for understanding fluid dynamics in scenarios where complex measurement equipment and computational fluid dynamics (CFD) simulations are impractical. This capability has broad applications in industrial monitoring [10], visual special effects [40], and scientific visualization [44]. Recent state-of-the-art approaches have explored integrating neural rendering with physics priors for fluid reconstruction from videos [9, 56]. However, these methods all require multi-view videos for reconstruction, which are often difficult to obtain in real-world scenarios.

In this paper, we address a novel problem setup: **single-video 3D fluid reconstruction and prediction**. Specifically, we aim to start with a single-view video and generate multiple synchronized novel-view videos to serve as references for 3D fluid reconstruction. At first glance, this problem appears ill-posed, as a single video provides limited information about the fluid's intricate 3D structure and dynamics, while infinitely many potential 3D fluid states could correspond to the observed frames. We identify three key challenges in

tackling this problem: (1) *Single-view video to multi-view video:* From a single-view video input, we aim to synthesize realistic videos of the same scene from novel viewpoints. This task represents a conditional video-to-video translation problem with substantial viewpoint changes, which remains an open challenge for existing video generation methods. (2) *Multi-view video to 4D reconstruction:* Using the generated multi-view videos, which may exhibit inconsistencies due to conflicts between synthesized views, we aim to reconstruct spatiotemporally consistent and physically plausible fluid flow motion. (3) *Reconstruction to prediction:* Building upon the reconstructed 4D fluid flow data, we seek to integrate physical models to predict future fluid motion. This reconstruction-to-prediction problem is particularly challenging due to the difficulty in identifying the inherent physical priors from the reconstructed data and enforcing these constraints in future prediction.

We propose a novel video-to-prediction framework ***FluidNexus*** to address these challenges. Our system consists of two components: (1) *A novel-view video synthesizer*, comprising a frame-wise novel-view diffusion model and a video diffusion refiner, which generates multi-view videos from a single-view input. The frame-wise model synthesizes individual frames from novel viewpoints without accounting for fluid dynamics, while the video diffusion refiner enhances these frames into a coherent and realistic video sequence. (2) *A two-layer particle representation*, which bridges differentiable physics and differentiable rendering using two sets of moving particles, reconstructs 4D fluid flow motion from multi-view video input. Specifically, our two-layer particle model integrates a group of **physical particles**, which implements a differentiable physics simulator based on Position-Based Fluid (PBF) [36], with a group of **visual particles**, which establishes a differentiable rendering pipeline leveraging 3D Gaussian Splatting [27] to link the input videos with the 4D reconstructions. The interaction between these two particle groups enables a sparse yet effective representation of the spatiotemporal fluid flow during reconstruction based on multi-view video input, while simultaneously enforcing physical constraints on fluid motions during prediction, where ground truth video reference is no longer available.

As shown in Figure 1, our FluidNexus framework simultaneously addresses the multifaceted challenges of video synthesis, 4D fluid reconstruction, and future motion prediction, all starting from a single video input. We evaluated our framework on existing benchmarks as well as two newly introduced fluid datasets, which feature more complex fluid motion dynamics and intricate environmental interactions. Our key contributions can be summarized as follows:

- A novel framework, FluidNexus, which enables single-video fluid reconstruction and prediction by bridging generative models and physics simulation.
- A reconstruction algorithm and a prediction algorithm that incorporate physics principles of 3D fluid dynamics and generative priors learned from fluid videos via differentiable simulation and rendering.
- Two new real fluid datasets featuring turbulent flow, textured background, and solid obstacles, which we use to train FluidNexus, and demonstrate novel view synthesis and future prediction capabilities from a single video.

2. Related Work

Video-based fluid analysis. Scientists and engineers have extensively studied fluid flow analysis through visible light measurements, primarily in controlled laboratories. These studies employ active sensing approaches (*e.g.*, laser scanners [21] and structural light [20]) or passive techniques like particle imaging velocimetry (PIV) [1, 12]. To allow broader applicability, recent methods have explored video-based fluid analysis with tomography [19, 41, 57] or neural/differentiable rendering [15]. They use multi-view videos as reference, often aided with differentiable physics simulation [11, 15] to constrain the solution within a domain, *e.g.*, GlobTrans [15] uses differentiable simulation with visual hull to constrain the reconstructed fluid. Recent methods [51] such as Physics-Informed Neural Fluid [9] and HyFluid [56] resort to physics-informed losses to help maintain physical plausibility. These methods require multiple synchronized videos as input, which are often unavailable. To address this limitation, we incorporate video generation to tackle single-video 3D fluid reconstruction and prediction.

Dynamic 3D reconstruction. Densely reconstructing dynamic 3D scenes has been challenging [3]. A recent breakthrough comes from representing scenes with dynamic radiance fields such as dynamic NeRFs [16, 17, 31, 38, 42, 43] and dynamic 3D Gaussians [32, 33, 35] and using differentiable (neural) rendering to optimize them from observed multi-view videos. By leveraging monocular depth estimation [26, 45], latest methods even allow reconstructing dynamic scenes from monocular videos [29, 49, 58]. Our approach also represents the fluid appearance with radiance fields. Different from these methods which do not consider physics modeling and thus unable to do prediction, our approach allows future prediction and interaction simulation.

Physics-based dynamic scene modeling. Traditional fluid simulation methods span a broad spectrum, from high-fidelity computational fluid dynamics (CFD) [2] to real-time approximations like stable fluids [46, 48] and position-based dynamics [36, 39]. While these methods excel at physical accuracy, they often struggle with realistic appearance and detail synthesis. Recent learning-based approaches [28] have demonstrated the ability to enhance simulation with learned fluid details [8] and dynamics [53]. In parallel, computer vision community has developed physics-integrated represen-

