SmokeSVD: Smoke Reconstruction from A Single View via Progressive Novel View Synthesis and Refinement with Diffusion Models

Chen Li[1†] Shanshan Dong[2†] Sheng Qiu[2*] Jianmin Han[2]
Yibo Zhao[1] Zan Gao[1] Taku Komura[3] Kemeng Huang[3]
[1] Tianjin University of Technology [2] Zhejiang Normal University [3] University of Hong Kong

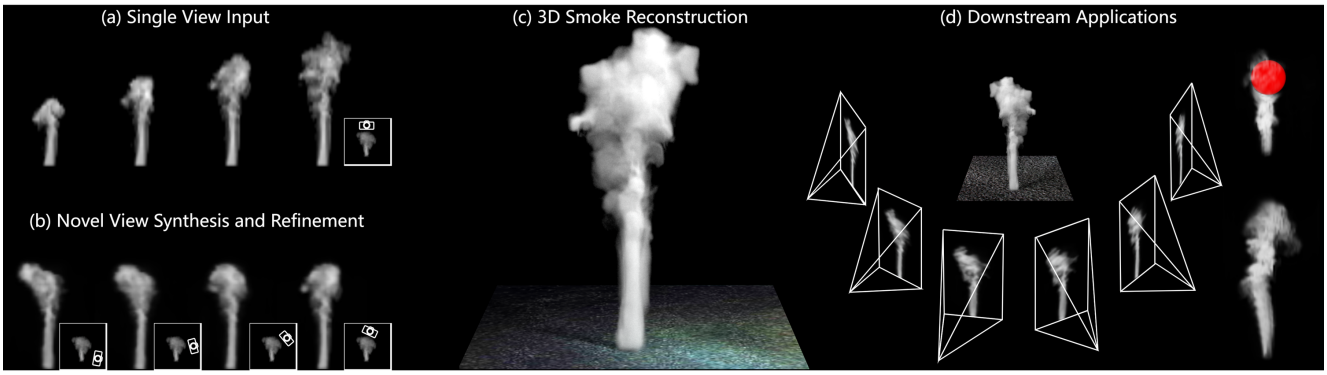

Figure 1. By leveraging physics-aware diffusion and refinement modules, our method progressively performs novel view synthesis (b) and 3D reconstruction (c) from a single-view input (a). When applied to downstream applications (d), our approach enables flexible novel view generation, re-simulation, and artist-driven control.

Abstract

Reconstructing dynamic fluids from sparse views is a long-standing and challenging problem, due to the severe lack of 3D information from insufficient view coverage. While several pioneering approaches have attempted to address this issue using differentiable rendering or novel view synthesis, they are often limited by time-consuming optimization under ill-posed conditions. We propose SmokeSVD, an efficient and effective framework to progressively reconstruct dynamic smoke from a single video by integrating the generative capabilities of diffusion models with physically guided consistency optimization. Specifically, we first propose a physically guided side-view synthesizer based on diffusion models, which explicitly incorporates velocity field constraints to generate spatio-temporally consistent side-view images frame by frame, significantly alleviating the ill-posedness of single-view reconstruction. Subsequently, we iteratively refine novel-view images and reconstruct 3D density fields through a progressive multi-stage process that renders and enhances images from increasing viewing angles, generating high-quality multi-view sequences. Finally, we estimate fine-grained density and velocity fields via differentiable advection by leveraging the Navier-Stokes equations. Our approach supports re-simulation and downstream applications while achieving superior reconstruction quality and computational efficiency compared to state-of-the-art methods.

*Corresponding Author. †Equal contributions.

1. Introduction

Smoke reconstruction and motion estimation from RGB videos has always been an important issue in a wide range of fields, including computer graphics and vision [35], atmospheric physics [2], optics [13], medicine [3]. Despite the rapid development of dynamic radiance fields, it is cumbersome and sometimes impractical for non-specialist to capture multi-view images of smoke phenomena in non-laboratory environments, impeding the widespread applications of relevant techniques, therefore, efficiently reconstructing and understanding smoke phenomena from highly sparse captured images [23] is of great value.

Existing solutions [6, 11, 24, 45] for sparse-view fluid capture integrate physically-based and geometric priors but are time-consuming. For single-view reconstruction, Franz et al. [8] introduced physical priors via differentiable rendering, but remains computationally expensive. Recent

works [4, 10] employ diffusion models to generate novel view videos, alleviating the ill-posed problem. However, combining multi-view diffusion models with sparse-view reconstruction faces two challenges: (1) limited multi-view consistency, where diffusion models produce low-quality inconsistent images [4, 49], and (2) insufficient incorporation of physical priors to guide generative models for complex smoke dynamics and external inflows.

In this paper, we propose SmokeSVD for efficient high-quality smoke reconstruction from single-view video. Inspired by recent 3D generation work [49], we first synthesize side-view sequences from front-view input using diffusion models guided by spatial and temporal priors. We then progressively generate novel views from near to far. Each iteration reconstructs a coarse 3D density field, then refines novel views using differentiable rendering and UNet3+ [17] for visual fidelity and temporal coherence. Finally, we reconstruct fine-grained density and velocity fields, and infer inflow states to support downstream applications.

Unlike recent sparse-view methods [10] that first generate multi-view images then reconstruct 3D, leading to shape-appearance ambiguity from insufficient consistency, we advocate a multi-stage strategy cyclically utilizing 2D diffusion synthesis, spatio-temporal refinement, and coarse/fine-grained 3D reconstruction. This exploits both high-quality 2D diffusion outputs and 3D volumetric consistency. Our progressive generation is guided by multi-view consistent optimization for temporally coherent sequences with minimal computation. Thus, SmokeSVD outperforms state-of-the-art in both quality and efficiency.

Our contributions are summarized as follows:

- We propose a novel and efficient smoke reconstruction framework from a single view by incorporating multi-stage 2D novel view synthesizer/refinement and coarse/fine-grained 3D reconstruction. The proposed framework allows us to rapidly infer velocity field and dynamic inflow states, supporting re-simulation of the input phenomena, or generation of new visual effects.
- We propose a method to synthesize a visually plausible side view image sequences based on front view sequences using the diffusion model. To guarantee reasonable smoke motion, we incorporate 3D predicted density and velocity fields as physical guidance into the denoising process for enhancing temporal consistency and producing physically-plausible smoke motion.
- We present a novel view refinement approach to progressively produce high-quality and consistent multi-view image sequences by injecting multiple view information and coarse 3D density field. Compared to direct multi-view diffusion models, our refinement approach achieves a better balance between computational efficiency and reconstruction robustness.

2. Related Work

Fluid Simulation and Reconstruction. Physically-based fluid simulation has a long history in computer graphics [22, 31, 33, 34, 46, 48]. Please refer to [35] for a comprehensive survey. As the inverse problem, fluid reconstruction are challenging [38, 39]. Conventional methods rely on specialized devices (e.g., Schlieren photography [1], structured light [12], light field probes [19]), or passive techniques [27, 41]. Gregson et al. [11] coupled fluid simulation into flow tracking to reconstruct temporally coherent velocity fields. Similarly, Eckert et al. [6, 7] adopted specific simulator components to infer unknown physical quantities.

Recently, neural rendering has gained attention in fluid reconstruction [25]. PINF [5] introduces a hybrid representation for dynamic fluid scenes with static obstacles. HyFluid [43] advocates hybrid neural fields to jointly infer density and velocity from multi-view videos. PICT [35] proposes a neural characteristic trajectory field with spatial-temporal NeRF. However, neural rendering faces challenges capturing high-frequency information from sparse views, often producing over-smooth results.

For single-view reconstruction, GlobTrans [8] employs strict differentiable physical priors. Franz et al. [9] applied central constraints with differentiable rendering to ensure smoke appearance in novel views. FluidNexus [10] reconstructs smoke by synthesizing multi-view videos. However, consistency issues may exist among viewpoints generated by [21]. Our method alleviates the ill-posed problem by generating side-view sequences and uses progressive refinement to ensure multi-view consistency.

Novel View Synthesis with 2D Diffusion Models. Since [15], diffusion models have been widely applied to multiple domains [16, 18, 30, 40, 44]. Through implicit representations [26] and sampling techniques [47], diffusion models achieve high quality and speed. Several studies have applied diffusion models to novel view synthesis [21, 32, 36]. Zero-1-to-3 [21] and 3DiM [36] concatenate conditional information as model inputs, while pose-guided diffusion uses cross-attention. However, end-to-end generation may lack consistency across viewpoints. To improve consistency, multiple works [20, 28, 29, 37, 42] have been proposed. Zero123++ [28] learns joint distribution by combining multi-view images into one. MVDream [29] enhances consistency via 3D self-attention and MLPs for camera information. Consistent123 [37] introduces cross-view and shared self-attention for structural consistency. ConsistNet [42] back-projects features into 3D space using multi-view geometry. ViVid-1-to-3 [20] reformulates it as video generation, introducing video diffusion priors. However, existing methods cannot be directly applied to smoke due to its complex physical properties.

3. Method

3.1. Overview

Our pipeline is illustrated in Fig. 2. Given a single-view video of T frames, we treat it as the front-view sequence $w^t_{\angle 0^\circ}$, where t is the frame number and $\alpha = \angle 0^\circ$ denotes the offset angle from front view. We propose a side-view synthesizer SvDiff based on diffusion models to synthesize side-view video $w^t_{p,\angle 90^\circ}$ from $w^t_{\angle 0^\circ}$ with reasonable spatial distribution, temporal evolution and appearance. Then, a coarse-grained density generator $\mathcal{G}^c_\rho$ generates a rough 3D density field $\rho_{r,c}$ from $w^t_{\angle 0^\circ}$ and $w^t_{p,\angle 90^\circ}$. We progressively rotate the camera along the horizontal plane to render novel view images (e.g., $w^t_{r,\angle 45^\circ}$, $w^t_{r,\angle 135^\circ}$), and refine them frame by frame with novel view refinement module NvRef. Benefiting from 3D spatial distribution constraint from ρ_c and temporal-spatial correlation from UNet3+, NvRef produces multi-view consistent images. With multiple views, we employ a fine-grained density generator $\mathcal{G}^f_\rho$ to reconstruct high-quality density field $\rho_{r,f}$, jointly estimating velocity fields $\mathbf{u}$ and inflow states ρ_{in} via differentiable advection operator $\mathcal{A}$, ensuring reconstruction satisfies long-term physical constraints. Finally, we can re-simulate the input smoke and support downstream applications, e.g., novel view synthesis, artist control.

3.2. Physically-Aware Side-View Synthesizer

While substantial progress has been made in generalizable novel-view synthesis, most approaches lack effective physically-aware priors for complex volumetric phenomena. Smoke poses unique challenges due to its semi-transparent appearance and complex dynamics. First, ensuring spatiotemporal consistency across synthetic sequences is difficult, as current methods often produce visual artifacts including temporal flickering and motion incoherence. Second, maintaining cross-view consistency between input frontal and generated side views requires sophisticated modeling of shared volumetric properties, as both views represent different projections of the same 3D volume with consistent spatial distributions and appearance.