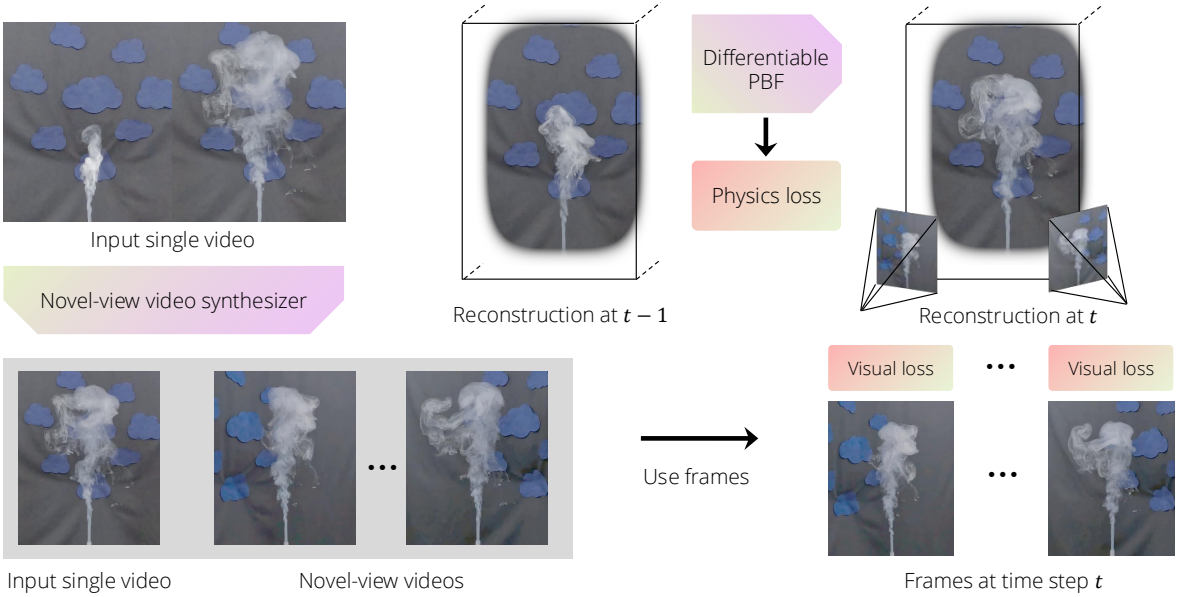

Figure 2. **FluidNexus in reconstruction.** From a single video, we synthesize multiple novel-view videos as references for 3D fluid reconstruction. We then sequentially optimize the two-layer particle fluid representations over time, using the multi-view video frames to compute the visual loss and the physics constraints to compute the physics loss. Our reconstruction output is the 3D fluid appearance and velocity fields over all input frames.

tations that combine physical constraints with differentiable rendering [13, 14, 30, 54], enabling motion synthesis and interaction of objects of different materials [60, 61]. However, they typically require multi-view inputs or pre-reconstructed 3D scenes. Our work bridges this gap by integrating video generation, allowing single-video fluid reconstruction and prediction.

Video generation. In the past few years, video generation models [22, 47], especially video diffusion models [4–7, 18, 23], have been rapidly developed. Recent video generation methods, such as Sora [7], have demonstrated great promise in simulating real-world complex physical events, including fluid dynamics and interaction. Yet, the generation is in 2D, and the controllability is limited. Nevertheless, the ability to simulate realistic fluid dynamics and fluid-object interaction motivates us to integrate video diffusion models into fluid reconstruction and prediction.

3. Approach

Problem statement. Given a fixed-viewpoint video $\mathcal{V}^0 = (I_1^0, I_2^0, \cdots, I_T^0)$ containing T frames $I_t^0 \in \mathbb{R}^{H\times W\times 3}$ of fluid dynamics (*e.g.*, a rising plume), we aim to reconstruct 3D fluid velocity and appearance over time T and predict future states beyond T.

Our solution. We propose FluidNexus, which consists of a novel-view video synthesizer (left of Figure 2) to create temporally consistent multi-view videos based on a single video input and a physics-integrated two-layer particle representation (right of Figure 2) to reconstruct and predict fluid motion based on multi-view video inputs. We elaborate these two components in the following.

3.1. Novel-view Video Synthesizer

Given an input video $\mathcal{V}^0 = (I_1^0, I_2^0, \cdots, I_T^0)$, we aim to generate C novel-view videos $\{\mathcal{V}^c\}_{c=1}^C$, where $\mathcal{V}^c = (I_1^c, I_2^c, \cdots, I_T^c)$. The key challenge lies in ensuring both spatial consistency across views and temporal coherence within each view. To address this, we design our video synthesizer with a frame-wise view synthesis model and a video refinement diffusion model.

Frame-wise Novel View Synthesis. To learn spatial coherence, we employ a camera view-conditioned image diffusion model to synthesize frames at novel viewpoints [34]. Given a single frame I_t^0 at timestep t and a camera transform matrix $\pi_c \in \mathbb{R}^{3\times 4}$ from the input view to the target view c, the diffusion model learns to synthesize the novel view:

$$\hat{I}_t^c = g(I_t^0, \pi_c), \tag{1}$$

where $\hat{I}_t^c$ denotes the synthesized frame and g denotes the diffusion model. The diffusion model performs denoising steps conditioned on the input frame I_t^0 and the camera transform matrix π_c to generate geometrically consistent novel views. Over $t = 1$ to T, this gives a rough video $\hat{\mathcal{V}}^c = (\hat{I}_1^c, \hat{I}_2^c, \cdots, \hat{I}_T^c)$ for a given viewpoint c.

Generative Video Refinement. Since each frame is generated independently of other frames, the generated frames lack temporal consistency. To address this, we introduce a generative video refinement approach conditioned on the

synthesized frames to generate temporally coherent fluid videos. Our refinement extends an image editing technique, SDEdit [37], to video diffusion:

$$(I_1^c, I_2^c, \cdots, I_T^c) = v(\hat{\mathcal{V}}^c | \lambda_{\text{SDEdit}}), \tag{2}$$

where v represents the video diffusion model and λ_{SDEdit} denotes the strength of generative refinement. The refinement works as follows: intuitively, instead of denoising a sampled pure noise for L steps to generate a video from scratch, it first creates an intermediate perturbed video $\widetilde{\mathcal{V}}^c$ by injecting mild noise to all frames in $\hat{\mathcal{V}}^c$, and then performs the latter $\lambda_{\text{SDEdit}} L$ denoising steps ($0 < \lambda_{\text{SDEdit}} < 1$). This generates a fluid video $\mathcal{V}^c$ that maintains spatial content in $\hat{\mathcal{V}}^c$ while ensuring temporal coherence by pulling it to the video manifold learned by the video diffusion model [37]. The strength of this generative refinement λ_{SDEdit} controls the balance between content preservation and temporal consistency. We leave further technical details and training details in the supplementary material.