We incorporate physical and visual priors into our side-view synthesizer SvDiff to address these challenges. SvDiff extends image generation diffusion models [15] to handle smoke sequences frame-by-frame for temporal coherence. Inspired by classifier-free guidance [14], we use side-view images of two previous frames $w^{t-1}_{\angle 90^\circ}, w^{t-2}_{\angle 90^\circ}$ and current front-view image $w^t_{\angle 0^\circ}$ as condition to train SvDiff:

$$c^t = w^t_{\angle 0^\circ} \oplus w^{t-1}_{\angle 90^\circ} \oplus w^{t-2}_{\angle 90^\circ}, \tag{1}$$

where $\oplus$ denotes concatenation. For initial frames ($t < 2$), we train another synthesizer with condition $c^0 = w^0_{\angle 0^\circ} \oplus w^1_{\angle 0^\circ}$. SvDiff is trained by minimizing:

$$\mathcal{L}_{noise} = \|\epsilon - \epsilon_\theta(w^t_{\angle 90^\circ}, c^t, s)\|^2. \tag{2}$$

During training, SvDiff synthesizes side-view image $w^t_{\angle 90^\circ}$ from ground truth side-view images $w^{t-1}_{\angle 90^\circ}, w^{t-2}_{\angle 90^\circ}$. However, during inference, SvDiff uses previously synthesized frames as input conditions, which progressively accumulates errors over time.

To reduce accumulated error and ensure long-term stability, we propose a multi-frame training scheme enabling SvDiff to learn from both historically generated and rendered images of reconstructed density fields, as shown in Fig. 3. We re-formulate Eq. 1 as $c^t = w^t_{\angle 0^\circ} \oplus w^{t-1}_{c,\angle 90^\circ} \oplus w^{t-1}_{r,\angle 90^\circ} \oplus w^{t-2}_{c,\angle 90^\circ} \oplus w^{t-1}_{r,\angle 90^\circ}$, where w_c is the synthesized side-view image from SvDiff. Since diffusion training predicts noise from the forward process, in multi-frame training we estimate generated images from noise. Based on Eq 2, the estimated clean image is:

$$w_{\angle 90^\circ} \approx w_{c,\angle 90^\circ} = \frac{w_{s,\angle 90^\circ} - \sqrt{1-\bar{\alpha}_s}\epsilon_\theta}{\sqrt{\bar{\alpha}_s}}, \tag{3}$$

where $w_{s,\angle\alpha}$ denotes a noisy image at diffusion step s and viewpoint α. When s is not labeled, it defaults to zero, indicating a clean image.

Unlike traditional diffusion models performing one forward process per batch, our multi-frame training performs multiple forward processes per batch. In each forward process, SvDiff estimates a clean image from the noisy image and uses it as condition for the next forward process. Through multiple forward diffusions, SvDiff learns from historically generated information, improving long-term stability.

To incorporate physical and visual priors and guide SvDiff toward physically faithful results, we introduce a guidance module imposing targeted constraints on denoising. We set a threshold TQ to determine when the guidance is applied: if $s \geq TQ$, the noise level is too high to extract meaningful physical information between consecutive frames, so the guidance is disabled; otherwise, the guidance module is activated and incorporated into the training objective. Specifically, the guidance consists of three loss terms: visual, velocity and spatial constraints, that collectively steer the model toward more accurate and realistic generation.

Visual Constraint. We use L_2 loss to measure difference between predicted clean image $\hat{x}^i_0$ and ground truth x^i_0, where i denotes the multi-frame training iteration index. This loss $\mathcal{L}_{img} = \|x^i_0 - \hat{x}^i_0\|^2$ penalizes pixel-wise discrepancies, ensuring high fidelity.

Velocity Constraint. To further ensure physically plausible smoke dynamics over time, we introduce velocity constraints between consecutive frames, penalizing both the divergence and abrupt changes in the velocity fields. To infer the 3D velocity field from 2D images, we first use a density generator $\mathcal{G}\rho$ (see Sec. 3.3) to reconstruct a coarse-grained 3D density field $\rho^i_{r,c}$ from the input front-view im-

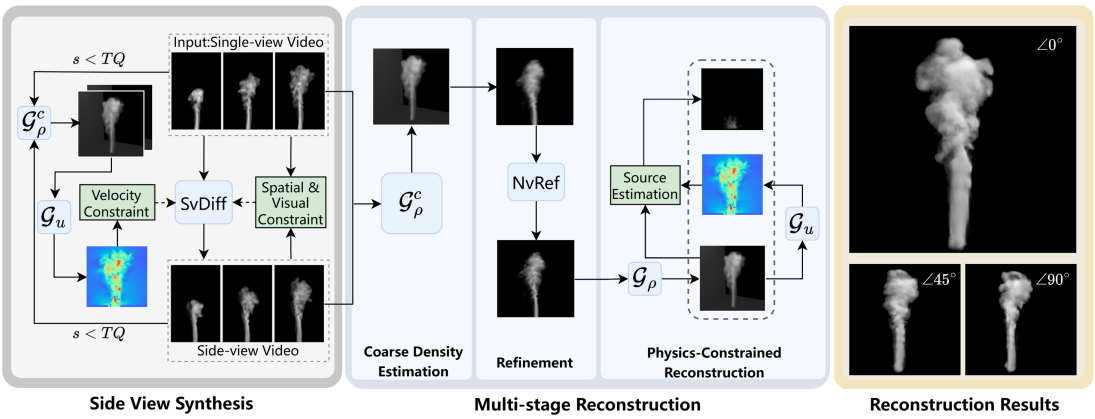

Figure 2. Overview of SmokeSVD. We categorize view angles into three types: input as front view ($\alpha = \angle 0°$), synthesized images from SvDiff as side view ($\alpha = \angle 90°$), and all others as novel views. Given front-view input, we first synthesize side-view sequences guided by spatial, visual, and velocity constraints via density and velocity reconstruction. We then iteratively estimate coarse 3D density and refine novel-view sequences, progressively introducing views from near to far. Our pipeline outputs 3D density, velocity fields, and dynamic inflow. Physical priors guide both SvDiff and NvRef modules for physically accurate and visually realistic results.

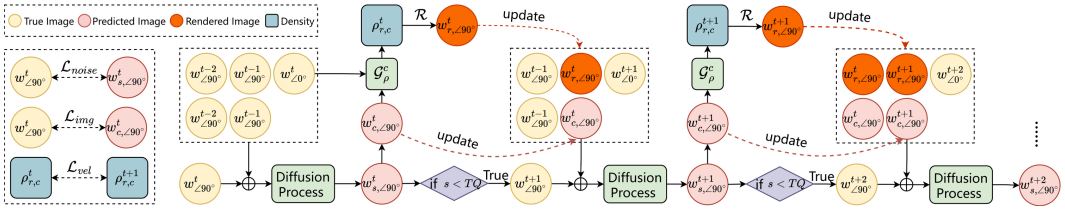

Figure 3. Frame-by-frame training of the side-view synthesizer via feature fusion of adjacent frames. In the forward diffusion process, a clean image $w_{c,\angle 90°}$ is estimated from the noisy image $w_{s,\angle t,90°}$, and this estimated clean image serves as one of the conditional images for the next forward diffusion process. The figure demonstrates the forward diffusion training process for three consecutive frames.

age and the predicted clean side-view image, defined as $\rho^i_{r,c} = \mathcal{G}_\rho(w^{i+t}_{\angle 0°}, w^{i+t}_{c,\angle 90°})$. Based on these reconstructed density fields from consecutive frames, we then employ a velocity generator $\mathcal{G}_u$ (see Sec. C.5 in supplementary) to estimate the velocity field as $\mathbf{u}^{i-1} = \mathcal{G}_u(\rho^{i-1}, \rho^i_{r,c})$. The velocity constraint consists of two terms:

$$\mathcal{L}_{vel} = \|\nabla \cdot \mathbf{u}^{i-1}\|^2 + \|\nabla \mathbf{u}^{i-1}\|^2, \tag{4}$$

where the first term enforces incompressibility and the second promotes smoothness, preventing temporal artifacts.

Spatial Constraint. To ensure that the generated side-view image $w_{c,\angle 90°}$ is consistent with the input image $w_{\angle 0°}$ in spatial distribution, we design a spatial distribution constraint based on the estimated clean image. The purpose of this loss term is to make SvDiff more attentive to the spatial distribution differences between $w_{c,\angle 90°}$ and $w_{\angle 0°}$, thereby guiding SvDiff to generate features that are closer to ground truth:

$$\mathcal{L}_{sp} = \|H(w_{c,\angle 90°}) - H(w_{\angle 0°})\|^2, \tag{5}$$

where $w_{c,\angle 90°}$ is the predicted clean image, $H()$ is the operation of summing each row of an image along the width direction. For an $H \times W$ image, this operation transforms it into a vector of size $H \times 1$.

The overall loss function can be formulated as:

$$\mathcal{L}_{SvDiff} = \lambda_{noise}\mathcal{L}_{noise} + \lambda_{img}\mathcal{L}_{img} + \lambda_{sp}\mathcal{L}_{sp} + \lambda_{vel}\mathcal{L}_{vel}. \tag{6}$$

By gradient steps on these losses, SvDiff generates physically accurate and visually realistic side-view predictions. Our multi-frame training strategy explicitly encourages temporal consistency, ensuring coherent and stable smoke motion.

3.3. Progressive Novel View Refinement

Based on 2D images from various views, we can train a density generator $\mathcal{G}_\rho$ to estimate a 3D density field of smoke as:

$$\rho^t_r = \mathcal{G}_\rho(I^t), \quad I^t = w^t_{\angle 0°} \oplus w^t_{p,\angle 90°} \oplus \cdots. \tag{7}$$

Here $\mathcal{G}_\rho$ adopts the UNet3+ architecture [17] and extends the 2D convolutions in UNet3+ to 3D convolutions. Please refer to Appendix for more details. Since estimating density along ray direction from 2D images is difficult, we design

the following loss for $\mathcal{G}_\rho$:

$$\mathcal{L}_{\mathcal{G}_\rho} = \lambda_\rho \|\rho_r^t - \rho^t\|^2 + \lambda_{in} \sum_{\alpha \in \mathbb{A}} \|\mathcal{R}(\rho_r^t, \alpha) - \mathcal{R}(\rho^t, \alpha)\|^2 + \lambda_{un} \sum_{\alpha \notin \mathbb{A}} \|\mathcal{R}(\rho_r^t, \alpha) - \mathcal{R}(\rho^t, \alpha)\|^2, \tag{8}$$

where ρ denotes the ground truth density, $\mathbb{A}$ denotes the set of input view angles (e.g., $\angle 0°, \angle 90°$), $\mathcal{R}(\rho, \alpha)$ is the differentiable rendering operator that renders density field ρ at the viewing angle α. The second and third terms correspond to images from input and unknown viewpoints, respectively. For the ScalarFlow dataset, we set λ_ρ to zero and use the reconstructed results from [7] as ρ for rendering. In our pipeline, when the number of input images is less than 16, we call it coarse-grained density generator $\mathcal{G}_\rho^c$; when the number of input images equals to 16, we call it fine-grained density generator $\mathcal{G}_\rho^f$.