Long Video Generation. Video diffusion models typically operate on a limited time window T' smaller than our target video length T due to computing limitations. Naively applying the video refinement to consecutive T'-frame segments leads to periodic jittering at segment boundaries. To address this, we train two video refiners: an unconditional $v_{\text{uncond}}((\hat{I}_1^c, \cdots, \hat{I}_{T'}^c))$ for generating the first segment $(I_1^c, \cdots, I_{T'}^c)$; And a $v_{\text{cond}}((\hat{I}_{T'+1}^c, \cdots, \hat{I}_{2T'-S}^c) | (I_{T'-S+1}^c, \cdots, I_{T'}^c))$ conditioned on a few S fixed previous frames for extending an existing segment. We recursively use the conditioned refiner v_{cond} to progressively extend the sequence by $T' - S$ frames at a time, maintaining consistency with previously generated frames.

3.2. Two-layer Particle Fluid Representation

While the generated multi-view videos $\{\mathcal{V}^c\}_{c=1}^C$ provide spatial-temporal references, they lack physical plausibility in fluid dynamics. To address this, our representation integrates Position-Based Fluid (PBF) [36] simulation, known for its efficiency and flexibility, with 3D Gaussian Splatting [27] to bridge simulated particles with the generated videos.

Particle-based Fluid Representation. Our fluid representation consists of two types of particles: physical particles that represent the fluid velocity field and density field in ambient 3D space, and visual particles that represent the fluid appearance and are passively advected by the velocity field. Physical particles at any timestep t are defined by their positions $\mathbf{p}_t \in \mathbb{R}^{3N_t^{\text{physical}}}$ where N_t^{physical} denotes the physical particle count at t, and the associated particle velocities $\mathbf{u}_t \in \mathbb{R}^{3N_t^{\text{physical}}}$ (not to be confused with the velocity field $\mathbf{V}_t : \mathbb{R}^3 \mapsto \mathbb{R}^3$). The velocity at any 3D point $\mathbf{x}$ at timestep t is computed through kernel-weighted interpolation:

$$\mathbf{V}_t(\mathbf{x} | \mathbf{u}_t, \mathbf{p}_t) = \frac{\sum_j \mathbf{u}_{t,j} K(\mathbf{x} - \mathbf{p}_{t,j})}{\sum_j K(\mathbf{x} - \mathbf{p}_{t,j})}, \tag{3}$$

where $K(\cdot)$ is a symmetric kernel function, and the density is also defined through kernel-weighted summation: $\rho_t(\mathbf{x}|\mathbf{p}_t) = \sum_j K(\mathbf{x} - \mathbf{p}_{t,j})$. Visual particles at timestep t are characterized by their attributes $\{\mathbf{x}_t, \mathbf{c}_t, \mathbf{s}_t, \mathbf{o}_t, \mathbf{r}_t\}$, representing position, color, scale, opacity, and orientation, respectively. Similarly, there are N_t^{visual} visual particles at t. They are rendered through rasterization following the 3D Gaussian Splatting [27].

Physical Constraints from Simulation. We generate physical constraints for both reconstruction and prediction through fluid simulation. The core idea is to use simulation to provide a physically plausible guess of the physical particles $\mathbf{p}_t^{\text{sim}}$, which, together with incompressibility, creates a physics loss to help solve for the velocity. We will only optimize $\mathbf{p}_t$ as it exclusively represents the velocity field[1] and density field. The simulation step can be written as: $\mathbf{p}_t^{\text{sim}} = \texttt{Sim}(\mathbf{u}_{t-1}, \mathbf{p}_{t-1})$, where we use $\mathbf{p}_t^{\text{sim}}$ for computing physics loss and for initializing $\mathbf{p}_t$. The physics loss consists of two terms:

$$\mathcal{L}_{\text{physics}} = \lambda_{\text{sim}} \mathcal{L}_{\text{sim}} + \mathcal{L}_{\text{incomp}}, \tag{4}$$

where $\mathcal{L}_{\text{sim}} = \|\mathbf{p}_t - \mathbf{p}_t^{\text{sim}}\|_2^2$ encourages the physical fields to be consistent with the simulation solution, and λ_{sim} denotes a weight. And we have the incompressibility loss:

$$\mathcal{L}_{\text{incomp}} = \sum_j \left(\frac{\rho_t(\mathbf{p}_{t,j})}{\rho_0} - 1\right)^2 + \lambda_{\text{next}} \sum_j \left(\frac{\rho_{t+1}(\mathbf{p}_{t+1,j})}{\rho_0} - 1\right)^2 + \lambda_{\text{v-incomp}} \sum_{i \neq j} (\max(0, \sigma - d_{t,ij}))^2, \tag{5}$$

where ρ_0 denotes the constant environmental air density, and $\mathbf{p}_{t+1} = \texttt{Sim}(\mathbf{u}_t, \mathbf{p}_t)$ denotes the physical particle positions at the next timestep by differentiable simulation. The first two terms encourage incompressibility around the physical particle positions at both current and next timesteps. The third term encourages incompressibility around the visual particles by maintaining a minimal distance σ between each pair of them at the current timestep $d_{t,ij} = \|\mathbf{x}_{t,i} - \mathbf{x}_{t,j}\|_2^2$, where the current positions

$$\mathbf{x}_t = \texttt{Adv}(\mathbf{V}_t, \mathbf{x}_{t-1}) \tag{6}$$

are estimated via advecting $\mathbf{x}_{t-1}$ using the current velocity field $\mathbf{V}_t$. We leave the details of the simulation operator `Sim` and advection operator `Adv` in the supplementary material.

[1]The velocity field $\mathbf{V}_t$ is a function of $\mathbf{u}_t$ and $\mathbf{p}_t$, and $\mathbf{u}_t = (\mathbf{p}_t - \mathbf{p}_{t-1})/\Delta t$ is also a function of $\mathbf{p}_t$ given fixed $\mathbf{p}_{t-1}$.