After generating side-view $w_{p,\angle 90°}^t$ from $w_{\angle 0°}^t$ with SvDiff, we employ $\mathcal{G}_\rho^c$ to produce rough density field $\rho_{r,c}^i$. Although $\mathcal{G}_\rho^c$ is trained using the rendered image loss $\|\mathcal{R}(\rho^t, \alpha) - \mathcal{R}(\rho_{r,c}^t, \alpha)\|^2$ to learn the smoke shape in novel views, in the absence of enough views, $\rho_{r,c}^t$ still exhibits blurriness in novel views.

To enhance details and reduce blurriness in $\rho_{r,c}$, we introduce novel view refinement module NvRef based on UNet3+:

$$\begin{aligned} res_\alpha^t &= \text{NvRef}\big(w_{r,\angle\alpha-\beta}^t \oplus w_{r,\angle\alpha+\beta}^t \oplus w_{r,\angle\alpha}^t \\ &\qquad \oplus \downarrow w_{f,\angle\alpha}^{t-1} \oplus \downarrow w_{f,\angle\alpha}^{t-2}\big), \\ w_{f,\angle\alpha}^t &= res_\alpha^t + w_{r,\angle\alpha}^t, \end{aligned} \tag{9}$$

where α is the target angle to be refined, β is the angular offset relative to α, $\downarrow$ is 2x downsampling operation, and res is the residual error.

NvRef is designed to maintain the spatial distribution consistency and perceptual similarity between ground truth and refined novel images, whose overall loss function is formulated as:

$$\begin{aligned} \mathcal{L}_{NvRef} &= \lambda_{mse}\|w_{f,\angle\alpha}^t - w_{\angle\alpha}^t\|^2 + \lambda_{l1}\|w_{f,\angle\alpha}^t - w_{\angle\alpha}^t\| \\ &+ \lambda_{res}\|Mean(res_\alpha^t)\|^2 + \lambda_{sp}\|H(w_{f,\angle\alpha}^t) - H(w_{\angle\alpha}^t)\|^2 \\ &+ \lambda_{psnr}\|PSNR(w_{f,\angle\alpha}^t) - PSNR(w_{\angle\alpha}^t)\|^2, \end{aligned} \tag{10}$$

where the first three terms penalize $L2$, $L1$ and residual error, the fourth is spatial constraint similar to SvDiff, and the last computes the peak signal-to-noise ratio (PSNR) discrepancy.

Subsequently, we iteratively invoke $\mathcal{G}\rho$ and NvRef to rotate the camera along the horizontal plane, progressively rendering and refining additional novel view images. In our experiments, we set the maximum number of views to 16 to achieve a balance between computational efficiency and reconstruction quality. Since rendered images from adjacent views tend to exhibit similar shapes and reduced blurriness, we further categorize these 16 views into four types, namely clear, near, mid, and far views, based on their relative positions to the front and side views, as illustrated in Fig. 4.

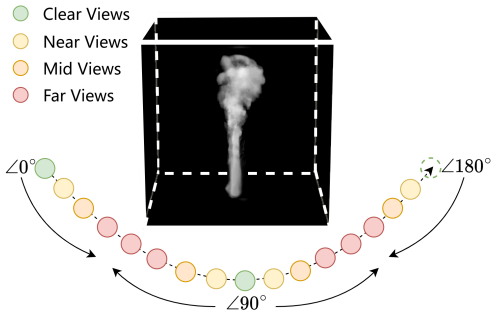

Figure 4. The progressive scheme for novel view refinement begins with clear views and incrementally rotates the camera to render and refine novel-view images from near, mid, and far views.

During multi-stage refinement process, we sequentially render images at near, mid, and far views from the density field reconstructed in the previous stage, and refine these images using NvRef. The refined images, together with the blurred images from the remaining views, are then used to reconstruct the density field for the next stage of refinement. By iteratively combining coarse 3D density estimation with targeted refinement of novel view images, our progressive novel view refinement strategy gradually expands the set of reliable views. Finally, we leverage multi-view information to jointly reconstruct the density, velocity, and inflow of the input smoke phenomena. See supplementary for details.

4. Evaluations and Ablation Study

4.1. Evaluation

Evaluation on ScalarFlow. To validate the applicability of our method in real-world scenarios, we conducted evaluations on the ScalarFlow dataset [7]. This dataset captures real-world smoke images using five cameras uniformly distributed along a $120°$ arc and provides 3D density and velocity fields. However, these 3D data cannot be directly used for quantitative comparison, so our subsequent evaluations are based solely on images.

In our experiments, we used one of the pre-processed images from the five viewpoints in the ScalarFlow as input to reconstruct smoke density fields at a resolution of $64 \times 112 \times 64$. For comparison, we interpolated the density fields reconstructed by all methods to the same resolution of 64^3 and rendered images at the input front view ($\angle 0°$) and side view ($\angle 90°$) using Houdini. We conducted qualitative comparisons with state-of-the-art methods, as shown

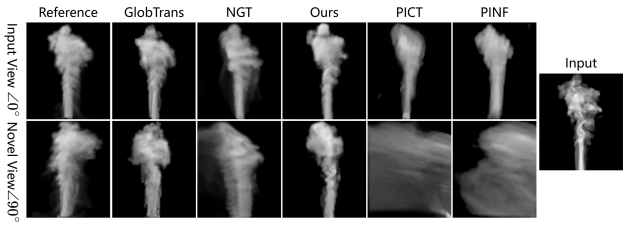

Figure 5. Qualitative comparison based on different methods on ScalarFlow. Our method matches the appearance pattern of the input image at the front view, and produces a reasonable shape in the side view.

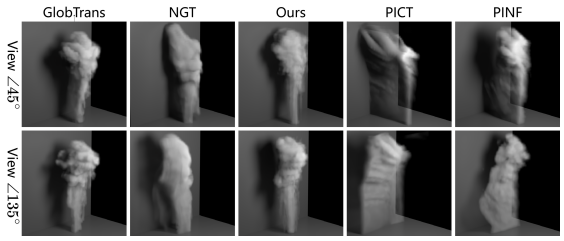

Figure 6. Qualitative comparison on ScalarFlow.

in Figs. 5 and 6. Due to limited single-view input, PICT and PINF exhibit varying degrees of blurring in the depth direction, even affecting the reconstruction quality at the front view. In contrast, GlobTrans achieves the best perceptual quality (as documented in Table 1) at the side view and performs well across multiple novel views, at the expense of heavy computational cost. The results of NGT match well with inputs through differentiable rendering and adversarial learning techniques, achieving the lowest root mean square error at novel views. However, it introduces artifacts in certain views (90° in Fig. 5) and presents overly smooth smoke at some angles (135° in Fig. 6).

These results indicate the difficulty of balancing reconstruction quality and computational efficiency from single-view input. Our method matches input images well while maintaining reasonable smoke appearance and rich details in novel views at minimal cost. From a perceptual quality perspective, our method performs excellently, second only to GlobTrans. However, as shown in Table 1, mean squared error cannot comprehensively measure novel view quality—PICT and PINF exhibit unreasonable appearance yet achieve similar MSE to our method.

Table 1. Quantitative comparison on ScalarFlow.

Algorithm	Input RMSE↓	SSIM ↑	PSNR↑	LPIPS↓	Side RMSE↓	STYLE↓	Time for 120 Steps
GlobTrans	**0.0101**	**0.9975**	**40.1560**	**0.0054**	**0.0352**	0.2167	>30h
NGT	0.0289	0.9539	31.0727	0.0655	0.0544	0.2499	5mins
PICT	0.0315	0.9252	30.5447	0.1332	0.0743	0.7259	/
PINF	0.0872	0.8715	21.3005	0.1020	0.1101	0.6335	/
Ours	0.0127	0.9868	38.0790	0.0223	0.0853	**0.2071**	15mins

Tables 2 and 3 compare our method with FluidNexus [10] and NeuSmoke [25]. Our method significantly outperforms both approaches on input view reconstruction across all metrics. Compared to NeuSmoke, we achieve substantial improvements on novel views, demonstrating that our progressive refinement strategy, which explicitly synthesizes side views, effectively alleviates single-view ambiguity better than implicit neural rendering from sparse views. For FluidNexus, while our novel view performance is slightly lower (as its multi-view diffusion inherently maintains cross-view consistency), we achieve superior input quality through progressive side-view refinement and avoid sensitivity to post-processing threshold selection. Our novel view refinement module further enhances quality through multi-view consistency constraints, producing accurate reconstructions without requiring hyperparameter tuning, demonstrating superior robustness. The qualitative comparison is shown in Figs. 7 and 8.

Table 2. Comparison with FluidNexus (various post-processing thresholds) on ScalarFlow. Averaged over five scenes, novel views from four non-frontal cameras.

Algorithm	Input RMSE↓	SSIM ↑	PSNR↑	LPIPS↓	Novel RMSE↓	SSIM ↑	PSNR↑	LPIPS↓
FN w/o th	0.0473	0.7924	26.6722	0.2192	0.0807	0.1651	21.9411	0.1881
FN th=0.05	0.0303	0.8858	30.8166	0.1912	0.0702	0.3187	23.2492	0.2665
FN th=0.1	0.0388	0.9159	30.7635	0.1217	**0.0565**	**0.8419**	**25.3569**	0.1575
FN th=0.15	0.0361	0.8968	29.3974	0.1402	0.0582	0.8435	25.1001	**0.1573**
FN th=0.2	0.0428	0.8757	27.8309	0.1628	0.0598	**0.8419**	23.9521	0.1669
ours	**0.0172**	**0.9764**	**35.5504**	**0.0586**	0.0690	0.7871	23.4393	0.1829

Table 3. Comparison with Neusmoke on ScalarFlow. Front and side views as input, three novel views for evaluation (averaged).

Algorithm	RMSE↓	SSIM ↑	PSNR↑	LPIPS↓
NeuSmoke	0.0514	0.8750	26.5031	0.1131
Ours	**0.0331**	**0.9038**	**30.0384**	**0.0991**

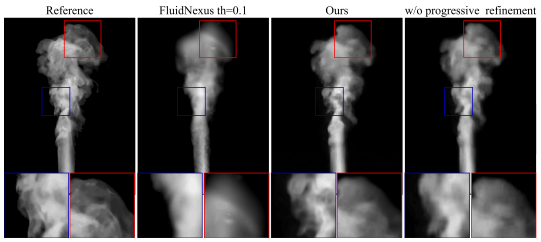

Figure 7. Comparison with FluidNexus on ScalarFlow.