3.3. Generative Reconstruction and Prediction

Reconstruction. Given multi-view videos $\mathcal{V}_{c=0}^{C}$ with their camera poses $\{\pi_c\}_{c=0}^{C}$, we reconstruct fluid velocity (represented by $\mathbf{p}_t$) and appearance (represented by $\{\mathbf{x}_t, \mathbf{c}_t, \mathbf{s}_t, \mathbf{o}_t, \mathbf{r}_t\}$) over time T. We formulate this as a sequential optimization from $t = 1$ to $t = T$, fixing all quantities at $t-1$ when solving for time t (we leave $t = 0$ in supplementary material.) Our overall loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{physics}} + \mathcal{L}_{\text{visual}} + \mathcal{L}_{\text{reg}}, \tag{7}$$

where the visual loss $\mathcal{L}_{\text{visual}} = \sum_{c=0}^{C} \mathcal{L}_1(I_t^c, I_t'^c) + \mathcal{L}_{\text{SSIM}}(I_t^c, I_t'^c)$ measures differences between reference frames I_t^c and rendered images $I_t'^c = \texttt{Render}(\pi_c, \mathbf{x}_t, \mathbf{c}_t, \mathbf{s}_t, \mathbf{o}_t, \mathbf{r}_t)$ with the image difference losses $\mathcal{L}_1$ and $\mathcal{L}_{\text{SSIM}}$ as in 3D Gaussian Splatting [27]. Here $\mathbf{x}_t$ is also obtained via Eq. 6. The regularization term $\mathcal{L}_{\text{reg}}$ encourages temporal consistency of $\{\mathbf{c}_t, \mathbf{s}_t, \mathbf{o}_t, \mathbf{r}_t\}$ with $\{\mathbf{c}_{t-1}, \mathbf{s}_{t-1}, \mathbf{o}_{t-1}, \mathbf{r}_{t-1}\}$ (details in supplementary material).

This sequential optimization is challenging due to potential error accumulation and the complex entanglement between physics and appearance. Therefore, we decompose it into two stages. First, we solve for dynamics while fixing appearance attributes:

$$\min_{\mathbf{p}_t} \mathcal{L}_{\text{physics}} + \mathcal{L}_{\text{visual}}, \quad t = 1, \ldots, T. \tag{8}$$

Then, with $\mathbf{p}_t$ fixed, we use Eq. 6 to get $\mathbf{x}_t$ and optimize other appearance attributes:

$$\min_{\mathbf{c}_t, \mathbf{s}_t, \mathbf{o}_t, \mathbf{r}_t} \mathcal{L}_{\text{visual}} + \mathcal{L}_{\text{reg}}, \quad t = 1, \ldots, T. \tag{9}$$

Limitations of Pure Physics Simulation in Prediction. In prediction, we solve for $\{\mathbf{p}_t, \mathbf{x}_t, \mathbf{c}_t, \mathbf{s}_t, \mathbf{o}_t, \mathbf{r}_t\}, t = T+1, \cdots, T_{\text{target}}$ with optionally new interactions (*e.g.*, inserting an object or adding a wind). While physics simulation can be highly accurate with known initial conditions and complete fluid properties, predicting future states for reconstructed fluids poses significant challenges. Due to inevitable inaccuracies in the reconstructed velocity and appearance fields, and the absence of modeling specific fluid attributes like temperature and viscosity, pure simulation struggles to capture the complex fluid dynamics of the observed fluid. These inaccuracies compound over time, leading to simplified fluid motion that deviates from the actual dynamics. Furthermore, visual characteristics of the observed fluid, such as scattering effects, cannot be faithfully reproduced through simulation of the reconstructed states alone.

Generative Fluid Simulation. We address these limitations through generative simulation that combines physics simulation with video generation. For predicting $t > T$, we first simulate rough fluid dynamics by $\mathbf{p}_t^{\text{pred}} = \texttt{Sim}(\mathbf{u}_{t-1}^{\text{pred}}, \mathbf{p}_{t-1}^{\text{pred}})$ and $\mathbf{x}_t^{\text{pred}} = \texttt{Adv}(\mathbf{V}_t^{\text{pred}}, \mathbf{x}_{t-1}^{\text{pred}})$ from $t = T+1$ to $t = T_{\text{target}}$, with $\mathbf{x}_T^{\text{pred}} = \mathbf{x}_T$ and $\mathbf{p}_T^{\text{pred}} = \mathbf{p}_T$. Then, we render rough multi-view videos $\{(\hat{I}_{T+1}^c, \cdots, \hat{I}_{T_{\text{target}}}^c)\}_{c=0}^{C}$ where $\hat{I}_t^c = \texttt{Render}(\pi_c, \mathbf{x}_t, \mathbf{c}_0, \mathbf{s}_0, \mathbf{o}_0, \mathbf{r}_0), t > T$, where $\mathbf{c}_0, \mathbf{s}_0, \mathbf{o}_0, \mathbf{r}_0$ are constant initialization values. These rough videos capture basic fluid dynamics but lack detailed dynamics or visual realism. We then refine these videos using our video refinement model as in Eq. 2 to obtain reference prediction videos $\{(I_{T+1}^c, \cdots, I_{T_{\text{target}}}^c)\}_{c=0}^{C}$ and apply the reconstruction algorithm described above to solve for $t = T+1, \cdots, T_{\text{target}}$. We leave interaction simulation details in the supplementary material. We also summarize the reconstruction algorithm and the prediction algorithm in the supplementary material.

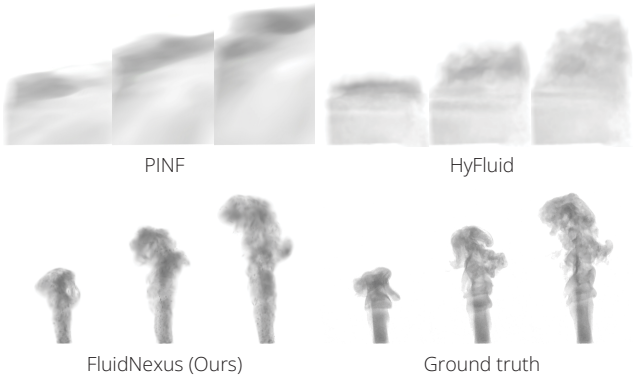

Figure 3. Novel view synthesis on ScalarFlow [11].

4. Experiments

Datasets. We use the widely-adopted ScalarFlow dataset [11] which contains 104 real plume scenes with five synchronized videos for each scene. Yet, all scenes in ScalarFlow do not have textured backgrounds or fluid interaction with other objects. Thus, we collect two real datasets for evaluating fluid reconstruction and prediction in more challenging setups. We leave more details in the supplementary material.

FluidNexus-Smoke includes 120 scenes. Each scene has a jet of smoke spurting from an ejector in front of a textured background. The ejector creates a burst of fast rising smoke that yields more vortical details compared to ScalarFlow scenes, making this dataset much more challenging. Each scene has 5 synchronized multi-view videos where the cameras are placed along a horizontal arc of approximately 120°. For each scene, we keep 6s of the fluid event with FPS 30 (*i.e.*, 180 frames).

FluidNexus-Ball includes 120 scenes. Compared to FluidNexus-Smoke, FluidNexus-Ball inserts a shiny ball above the smoke ejector, creating fluid-solid interactions.

Baselines. We compare FluidNexus with two representative state-of-the-art approaches on 3D fluid reconstruction:

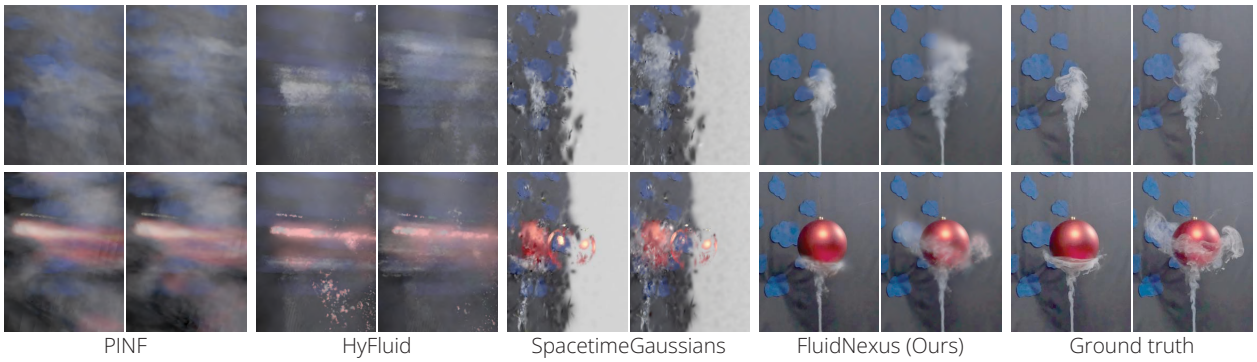

Figure 4. Qualitative results of novel view synthesis on our collected datasets.

Model	Novel View Synthesis			Input View Future Prediction			Novel Views Future Prediction			Re-simulation			
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	$\nabla \cdot \mathbf{V}$ ↓
PINF [9]	22.68	0.7597	0.1926	20.48	0.6689	0.2737	20.66	0.6709	0.2704	22.16	0.7409	0.1970	0.0297
HyFluid [56]	22.23	0.7645	0.2275	26.84	0.9072	0.1776	20.29	0.6280	0.3418	22.26	0.7643	0.2272	0.0619
STG [32]	19.85	0.7063	0.2790	21.79	0.8759	0.2142	18.51	0.7011	0.3697	19.73	0.7033	0.2880	0.0973
FluidNexus (Ours)	**32.45**	**0.9544**	**0.1299**	**28.51**	**0.9159**	**0.1754**	**26.83**	**0.8952**	**0.2052**	**32.44**	**0.9543**	**0.1299**	**0.0126**

Table 1. Quantitative results on ScalarFlow [11].

Physics-Informed Neural Fluid (PINF) [9] and HyFluid [56]. In addition, we include a state-of-the-art 4D dynamic reconstruction model SpacetimeGaussians (STG) [32].

Task settings. For each scene in each dataset, we use the video from the middle camera as input, and the other 4 videos as ground truth. For 3D fluid reconstruction, we evaluate the visual appearance via novel view synthesis, and we evaluate the velocity via re-simulation (*i.e.*, recreating the fluid dynamics by progressively advecting reconstructed fluid at the initial timestep). For these two tasks, we use the first 120 frames. For 3D fluid prediction, we focus on future prediction and interaction simulation. To create ground truth for future prediction, we use the first 120 frames for reconstruction and predict the next 60 frames for our dataset.

Metrics. For novel view synthesis and future prediction, we compute PSNR, SSIM [52] and LPIPS [59] against the groundtruth frames. For re-simulation, we also compute the divergence of the velocity fields $\nabla \cdot \mathbf{V}$ averaged over time and 3D dense grid points to measure the incompressibility and the quality of the reconstructed velocity field.

Implementation details. We train a Zero123 [34] model as our frame-wise novel view synthesis model g, which generates frames at a resolution of 256×256. We use CogVideoX-5b [55] as our generative video refinement model v, which has a time window of T'=49 frames. It generates videos at a resolution of 720×480. We use the public pretrained model and fine-tune it with LoRA [25] on each of the fluid datasets. We use CogVLM2 [24, 50] to generate captions for our datasets and the ScalarFlow [11]. We set λ_{SDEdit}=0.5 for refining frames during reconstruction and λ_{SDEdit}=0.75 during prediction. For our dataset, we used 100 scenes as fine-tuning training samples, while for ScalarFlow [11], we used 94 scenes as training samples. We held out 20 scenes and 10 scenes, respectively, for evaluation. We set the loss weights to $\lambda_{\text{sim}} = 0.1$, $\lambda_{\text{next}} = 0.1$, and $\lambda_{\text{v-incomp}} = 0.1$ for all experiments. For more details, please refer to supplementary material.

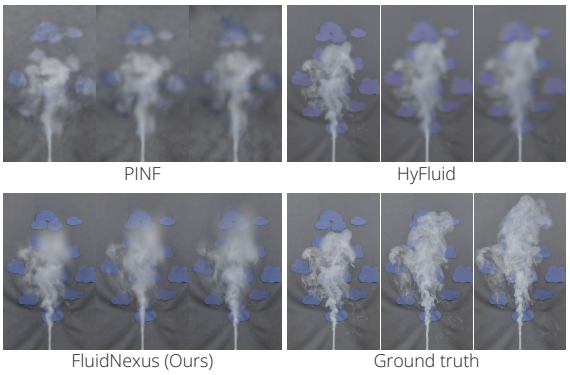

Figure 5. Qualitative results of future prediction on input view.

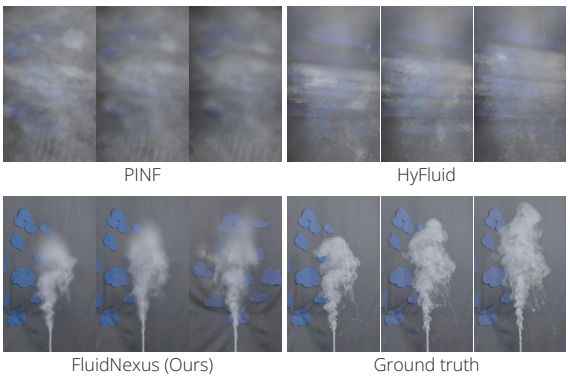

Figure 6. Qualitative results of future prediction on novel view.