Evaluation on Synthetic Data. We evaluated our method on a synthetic smoke dataset generated with the rendering operator [8]. The synthetic dataset provides precise 3D physical fields and smooth motion compared to real-world scenes. Table 4 shows performance comparison with baseline methods using image metrics.

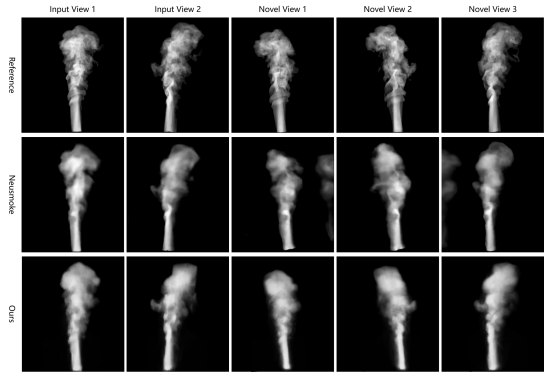

Figure 8. Comparison with NeuSmoke on ScalarFlow.

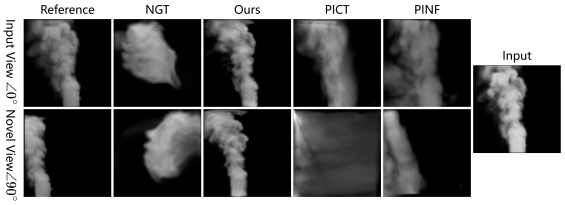

Figure 9. Qualitative comparison on the synthetic dataset.

Fig. 9 shows qualitative comparison with state-of-the-art methods. Similar to ScalarFlow results, PICT and PINF exhibit blurriness in side views. Additionally, NGT's inaccurate inflow estimation causes reconstructed density to gradually deviate from input over time. See Sec. E in supplementary for more complex phenomena.

Table 4. Quantitative comparison on the synthetic dataset.

Algorithm	Input RMSE↓	SSIM ↑	PSNR↑	LPIPS↓	Side RMSE↓	STYLE↓
NGT	0.1844	0.7754	15.6521	0.2227	0.2714	1.2242
PICT	0.1625	0.7608	16.2969	0.2153	0.2913	1.5585
PINF	0.2286	0.6293	13.2970	0.2259	**0.2468**	1.1321
Ours	**0.0395**	**0.9645**	**28.1332**	**0.0293**	0.3821	**1.0790**

Generalization Performance. To evaluate generalization, we apply our method to smoke without inflow and horizontal plume (Figs. 10 and 11), unseen in training. Results show effectiveness on these previously unseen scenarios.

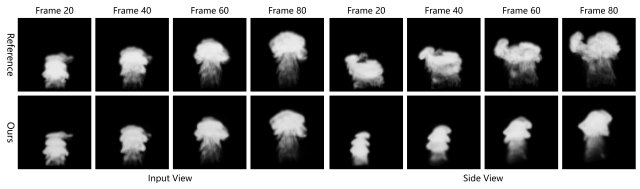

Figure 10. Reconstruction results for a bunny-shaped smoke scenario without inflow.

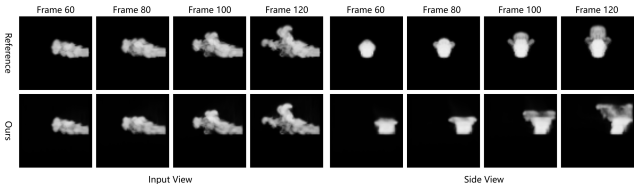

Figure 11. Reconstruction result for a horizontal plume scenario.

4.2. Ablation Study

Ablation on Side-view Synthesizer. To evaluate physical priors in SvDiff, we remove noise threshold, velocity loss, gradient loss, and 3D reconstruction ("w/o threshold", "w/o vel", "w/o grad", "w/o reconstruction"). Table 5 shows removing these constraints degrades performance. Note that velocity-based temporal correction slightly reduces input view LPIPS.

Table 5. Ablation studies on SvDiff.

Algorithm	Input RMSE↓	SSIM ↑	PSNR↑	LPIPS↓	Side RMSE↓	STYLE↓
w/o threshold	0.0089	0.9946	41.8412	0.0096	0.0990	0.2139
w/o vel	0.0100	0.9929	41.6814	0.0069	0.1032	0.2074
w/o grad	0.0091	0.9940	42.0804	**0.0061**	0.1025	0.2025
w/o divergence	0.0136	0.9886	40.9043	0.0114	0.1816	0.4831
w/o reconstruction	0.0106	0.9934	41.4763	0.0077	0.1025	0.3118
Ours	**0.0062**	**0.9955**	**44.5518**	0.0075	**0.0899**	**0.1892**

Fig. 13 visualizes the divergence of reconstructed velocity fields to demonstrate the velocity term's impact. Incorporating velocity loss produces smoother and more stable smoke dynamics, preventing artifact flickering. To evaluate visual priors, we ablated rendered density images as SvDiff input. Fig. 14 shows omitting these images causes noticeable errors in long-term synthesis.

Ablation on Novel View Refinement. To assess the impact of novel view refinement, we performed ablation studies by (1) removing the entire refinement process, (2) replacing the multi-stage progressive refinement with a single-pass refinement for all novel views, and (3) remove residual loss. These variants are denoted as "w/o Refinement", "w/o Progressive," "w/o Res Loss", respectively. The quantitative and qualitative results are presented in Table 6, Figs. 12 and 15. Our progressive refinement approach achieves richer visual details and appearance consistency.

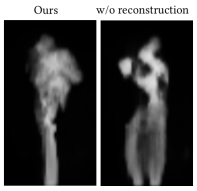

Figure 14. Ablation on rendered density input for side-view synthesis.

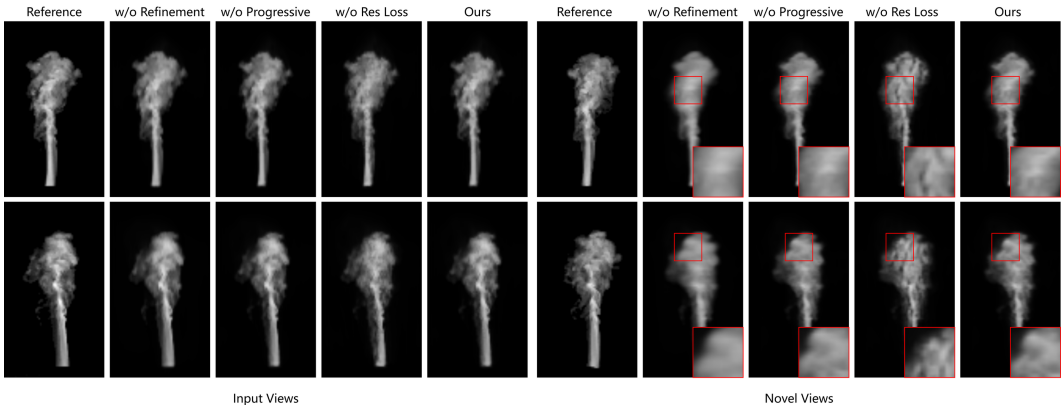

Figure 12. Ablation on novel view refinement. From top to bottom: reference, results without refinement, without progressive refinement, without res loss and with NvRef. Red boxes show close-ups.

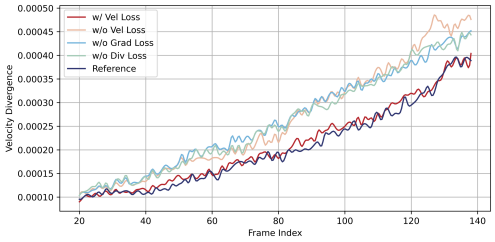

Figure 13. Comparison of the divergence of reconstructed velocity fields by SvDiff with different loss functions at various time steps.

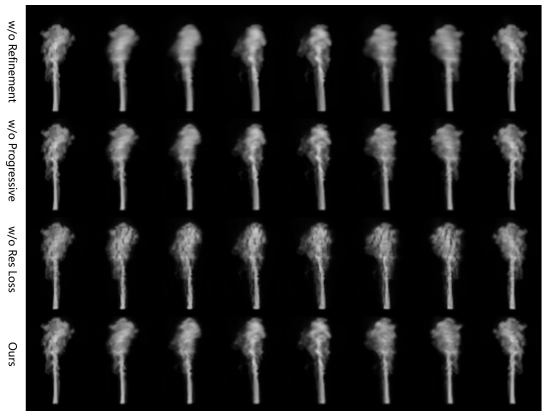

Figure 15. Refined results across novel views. Each row shows renderings uniformly distributed from $\angle 0^\circ$ to $\angle 175^\circ$.

Ablation on Key Components. To evaluate key components, we conduct two ablation studies: (1) removing novel view refinement, and (2) replacing our side-view synthesizer with NGT [9]. Fig. 16 shows novel views before and after refinement, demonstrating that refinement produces richer details and reduces blurriness.

Table 6. Ablation on novel view refinement. Views 0 (front) and 3 (side) as input, remaining views for evaluation.

Algorithm	MSE ↓	SSIM ↑	PSNR ↑	LPIPS ↓
w/o Refinement	0.0196	0.7454	18.7490	0.1808
w/o Progressive	0.0192	**0.7559**	<u>18.7902</u>	**0.1704**
w/o Res Loss	**0.0168**	0.7126	18.5066	0.1789
Ours	<u>0.0190</u>	<u>0.7559</u>	**18.7978**	<u>0.1757</u>

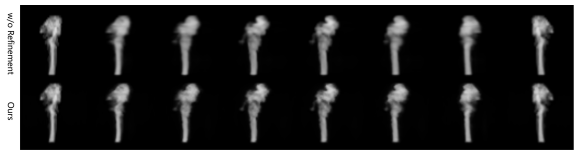

Figure 16. Ablation study of the refinement model. Each row shows renderings uniformly distributed from $\angle 0^\circ$ to $\angle 175^\circ$.

5. Conclusion and Future Work

We present a framework for 3D smoke reconstruction from single-view input by integrating physical priors and spatiotemporal constraints. Our approach overcomes single-view ambiguity through a diffusion-based side-view synthesizer and novel view refinement module, providing rich multi-view information for density and velocity reconstruction. Experiments on synthetic and real-world datasets demonstrate superior balance between quality and efficiency. Our framework maintains accurate input matching while preserving reasonable smoke appearance and rich details in novel views. Future work could address more complex fluids, vertical multi-view fusion, and higher-order physical constraints.

Acknowledgements

This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant

No.ZCLQN26F0204, the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No.VRLAB2025C05), National Natural Science Foundation of China (No.U25A20444, No.62372325, No.62402255, No.62502344), Natural Science Foundation of Tianjin Municipality (No.23JCZDJC00280), Shandong Provincial Natural Science Foundation (No.ZR2024QF020), Shandong Province National Talents Supporting Program (No.2023GJJLJRC-070), Young Talent of Lifting engineering for Science and Technology in Shandong (No.SDAST2024QTB001), Shandong Project towards the Integration of Education and Industry (No.2024ZDZX11).

References

[1] Bradley Atcheson, Ivo Ihrke, Wolfgang Heidrich, Art Tevs, Derek Bradley, Marcus Magnor, and Hans-Peter Seidel. Time-resolved 3d capture of non-stationary gas flows. *ACM Transactions on Graphics (TOG)*, 27(5):1–9, 2008. 2

[2] CM Carrico, MD Petters, SM Kreidenweis, AP Sullivan, GR McMeeking, EJT Levin, G Engling, WC Malm, and JL Collett Jr. Water uptake and chemical composition of fresh aerosols generated in open burning of biomass. *Atmospheric Chemistry and Physics*, 10(11):5165–5178, 2010. 1

[3] Long Chen, Wen Tang, Nigel W John, Tao Ruan Wan, and Jian Jun Zhang. De-smokegcn: generative cooperative networks for joint surgical smoke detection and removal. *IEEE Transactions on Medical Imaging*, 39(5):1615–1625, 2019. 1

[4] Yabo Chen, Jiemin Fang, Yuyang Huang, Taoran Yi, Xiaopeng Zhang, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Cascade-zero123: One image to highly consistent 3d with self-prompted nearby views. In *European Conference on Computer Vision*, pages 311–330. Springer, 2024. 2

[5] 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 Transactions on Graphics (ToG)*, 41(4):1–14, 2022. 2

[6] M-L Eckert, Wolfgang Heidrich, and Nils Thuerey. Coupled fluid density and motion from single views. *Computer Graphics Forum*, 37(8):47–58, 2018. 1, 2

[7] 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 Transactions on Graphics (TOG)*, 38(6):1–16, 2019. 2, 5

[8] 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. 1, 2, 6

[9] Erik Franz, Barbara Solenthaler, and Nils Thuerey. Learning to estimate single-view volumetric flow motions without 3d supervision. *arXiv preprint arXiv:2302.14470*, 2023. 2, 8

[10] Yue Gao, Hong-Xing Yu, Bo Zhu, and et al. Fluidnexus: 3d fluid reconstruction and prediction from a single video. *arXiv preprint arXiv:2503.04720*, 2025. 2, 6

[11] James Gregson, Ivo Ihrke, Nils Thuerey, and Wolfgang Heidrich. From capture to simulation: connecting forward and inverse problems in fluids. *ACM Transactions on Graphics (TOG)*, 33(4):1–11, 2014. 1, 2

[12] Jinwei Gu, Shree K Nayar, Eitan Grinspun, Peter N Belhumeur, and Ravi Ramamoorthi. Compressive structured light for recovering inhomogeneous participating media. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):1–1, 2012. 2

[13] Wenyu Han, Fuhao Zhang, Wensong Liu, Shunyao Huang, Can Gao, Zhiyin Ma, Fengnian Zhao, David LS Hung, Xuesong Li, and Min Xu. Three-dimensional reconstruction of smoke aerosols based on simultaneous multi-view imaging and tomographic absorption spectroscopy. *Optics Letters*, 50(4):1385–1388, 2025. 1

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3

[16] Jonathan Ho, Tim Salimans, Alexey Gritsenko, and et al. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[17] Huimin Huang, Lanfen Lin, Ruofeng Tong, and et al. Unet 3+: A full-scale connected unet for medical image segmentation. In *International Conference on Acoustics, Speech and Signal Pprocessing*, pages 1055–1059. IEEE, 2020. 2, 4

[18] Rongjie Huang, Jiawei Huang, Dongchao Yang, and et al. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR, 2023. 2

[19] Yu Ji, Jinwei Ye, and Jingyi Yu. Reconstructing gas flows using light-path approximation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2507–2514, 2013. 2

[20] Jeong-gi Kwak, Erqun Dong, Yuhe Jin, Hanseok Ko, Shweta Mahajan, and Kwang Moo Yi. Vivid-1-to-3: Novel view synthesis with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6775–6785, 2024. 2

[21] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 2

[22] Shusen Liu, Xiaowei He, Yuzhong Guo, Yue Chang, and Wencheng Wang. A dual-particle approach for incompressible sph fluids. *ACM Transactions on Graphics*, 43(3):1–18, 2024. 2

[23] Zhengyan Liu, Yong Hu, and Yue Qi. Modeling of smoke from a single view. In *2011 International Conference on Virtual Reality and Visualization*, pages 291–294. IEEE, 2011. 1

[24] Makoto Okabe, Yoshinori Dobashi, Ken Anjyo, and Rikio Onai. Fluid volume modeling from sparse multi-view images by appearance transfer. *ACM Transactions on Graphics (TOG)*, 34(4):1–10, 2015. 1

[25] Jiaxiong Qiu, Ruihong Cen, Zhong Li, Han Yan, Ming-Ming Cheng, and Bo Ren. Neusmoke: Efficient smoke reconstruction and view synthesis with neural transportation fields. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 2, 6

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, and et al. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[27] Jan FG Schneiders and Fulvio Scarano. Dense velocity reconstruction from tomographic ptv with material derivatives. *Experiments in fluids*, 57(9):139, 2016. 2

[28] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, and et al. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 2

[29] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2

[30] Yi Shi, Jingbo Wang, Xuekun Jiang, Bingkun Lin, Bo Dai, and Xue Bin Peng. Interactive character control with autoregressive motion diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4):1–14, 2024. 2

[31] Jos Stam and Eugene Fiume. Turbulent wind fields for gaseous phenomena. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, pages 369–376, 1993. 2

[32] Hung-Yu Tseng, Qinbo Li, Changil Kim, and et al. Consistent view synthesis with pose-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16773–16783, 2023. 2

[33] Zaili Tu, Chen Li, Zipeng Zhao, and et al. A unified mpm framework supporting phase-field models and elastic-viscoplastic phase transition. *ACM Transactions on Graphics*, 43(2):1–19, 2024. 2

[34] Sinan Wang, Yitong Deng, Molin Deng, and et al. An eulerian vortex method on flow maps. *ACM Transactions on Graphics (TOG)*, 43(6):1–14, 2024. 2

[35] 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, New York, NY, USA, 2024. Association for Computing Machinery. 1, 2

[36] Daniel Watson, William Chan, Ricardo Martin Brualla, and et al. Novel view synthesis with diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[37] Haohan Weng, Tianyu Yang, Jianan Wang, and et al. Consistent123: Improve consistency for one image to 3d object synthesis. *arXiv preprint arXiv:2310.08092*, 2023. 2

[38] Xueguang Xie, Yang Gao, Fei Hou, Tianwei Cheng, Aimin Hao, and Hong Qin. Fluid inverse volumetric modeling and applications from surface motion. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2

[39] Xueguang Xie, Yang Gao, Fei Hou, and et al. Dynamic ocean inverse modeling based on differentiable rendering. *Computational Visual Media*, 10(2):279–294, 2024. 2

[40] Zhen Xing, Qijun Feng, Haoran Chen, and et al. A survey on video diffusion models. *ACM Computing Surveys*, 57(2): 1–42, 2024. 2

[41] Jinhui Xiong, Ramzi Idoughi, Andres A Aguirre-Pablo, and et al. Rainbow particle imaging velocimetry for dense 3d fluid velocity imaging. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017. 2

[42] Jiayu Yang, Ziang Cheng, Yunfei Duan, and et al. Consistnet: Enforcing 3d consistency for multi-view images diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7079–7088, 2024. 2

[43] Hong-Xing Yu, Yang Zheng, Yuan Gao, and et al. Inferring hybrid neural fluid fields from videos. *Advances in Neural Information Processing Systems*, 36:63595–63608, 2023. 2

[44] Xin Yu, Ze Yuan, Yuan-Chen Guo, Ying-Tian Liu, Jianhui Liu, Yangguang Li, Yan-Pei Cao, Ding Liang, and Xiaojuan Qi. Texgen: a generative diffusion model for mesh textures. *ACM Transactions on Graphics (TOG)*, 43(6):1–14, 2024. 2

[45] Guangming Zang, Ramzi Idoughi, Congli Wang, and et al. Tomofluid: Reconstructing dynamic fluid from sparse view videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1870–1879, 2020. 1

[46] Meng Zhang, Shiguang Liu, Hanqiu Sun, and et al. Hybrid vortex model for efficiently simulating turbulent smoke. In *Proceedings of the 13th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*, pages 71–79, 2014. 2

[47] Qinsheng Zhang, Molei Tao, and Yongxin Chen. gddim: generalized denoising diffusion implicit models. In *International Conference on Learning Representations*, 2023. 2

[48] Junwei Zhou, Duowen Chen, Molin Deng, and et al. Eulerian-lagrangian fluid simulation on particle flow maps. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024. 2

[49] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, and et al. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10324–10335, 2024. 2

"SmokeSVD: Smoke Reconstruction from A Single View via Progressive Novel View Synthesis and Refinement with Diffusion Models" Supplemental Document

Chen Li[1†] Shanshan Dong[2†] Sheng Qiu[2*] Jianmin Han[2]
Yibo Zhao[1] Zan Gao[1] Taku Komura[3] Kemeng Huang[3]
[1] Tianjin University of Technology [2] Zhejiang Normal University [3] University of Hong Kong

A. Overview

In this supplementary material, we provide additional background, detailed descriptions of the technical approach, implementation specifics, evaluation results, and ablation studies. We also discuss the limitations of our work and outline potential directions for future research.