Model	Novel View Synthesis			Input View Future Prediction			Novel Views Future Prediction			Re-simulation			
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	$\nabla \cdot \mathbf{V}$ ↓
PINF [9]	22.40	0.8002	0.5089	26.48	0.7299	0.2418	22.66	0.8234	0.5931	21.97	0.7992	0.5029	0.0451
HyFluid [56]	22.64	0.7948	0.4764	21.14	0.7044	0.5417	21.95	0.8334	0.6013	22.34	0.7615	0.4937	0.0573
STG [32]	19.94	0.6875	0.3673	23.94	0.8408	0.2639	18.34	0.6325	0.4116	19.48	0.6867	0.4818	0.0323
FluidNexus (Ours)	**30.62**	**0.9209**	**0.1707**	**27.79**	**0.8747**	**0.2337**	**25.74**	**0.8609**	**0.2675**	**30.61**	**0.9208**	**0.1707**	**0.0246**

Table 2. Quantitative results on our FluidNexus-Smoke.

Model	Novel View Synthesis			Input View Future Prediction			Novel Views Future Prediction			Re-simulation			
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	$\nabla \cdot \mathbf{V}$ ↓
PINF [9]	20.50	0.7556	0.4611	24.70	0.8162	0.5079	18.26	0.6199	0.4458	20.05	0.7111	0.5127	0.0441
HyFluid [56]	20.71	0.7251	0.4978	20.93	0.7005	0.5019	18.28	0.7239	0.4380	20.60	0.7152	0.5192	0.0518
STG [32]	18.70	0.6793	0.3866	25.80	0.8629	0.2251	17.11	0.6156	0.4131	18.46	0.6819	0.3843	0.0531
FluidNexus (Ours)	**29.89**	**0.9107**	**0.1773**	**27.70**	**0.8733**	**0.2019**	**24.82**	**0.8431**	**0.2607**	**29.88**	**0.9101**	**0.1729**	**0.0280**

Table 3. Quantitative results on our FluidNexus-Ball.

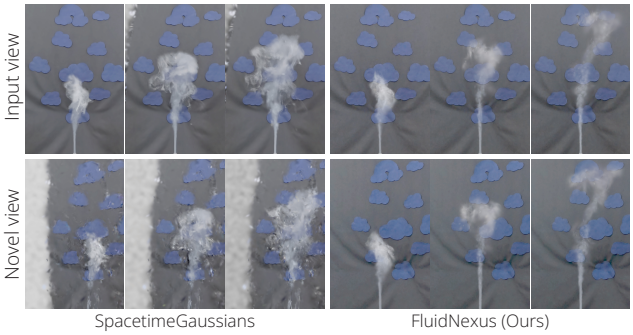

Figure 7. Qualitative results of wind-fluid interactive simulation results rendered at the input view or the novel view.

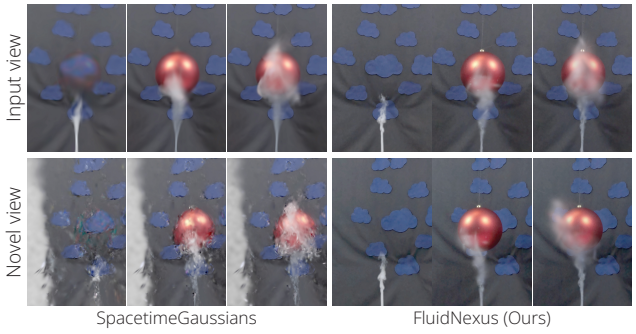

Figure 8. Qualitative results of object-fluid interactive simulation results rendered at the input view or the novel view.

Settings	Novel View Synthesis			Settings	Re-simulation		
	PSNR ↑	SSIM ↑	LPIPS ↓		PSNR ↑	SSIM ↑	LPIPS ↓
w/o NVS	21.88	0.5909	0.6117	w/o $\mathcal{L}_{\text{physics}}$	28.00	0.9016	0.1913
w/o GVR	29.12	0.9060	0.1904	w/o $\mathcal{L}_{\text{incomp}}$	28.76	0.9148	0.1854
w/o LVG	30.14	0.9196	0.1819	w/o $\mathcal{L}_{\text{sim}}$	29.70	0.9175	0.1905
FluidNexus	**30.62**	**0.9209**	**0.1707**	FluidNexus	**30.61**	**0.9208**	**0.1707**

Table 4. Ablation studies on FluidNexus-Smoke.

4.1. Comparison to baselines

Novel view synthesis. We showcase the novel view synthesis results on the ScalarFlow dataset [11] in Fig. 3 and the results on our new datasets in Fig. 4. We observe that PINF [9] and HyFluid [56], both designed for multi-view fluid reconstruction using physics losses and neural rendering, fail to synthesize reasonable novel views, as they cannot disambiguate the thickness of the fluid. In contrast, FluidNexus allows plausible novel view synthesis on all three datasets. We show quantitative metrics in Tab. 1, Tab. 2, and Tab. 3, where FluidNexus also outperforms all baselines in quantitative measurements. Note that even for scenes with textured backgrounds and scenes containing fluid-object interaction, FluidNexus achieves reasonable 3D fluid reconstruction (Fig. 4).

Future prediction. We showcase input-view future prediction results on the FluidNexus-Smoke in Fig. 5. We compared our method with PINF [9] and HyFluid [56], noting that their future predictions often mimic the last observed input frame due to an inadequate 3D velocity field, leading to physically inaccurate fluid advection. The single-view input also limits appearance information, resulting in lower-quality predictions. As shown in Fig. 6, future predictions from novel viewpoints reveal that other methods fail to deliver satisfactory results. Tab. 1, Tab. 2, and Tab. 3 further demonstrate that while these methods perform reasonably well in the input view, our approach consistently outperforms them across all views.

Re-simulation. Tab. 1, Tab. 2, and Tab. 3 present quantitative evaluations of the appearance consistency and physical correctness of the reconstructed velocity fields. Our method FluidNexus, which reconstructs fluid particle positions and velocity fields at each time step under physical constraints, achieves lower divergence of velocity field and fully reproduces the reconstruction results. In contrast, other methods show significantly poorer performance.