B. Preliminary

Navier-Stokes Equation. Generally, fluid motion is governed by the well-known incompressible Navier-Stokes equations:

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} = -\frac{\nabla p}{\rho} + \nu\nabla^2\mathbf{u} + \mathbf{f}, \tag{1}$$

$$\nabla\cdot\mathbf{u} = 0, \tag{2}$$

where $\mathbf{u}$ is the velocity, ρ is the density, p is the pressure, $\mathbf{f}$ is the external force, and ν is the viscosity coefficient, which is usually set to zero for smoke phenomena. Eq. 1 is the momentum equation, which describes the time rate of velocity change, while Eq. 2 is the mass conservation equation to preserve the incompressibility. To formalize, density evolution follows the transport equation:

$$\frac{\partial \rho}{\partial t} + \mathbf{u}\cdot\nabla\rho = 0. \tag{3}$$

Diffusion Models. Diffusion probabilistic models (DDPM) consist of two processes: a forward diffusion process and a reverse inference process. During the training stage, given a data point $x_0 \sim q(x)$ sampled from the real data distribution, the forward process adds Gaussian noise to the sample x_0 over S time steps, constructing a Markov chain diffusion process:

$$q(x_s|x_{s-1}) = \mathcal{N}(x_s; \sqrt{1-\beta_s}x_{s-1}, \beta_s I), \tag{4}$$

$$q(x_{1:S}|x_0) = \prod_{s=1}^{S} q(x_s|x_{s-1}), \tag{5}$$

where $\mathcal{N}$ denotes a Gaussian distribution, β_s denotes a fixed or learnable variance schedule parameter that controls the noise intensity added at each step, x_s denotes the noisy image at time step s (selected from the total steps S), which can be expressed as:

$$x_s = \sqrt{\bar{\alpha}_s}x_0 + \sqrt{1-\bar{\alpha}_s}\epsilon, \tag{6}$$

where $\alpha_s = 1-\beta_s$, $\bar{\alpha}_s := \prod_{s=1}^{s}\alpha_s$, and $\epsilon \sim \mathcal{N}(0, I)$. The model is trained to minimize the following loss function:

$$\|\epsilon - \epsilon_\theta(x_s, s)\|^2. \tag{7}$$

During the generation stage, the diffusion model samples a Gaussian random noise $x_S \sim \mathcal{N}(0, I)$, and utilizes the predefined variance σ_s and random noise ϵ_s to gradually denoise it to until x_0. This process is formulated as:

$$\begin{aligned} x_{s-1} = \sqrt{\bar{\alpha}_{s-1}}\Big(\frac{x_s - \sqrt{1-\bar{\alpha}_s}\epsilon_\theta^{(s)}(x_s)}{\bar{\alpha}_t}\Big) \\ + \sqrt{1-\bar{\alpha}_{s-1}-\sigma_s^2}\cdot\epsilon_\theta^{(s)} + \sigma_s\epsilon_s, \end{aligned} \tag{8}$$

where $s = S, ..., 1$, and ϵ_θ is estimated noise from x_s.

C. Technical Details

C.1 Mathematical Symbols

Key mathematical symbols used in the paper are documented in Table 1.

C.2 Multi-frame Training Algorithm

If the previously synthesized frame is not used as one of the input conditions, the generated results exhibit significant cumulative errors, as shown in Fig. 1. To address this issue, we propose a multi-frame training algorithm, summarized in Alg. 1, which incorporates the estimated clean image from the previous time step as a conditional input for the subsequent forward diffusion process.

*Corresponding Author. †Equal contributions.

Table 1. Key Mathematical Symbols

Symbol	Meaning
w_{α}^{t}	The smoke image at the tth frame and α viewing angle
$w_{c,\alpha}^{t}$	The clean image
$w_{r,\alpha}^{t}$	The rendered result for reconstructed density field
$w_{f,\alpha}^{t}$	The refined image
α	$\alpha = \angle 0^{\circ}$ for the input front view, $\alpha = \angle 90^{\circ}$ for the side view
I^{t}	The set of images from multiple views at the tth frame
ρ	Density field
$\hat{\rho}$	Advected density field
$\rho_{r,c}$	Coarse-grained reconstructed density field
$\rho_{r,f}$	Fine-grained reconstructed density field
$\mathbf{u}$	Velocity field
$\mathbf{u_r}$	Reconstructed velocity field
ρ_{in}	Inflow state
$\mathcal{A}$	Differentiable advection operator
$\mathcal{R}$	Differentiable rendering operator
SvDiff	Side-view synthesizer based on diffusion models
NvRef	Novel refinement module
$\mathcal{G}_{\rho}^{c}$	Coarse-grained density generator
$\mathcal{G}_{\rho}^{f}$	Fine-grained density generator
$\mathcal{G}_{u}$	Velocity generator

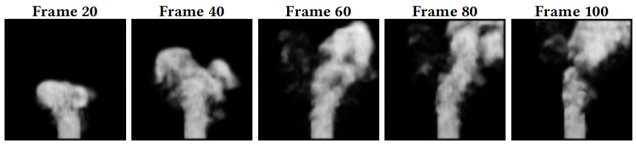

Figure 1. Side-view generation results affected by cumulative error.

C.3 Progressive Refinement

As shown in Fig. 3, $\rho_{r,c}^{t}$ appears blurry in novel views due to limited available information. To address this, we introduce a progressive refinement module that incrementally enhances the blurred novel images, improving clarity from near to far views, as summarized in Alg. 2.

Algorithm 1 Multi-frame Training Algorithm for SvDiff.

Require: Number of iterations it, noise steps S, noise threshold TQ

repeat
  Sample $s \sim \text{Uniform}(\{1, \dots, S\})$
  $\rho^{t-1} = \mathcal{G}_{\rho}(w_{\angle 0^{\circ}}^{t-1}, w_{\angle 90^{\circ}}^{t-1})$
  for $i = 0, 1, 2, \dots, it$ **do**
    Condition c^{i} : $w_{c,\angle 90^{\circ}}^{i+t-2}$, $w_{r,\angle 90^{\circ}}^{i+t-2}$, $w_{c,\angle 90^{\circ}}^{i+t-1}$, $w_{r,\angle 90^{\circ}}^{i+t-1}$, $w_{\angle 0^{\circ}}^{i+t}$
    Clean image sample $x_0^i : w_{\angle 90^{\circ}}^{i+t}$
    Sample $\epsilon \sim \mathcal{N}(0, I)$
    $x_s^i = \sqrt{\bar{\alpha}_s} x_0^i + \sqrt{1-\bar{\alpha}_s}\epsilon$
    $\hat{\epsilon} = \epsilon_{\theta}(x_s^i, c^i, s)$
    $\mathcal{L}_{noise} = \|\epsilon - \hat{\epsilon}\|^2$
    if $s < TQ$ **then**
      $\hat{x}_0^i = \dfrac{x_s^i - \sqrt{1-\bar{\alpha}_s}\hat{\epsilon}}{\sqrt{\bar{\alpha}_s}}$
      $\rho_{r,c}^i = \mathcal{G}_{\rho}(w_{\angle 0^{\circ}}^{i+t}, w_{c,\angle 90^{\circ}}^{i+t})$
      $\mathbf{u}^{i-1} = \mathcal{G}_u(\rho^{i-1}, \rho_{r,c}^i)$
      $w_{c,\angle 90^{\circ}}^{i+t} = \hat{x}_0^i, w_{r,\angle 90^{\circ}}^{i+t} = \mathcal{R}(\rho_{r,c}^i), \rho^{i-1} = \rho_{r,c}^i$
      $\mathcal{L}_{img} = \|x_0^i - \hat{x}_0^i\|^2$
      $\mathcal{L}_{vel} = \|\nabla \cdot \mathbf{u}^{i-1}\|^2 + \|\nabla \mathbf{u}^{i-1}\|^2$
      $\mathcal{L}_{sp} = \|H(w_{c,\angle 90^{\circ}}^{i+t}) - H(w_{\angle 0^{\circ}}^{i+t})\|^2$
    else
      break
    end if
  end for
  Take gradient step on $\mathcal{L}_{SvDiff}$
until converged

C.4 Density Generator

To provide 3D input from 2D images, we transform the image through expansion to match the required dimensions, and concatenate them from multiple viewpoints, as shown in Fig. 4. To be specific, $\mathcal{G}_{\rho}$ adopts the UNet3+ architecture with 3D convolutions.

C.5 Velocity Estimation

To reconstruct temporal and physically reasonable smoke dynamics, we establish a velocity generator $\mathcal{G}_u$ to estimate the velocity field based on two density fields of consecutive frames:

$$\mathbf{u}_r^t = \mathcal{G}_u(\rho^t, \rho^{t+1}), \tag{9}$$

which is supervised by $\mathcal{L}_u = \|\mathbf{u_r} - \mathbf{u}\|^2$. Additionally, to satisfy the divergence-free requirement in Eq. 2, we introduce another divergence loss as $\mathcal{L}_{div} = \|\nabla \cdot \mathbf{u_r} - \nabla \cdot \mathbf{u}\|^2$.

To ensure long-term robustness and reduce the adverse impact of the reconstruction errors in density, we employ

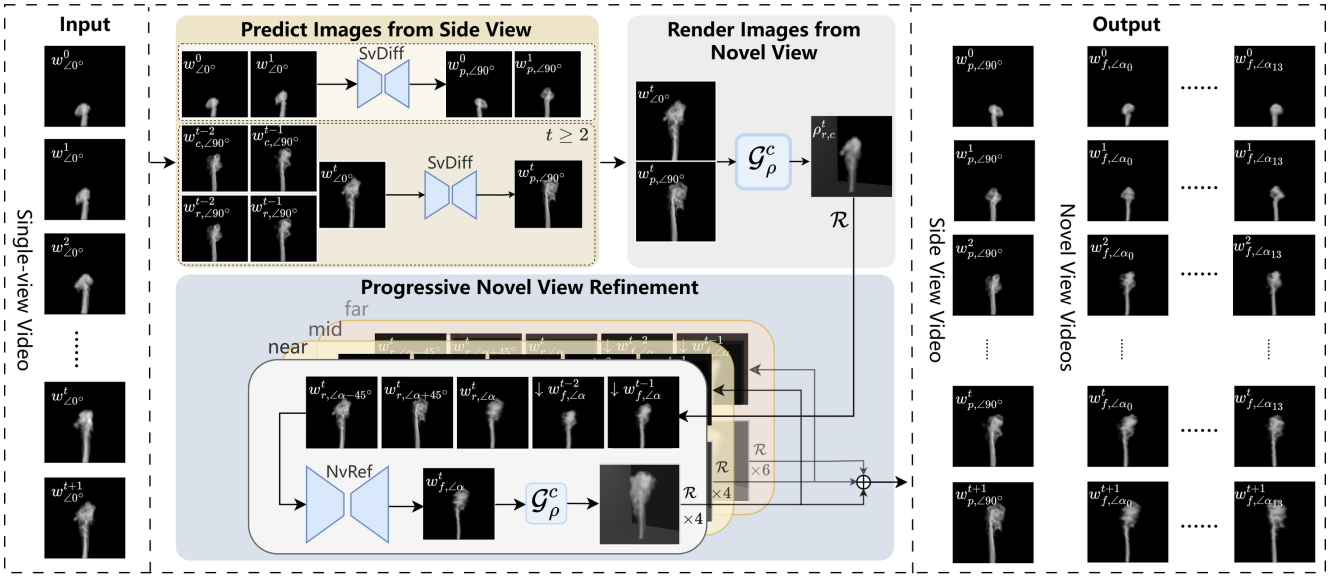

Figure 2. Procedure for side-view synthesis and novel view refinement. First, SvDiff predicts side-view images from input and previously generated images (when $t \geq 2$). Next, we reconstruct coarse density with $\mathcal{G}_\rho^c$ using front and side views, and render nearby novel views. Then, we iteratively refine novel views and reconstruct density, progressively extending from near to mid and far views, yielding multiple high-quality views for fine-grained reconstruction.