4.2. Interaction Simulation

To showcase the superiority and robustness of our PBF-based differentiable simulator combined with generative models, we present scenarios involving interactions with external force or additional object. We use *the same input video* for all methods. After reconstruction, we simulate adding wind or a ball to the reconstructed scenes. Our explicit PBD-based [39] representation allows direct application of forces or object interaction constraints to particle positions. In contrast, implicit NeRF-based [38] methods like PINF [9]

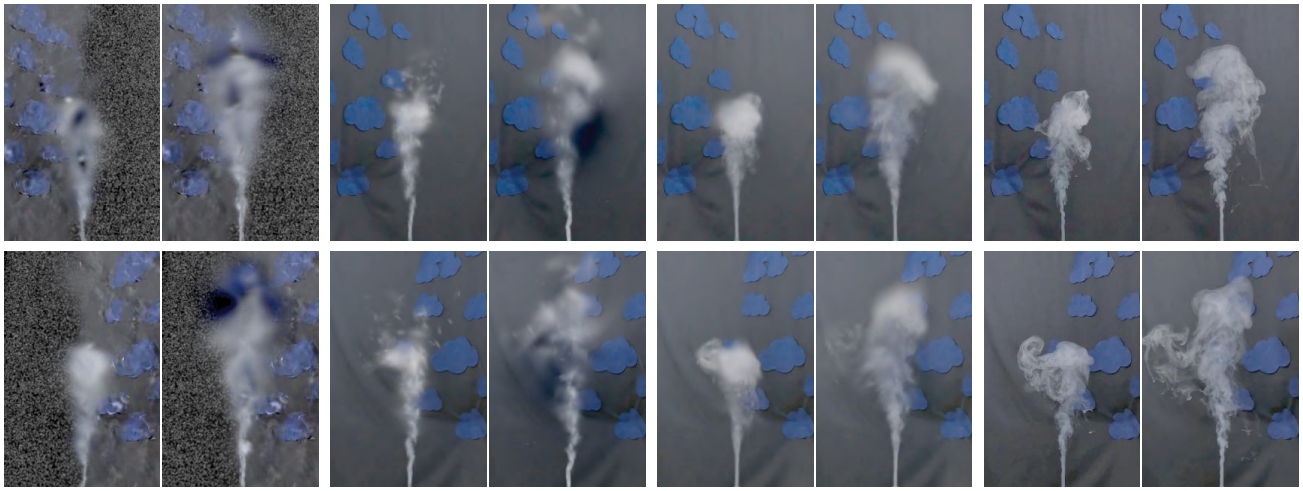

Figure 9. Ablation studies of FluidNexus. Results in the top row are from a novel view, and results in the bottom row are from another novel view. We leave a visual comparison of all variants in the supplementary material.

and HyFluid [56] must reconstruct entire volumes and solve higher-order PDEs, making their process less stable, which is much difficult comparing to our model. Thus, we compare only with SpacetimeGaussians [32] in this context. While no ground truth videos are available with and without the extra force or rigid body, qualitative results Fig. 7 and Fig. 8 show that our method produces more realistic outcomes.

4.3. Ablation studies

We conduct ablation study experiments on the FluidNexus-Smoke dataset to evaluate the core components of our proposed FluidNexus.

Video generation. We evaluate three variants: Firstly, we remove the novel-view video synthesizer from FluidNexus, denoted as "w/o NVS". Secondly, we remove the generative video refinement, denoted as "w/o GVR". Lastly, we remove the long video generation support (*i.e.*, we use only the unconditional video refinement diffusion model and refine each 49-frame segment at a time), denoted as "w/o LVG". We show the novel view synthesis results in the left panel of Tab. 4. From Tab. 4 we observe that the performance degrades drastically when we remove the novel-view video synthesizer, as it provides the generative priors for a reasonable reconstruction especially for the empty air space and the background region. This can be further observed in Fig. 9. The noisy floaters in mid-air and the distorted background textures are due to the lack of reconstruction reference from different viewpoints. The generative video refinement and long video generation are also important as they prevent jittering. This can be clearly observed when viewing the video results (attached in the supplementary material). They only provide mild improvements in terms of novel view synthesis metrics, as these metrics do not measure temporal consistency.

Physics constraints. We evaluate three variants. We remove the physics loss and denote the first variant as "w/o $\mathcal{L}_{\text{physics}}$". We remove the incompressibility loss and denote the second variant as "w/o $\mathcal{L}_{\text{incomp}}$". We remove the simulation supervision and denote the third variant as "w/o $\mathcal{L}_{\text{sim}}$". We show the re-simulation results in the right panel of Tab. 4. We conducted re-simulation experiments to evaluate the impact of physical constraints on the reconstructed velocity fields. Without these constraints, image loss dominates, producing velocity fields that lack physical accuracy. As shown in Tab. 4, these incorrect velocity fields greatly compromise the re-simulation results. Furthermore, Fig. 9 illustrates that the absence of physical constraints results in inaccurate particle trajectories, leading to noticeable artifacts.

5. Conclusion

We presented FluidNexus, a novel framework that enables 3D fluid reconstruction and prediction from a single video by bridging video generation with physics simulation. Through extensive experiments on two new challenging fluid datasets, we show that video generation provides significant support for future prediction tasks. Our work facilitates further research on single-video fluid analysis.

Limitations. One limitation of FluidNexus is that the reconstruction results on challenging datasets, including FluidNexus-Smoke and FluidNexus-Ball, are blurry, missing fine fluid details. The other major limitation is that FluidNexus does not generalize to different backgrounds or different objects, as the novel view video synthesizer is trained on limited scenes. Collecting large-scale multi-view fluid video datasets, either synthetic or real, can be a promising direction to improve generalization.

Acknowledgments. The work was in part supported by ONR MURI N00014-24-1-2575, ONR MURI N00014-22-1-2740, and NSF RI #2211258 and #2338203.