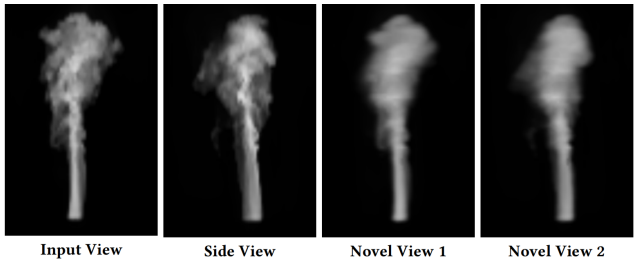

Figure 3. Rendering results of coarse-grained density field, which exhibits blurriness in novel views.

a differentiable advection operator $\mathcal{A}$ based on Eq. 3, to formulate an advection loss term for the velocity generator. The advection operator $\mathcal{A}$ transports the density field ρ based on the velocity field $\mathbf{u}$, expressed as:

$$\hat{\rho}^t = \mathcal{A}(\rho^{t-1}, \mathbf{u}_r^{t-1}, \rho_{in}, dt), \tag{10}$$

where the density field obtained through velocity-based advection is called the advected density field, denoted as $\hat{\rho}$, ρ_{in} is the dynamic inflow, and dt is the time step. Similar to the density generator, we employ the following 3D density-based and 2D image-based advection loss terms:

$$\mathcal{L}_{advect} = \lambda_{\hat{\rho}} \|\rho - \hat{\rho}\|^2 + \lambda_{\mathcal{R}} \|\mathcal{R}(\rho) - \mathcal{R}(\hat{\rho})\|^2. \tag{11}$$

Based on the advected density field $\hat{\rho}$, we modify the input of $\mathcal{G}_u$ to ensure that the velocity field can be corrected through the advected density field, with the formula being:

$$\mathbf{u}_r^t = \mathcal{G}_u(\hat{\rho}^t, \rho^{t+1}). \tag{12}$$

Algorithm 2 Progressive Novel View Refinement.

Require: Current frame t; coarse density ρ_c^t; near/mid/far view sets nv, mv, fv; angular offset β; refined images from previous frames w_f^{t-1}, w_f^{t-2}

ViewSets $\leftarrow \{nv,\ mv,\ fv\}$
for each view set V in *ViewSets* **do**
  # Rendering and refinement for the same view type
  for each view angle α in V **do**
    $w_{r,\alpha}^t = \mathcal{R}(\rho_c^t,\ \alpha)$
    $w_{r,\alpha-\beta}^t = \mathcal{R}(\rho_c^t,\ \alpha - \beta)$
    $w_{r,\alpha+\beta}^t = \mathcal{R}(\rho_c^t,\ \alpha + \beta)$
    $w_{f,\alpha}^t = \text{NvRef}\Big(w_{r,\alpha-\beta}^t \oplus w_{r,\alpha+\beta}^t \oplus w_{r,\alpha}^t$
      $\oplus \downarrow w_{f,\alpha}^{t-1} \oplus \downarrow w_{f,\alpha}^{t-2}\Big)$
  end for
  # Density reconstruction using all refined images obtained
  $\rho_c^t = \mathcal{G}_\rho(\text{all refined imgs})$
end for
 # After the final iteration
$\rho_f^t \leftarrow \rho_c^t$

C.6 Inflow Estimation

The inflow state has a tremendous impact on the visual pattern of smoke phenomena, which cannot be ignored in smoke reconstruction. In long-term evolution, underestimating the inflow will lead to an inability to fill the smoke

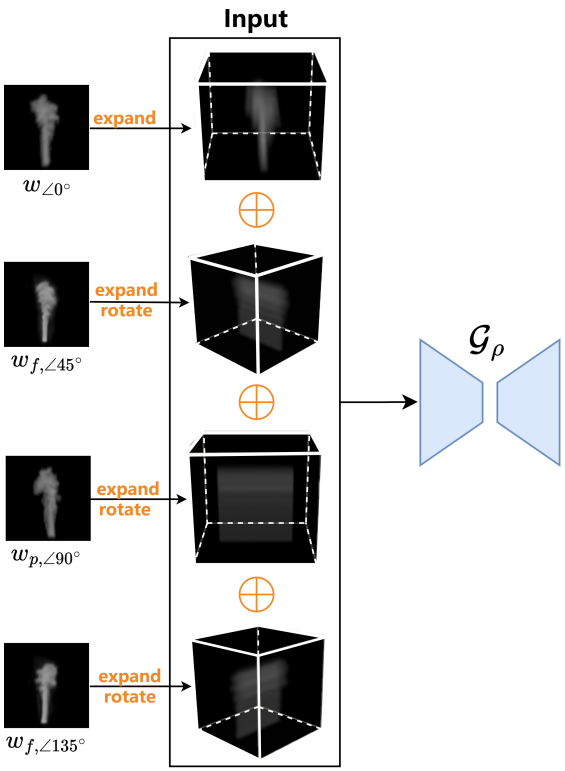

Figure 4. The architecture of density generator. The illustration depicts the case with four input images.

volume in later time steps, while overestimating can cause obvious instability, ultimately failing to match the input images [2].

To address this issue, we propose to estimate the inflow state frame-by-frame, that determines the inflow of current frame based on two adjacent density fields $\hat{\rho}^t$ and ρ^{t+1}, the velocity field $\mathbf{u}^t$, and the input image $w^{t+1}_{\angle 0^\circ}$. Specifically, for each frame, we initialize a random smoke source ρ_{in} and iteratively optimize the inflow source by minimizing the following loss function:

$$\begin{aligned}\mathcal{L}_s = &\|\rho_r^{t+1} - \mathcal{A}(\hat{\rho}_r^t, \mathbf{u}_r^t, \rho_{in}^t, dt)\|^2 \\ &+ \|w^{t+1}_{\angle 0^\circ} - \mathcal{R}(\mathcal{A}(\hat{\rho}_r^t, \mathbf{u}_r^t, \rho_{in}^t, dt), \angle 0^\circ)\|^2 \\ &+ \|\rho_{in}^{t-1} - \rho_{in}^t\|^2.\end{aligned} \tag{13}$$

Additionally, to prevent overestimation of the inflow source, we enforce zeroing out portions of the source that exceed a height threshold.

By incorporating the velocity and inflow estimation with density evolution [7], we can impose strong physical constraints to augment the temporal coherence and visual realism of SmokeSVD, thus effectively removing long-term flickers and non-physical artifacts in reconstructed smoke dynamics.

D. Implementation Details and Experimental Settings

Implementation Details. Our method is trained in two stages. In the first stage, we train SvDiff and NvRef based on the multi-frame training scheme to estimate clean images. We employ DDIM (Denoising Diffusion Implicit Models) sampling [8] described in Eq. 8 to accelerate the sampling process. Simultaneously, we also train the density generator $\mathcal{G}_\rho$ and the velocity generator $\mathcal{G}_u$. Our density generator $\mathcal{G}_\rho$ outputs smoke density fields with resolutions of 64^3 (for synthetic datasets) or $64 \times 112 \times 64$ (for real-world datasets). In the second stage, we fine-tune the velocity generator $\mathcal{G}_u$ based on the pre-trained density generator $\mathcal{G}_\rho$. All the aforementioned experiments were conducted on an NVIDIA GeForce RTX 3090 (24GB) GPU, while the performance was tested on an NVIDIA GeForce RTX 2080 Ti (11GB) GPU. Since optimization-based and neural radiance field (NeRF) methods require training for a few hours, far exceeding the minute-level time consumption of our proposed method, their specific time cost is not listed in the table.

Dataset. Based on the Eulerian method [6], we generated the required synthetic dataset by randomly modifying the wind fields, thermal fields, and the size and position of inflow regions in the scenarios. A total of 100 scenarios were generated, with each scene containing 150 frames. Additionally, we used post-processed images from the first 20 scenes of the ScalarFlow dataset [2] to train and evaluate our model.

Benchmarks. We compared our method with existing techniques that accept single-view videos as input for 3D smoke reconstruction, selecting GlobTrans [3], NGT [4], PICT [9], and PINF [1] as benchmarks. In our experiments, we modified the inputs of PICT and PINF to support single-view video input. Among these methods, GlobTrans reconstructs 3D smoke based on direct optimization algorithms, while PICT and PINF are based on Neural Radiance Fields (NeRF). These methods all require optimization for individual scenario, resulting in expensive time consumption and re-optimization requirement when changing scenarios. In contrast, the NGT method uses a trained neural network to estimate a single motion of smoke, avoiding direct optimization of the entire scenario, thereby significantly improving reconstruction speed and applicability.

Evaluation Metric. For image-related tasks (including novel view generation, refinement, and rendered images from reconstructed density fields), we use Mean Square Error (MSE), Root Mean Square Error (RMSE), Peak

Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [10], Fréchet Inception Distance (FID) [5], Learned Perceptual Image Patch Similarity (LPIPS) [11], and STYLE similarity to measure the similarity between generated images and ground truth images. The STYLE similarity is defined as the $L1$ difference between the Gram matrices of features extracted from the generated results and the ground truth using VGG19. Additionally, we evaluate the feature consistency between generated images and ground truth images with $\mathcal{L}_{sp}$. For reconstruction tasks, we use RMSE of density fields, divergence and gradient of velocity fields to measure the similarity between reconstructed and ground truth physical fields.

E. More Evaluations

Results on Synthetic Dataset. Fig. 5 demonstrates the qualitative performance of our method on the synthetic dataset, where the density field resolution of the reconstructed scenario is 64^3. By generating novel view images, our method significantly alleviates the ill-posed problem in single-view video based reconstruction, and the rendering results of reconstructed density fields perform well across different views.

Side-View Quality. We employ optical flow analysis as temporal consistency metrics (Table 2). We achieve performance closest to GT (15min vs. GT's 30 hours): Max (2nd best) indicates minimal flickers, Avg shows reasonable dynamics comparable to NGT/GT, and Std validates consistency. Note that PICT's low metrics stem from depth-blur eliminating motion detail.

Table 2. Optical flow statistics over 120 frames on ScalarFlow.

Metric	Reference	GT	NGT	PINF	PICT	FluidNexus	Ours
Max.	0.0896	**0.0953**	0.1272	0.1861	0.5890	0.2166	<u>0.1208</u>
Avg.	0.0593	<u>0.0639</u>	**0.0630**	0.1274	0.0253	0.0765	0.0767
Std Dev	0.0091	<u>0.0121</u>	0.0170	0.0185	**0.0116**	0.0348	0.0158

More Generalization Performance. We also test with multi-plume collisions and dry ice, as shown in Figs. 6 and 7. Our method performs well on various smoke shapes, which are fundamentally different from the single-source smoke scenes in our training dataset.