References

[1] Ronald J Adrian and Jerry Westerweel. *Particle image velocimetry*. Cambridge university press, 2011. 2

[2] John David Anderson and John Wendt. *Computational fluid dynamics*. Springer, 1995. 2

[3] Aayush Bansal, Minh Vo, Yaser Sheikh, Deva Ramanan, and Srinivasa Narasimhan. 4d visualization of dynamic events from unconstrained multi-view videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5366–5375, 2020. 2

[4] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024. 3

[5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[6] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[7] Tim Brooks, Bill Peebles, Connor Homes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. 3

[8] Mengyu Chu and Nils Thuerey. Data-driven synthesis of smoke flows with cnn-based feature descriptors. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017. 2

[9] Mengyu Chu, Lingjie Liu, Quan Zheng, Erik Franz, Hans-Peter Seidel, Christian Theobalt, and Rhaleb Zayer. Physics informed neural fields for smoke reconstruction with sparse data. *ACM TOG*, 2022. 1, 2, 6, 7

[10] Xili Duan, Bren Phillips, Thomas McKrell, and Jacopo Buongiorno. Synchronized high-speed video, infrared thermometry, and particle image velocimetry data for validation of interface-tracking simulations of nucleate boiling phenomena. *Experimental Heat Transfer*, 2013. 1

[11] Marie-Lena Eckert, Kiwon Um, and Nils Thuerey. Scalarflow: a large-scale volumetric data set of real-world scalar transport flows for computer animation and machine learning. *ACM TOG*, 2019. 2, 5, 6, 7

[12] Gerrit E Elsinga, Fulvio Scarano, Bernhard Wieneke, and Bas W van Oudheusden. Tomographic particle image velocimetry. *Experiments in fluids*, 2006. 2

[13] Yutao Feng, Xiang Feng, Yintong Shang, Ying Jiang, Chang Yu, Zeshun Zong, Tianjia Shao, Hongzhi Wu, Kun Zhou, Chenfanfu Jiang, et al. Gaussian splashing: Dynamic fluid synthesis with gaussian splatting. *arXiv preprint arXiv:2401.15318*, 2024. 3

[14] Yutao Feng, Yintong Shang, Xuan Li, Tianjia Shao, Chenfanfu Jiang, and Yin Yang. Pie-nerf: Physics-based interactive elastodynamics with nerf. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4450–4461, 2024. 3

[15] Erik Franz, Barbara Solenthaler, and Nils Thuerey. Global transport for fluid reconstruction with learned self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1632–1642, 2021. 2

[16] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 2

[17] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5712–5721, 2021. 2

[18] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023. 3

[19] James Gregson, Ivo Ihrke, Nils Thuerey, and Wolfgang Heidrich. From capture to simulation: connecting forward and inverse problems in fluids. *ACM TOG*, 2014. 2

[20] Jinwei Gu, Shree K Nayar, Eitan Grinspun, Peter N Belhumeur, and Ravi Ramamoorthi. Compressive structured light for recovering inhomogeneous participating media. *IEEE TPAMI*, 2012. 2

[21] Tim Hawkins, Per Einarsson, and Paul Debevec. Acquisition of time-varying participating media. *ACM TOG*, 2005. 2

[22] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[23] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 3

[24] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 6

[25] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 6

[26] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. *arXiv preprint arXiv:2312.02145*, 2023. 2

[27] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. 2, 4, 5

[28] Byungsoo Kim, Vinicius C Azevedo, Nils Thuerey, Theodore Kim, Markus Gross, and Barbara Solenthaler. Deep fluids: A generative network for parameterized fluid simulations. In *CGF*, 2019. 2

[29] Jiahui Lei, Yijia Weng, Adam Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. *arXiv preprint arXiv:2405.17421*, 2024. 2

[30] Xuan Li, Yi-Ling Qiao, Peter Yichen Chen, Krishna Murthy Jatavallabhula, Ming Lin, Chenfanfu Jiang, and Chuang Gan. Pac-nerf: Physics augmented continuum neural radiance fields for geometry-agnostic system identification. In *ICLR*, 2023. 3

[31] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, 2021. 2

[32] Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. Spacetime gaussian feature splatting for real-time dynamic view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8508–8520, 2024. 2, 6, 7, 8

[33] Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21136–21145, 2024. 2

[34] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. In *CVPR*, 2023. 3, 6

[35] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision (3DV)*, pages 800–809. IEEE, 2024. 2

[36] Miles Macklin and Matthias Müller. Position based fluids. *ACM Transactions on Graphics (TOG)*, 2013. 2, 4

[37] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 4

[38] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 7

[39] Matthias Müller, Bruno Heidelberger, Marcus Hennix, and John Ratcliff. Position based dynamics. *Journal of Visual Communication and Image Representation*, 18(2):109–118, 2007. 2, 7

[40] Makoto Okabe, Ken Anjyor, and Rikio Onai. Creating fluid animation from a single image using video database. In *Computer Graphics Forum*, 2011. 1

[41] Makoto Okabe, Yoshinori Dobashi, Ken Anjyo, and Rikio Onai. Fluid volume modeling from sparse multi-view images by appearance transfer. *ACM TOG*, 2015. 2

[42] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv:2106.13228*, 2021. 2

[43] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, 2021. 2

[44] Maziar Raissi, Alireza Yazdani, and George Em Karniadakis. Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science*, 2020. 1

[45] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(3), 2022. 2

[46] Andrew Selle, Ronald Fedkiw, Byungmoon Kim, Yingjie Liu, and Jarek Rossignac. An unconditionally stable maccormack method. *Journal of Scientific Computing*, 2008. 2

[47] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3

[48] J STAM. Stable fluid. In *Proceedings of ACM SIGGRAPH*, pages 121–128, 1999. 2

[49] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. *arXiv preprint arXiv:2407.13764*, 2024. 2

[50] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2023. 6

[51] Yiming Wang, Siyu Tang, and Mengyu Chu. Physics-informed learning of characteristic trajectories for smoke reconstruction. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2

[52] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE TIP*, 2004. 6

[53] Steffen Wiewel, Moritz Becher, and Nils Thuerey. Latent space physics: Towards learning the temporal evolution of fluid flow. In *CGF*, 2019. 2

[54] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4389–4398, 2024. 3

[55] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 6

[56] Hong-Xing Yu, Yang Zheng, Yuan Gao, Yitong Deng, Bo Zhu, and Jiajun Wu. Inferring hybrid neural fluid fields from

videos. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 6, 7, 8

[57] Guangming Zang, Ramzi Idoughi, Congli Wang, Anthony Bennett, Jianguo Du, Scott Skeen, William L Roberts, Peter Wonka, and Wolfgang Heidrich. Tomofluid: Reconstructing dynamic fluid from sparse view videos. In *CVPR*, 2020. 2

[58] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 2

[59] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[60] Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y. Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T. Freeman. PhysDreamer: Physics-based interaction with 3d objects via video generation. In *European Conference on Computer Vision*. Springer, 2024. 3

[61] Licheng Zhong, Hong-Xing Yu, Jiajun Wu, and Yunzhu Li. Reconstruction and simulation of elastic objects with spring-mass 3d gaussians. In *European Conference on Computer Vision*, pages 407–423. Springer, 2024. 3