Interactive Simulation. Our reconstructed physical fields enable the re-simulation of input videos, and the generation of new smoke phenomena with controllable effects and enhanced detail, as shown in Figs. 8 and 9. In Fig. 9, we demonstrate re-simulation results in which a newly added spherical obstacle (top row) or external force field (bottom) is introduced by projecting the reconstructed velocity field onto a new simulation domain.

Compatibility with 3D Gaussian Splatting. Once sufficient novel views have been generated, our method can be seamlessly integrated with downstream applications such as 3D Gaussian Splatting (3DGS). As shown in Figs. 10 and 11, thanks to the multi-view consistency and well-structured spatiotemporal features provided by our approach, 3DGS is able to reproduce physically and visually plausible smoke sequences without the need for additional temporal processing.

F. More Ablation Studies

Effect of Frame Numbers. We adopted a multi-frame training strategy to train the side-view synthesizer (SvDiff) and the novel view refinement module (NvRef). Taking SvDiff as an example, in the early stages of training, We fed SvDiff one image for a single forward diffusion process; subsequently, we gradually increased the number of training frames and forward diffusion times until the synthesis quality met the expectation. To determine the final number of training frames and forward diffusion timesteps, we tested different hyperparameter settings for SvDiff. Since the number of training frames equals the number of forward diffusion times, we named these hyperparameter settings based on the number of frames (e.g., SvDiff-F1, SvDiff-F2), as shown in Fig. 12. As the number of training frames increased, the synthetic results gradually became more reasonable. For example, the SvDiff-F1 in Fig. 12 did not use the multi-frame information to estimate clean images, so due to the cumulative error, subsequent synthetic frames gradually deviated from reasonable smoke appearance. According to the results in Table 3, we found that the SvDiff based on four forward diffusions (SvDiff-F4) achieves the best. Both qualitative and quantitative evaluations indicate that the multi-frame training strategy based on estimated clean images plays a crucial role in the long-term generation process of diffusion models.

Table 3. Quantitative comparison of SvDiff with different frame numbers on the synthetic dataset. We report $\mathcal{L}_{sp}$, LPIPS, and SSIM to measure the differences between synthetic images and reference images, and warp error to measure pixel-level distortion between consecutive frames based on mean squared error (MSE).

Algorithm	$\mathcal{L}_{sp}\downarrow$	Warp Error$\downarrow$	LPIPS$\downarrow$	SSIM$\uparrow$
reference	/	0.0981	/	/
SvDiff-F1	1.2601	0.2003	0.3873	0.4364
SvDiff-F2	1.2673	0.1819	0.3742	<u>0.5077</u>
SvDiff-F3	1.0422	**0.0915**	0.3910	0.4997
SvDiff-F4	**0.3475**	0.1481	**0.3384**	**0.5729**
SvDiff-F5	<u>0.7081</u>	<u>0.1259</u>	<u>0.3779</u>	0.5052

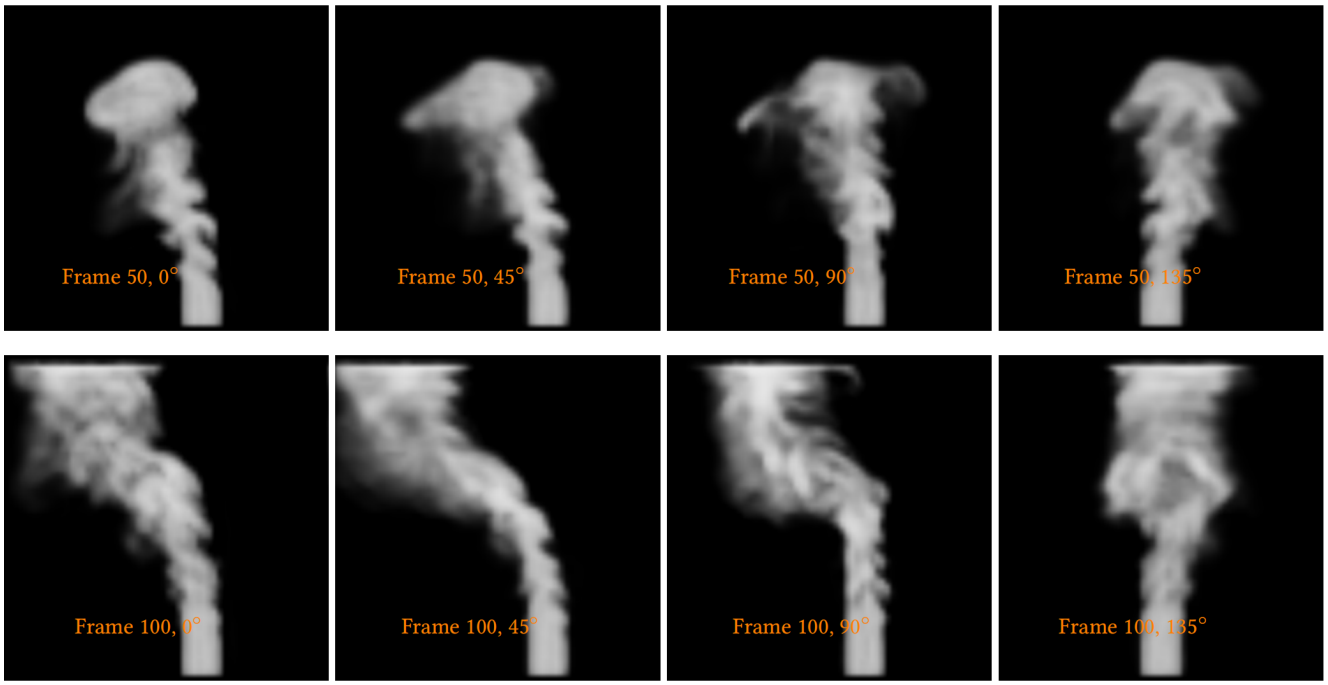

Figure 5. The rendering results of reconstructed density field at multiple views based on our proposed method.

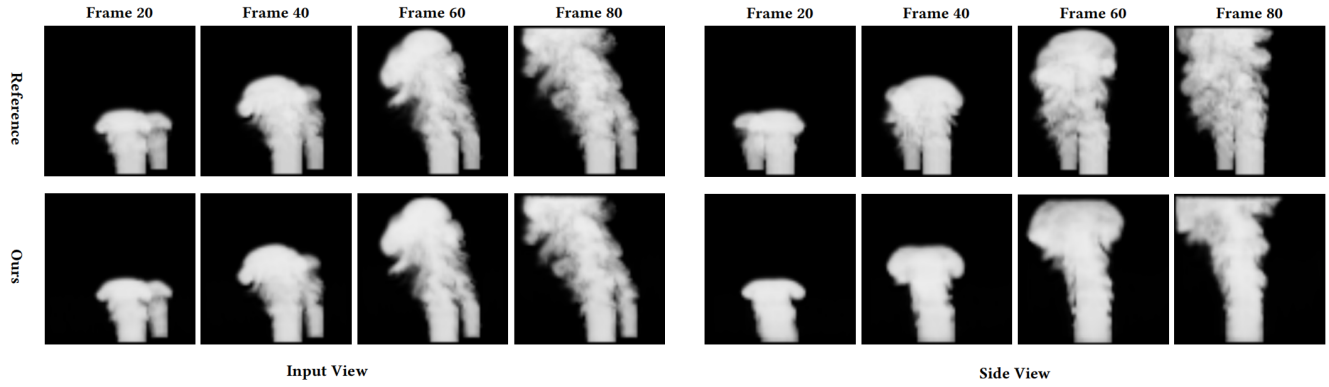

Figure 6. Reconstruction result for a multi-plume scenario, shown from both input and side views.

Effect of View Numbers. Our density generator can accept up to 16 smoke images from different viewpoints, with these views evenly distributed along a 180° arc. To determine the optimal number of input views for fine-grained density reconstruction, we trained several density generators using 2, 4, 8, and 16 input images (denoted as $2-, 4-, 8-, 16-\mathcal{G}\rho$), and evaluated their performance. The quantitative results are presented in Table 4. In the experiment, when the number of input images was less than 16, images from other novel views were masked. All image metrics were evaluated based on 16 real viewpoints, and the quantitative analysis indicates that as the number of input views increases, the reconstruction quality gradually improves. Therefore, in the coarse-grained density reconstruction stage, we used only a subset of views as input, whereas in the fine-grained stage, all 16 input views were utilized to provide richer information for high-quality reconstruction.

Table 4. Quantitative evaluation of density generators with different numbers of input views on the synthetic dataset. The last five metrics are evaluated based on images from 16 views.

View Num	ρ RMSE↓	RMSE ↓	SSIM↑	PSNR↑	LPIPS↓	FID↓
2	0.0356	0.0206	0.9795	37.0561	0.0417	31.0919
4	0.0256	0.0100	0.9915	43.1682	0.0205	9.7665
8	<u>0.0186</u>	<u>0.0058</u>	<u>0.9960</u>	<u>47.2533</u>	<u>0.0099</u>	<u>2.5882</u>
16	**0.0148**	**0.0043**	**0.9974**	**49.6970**	**0.0050**	**1.3745**

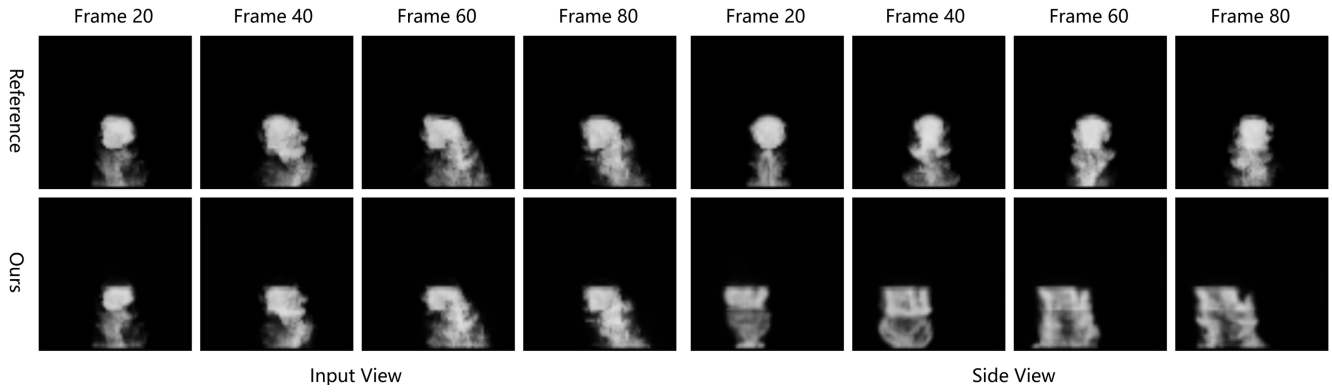

Figure 7. Reconstruction result for a dry ice scenario.

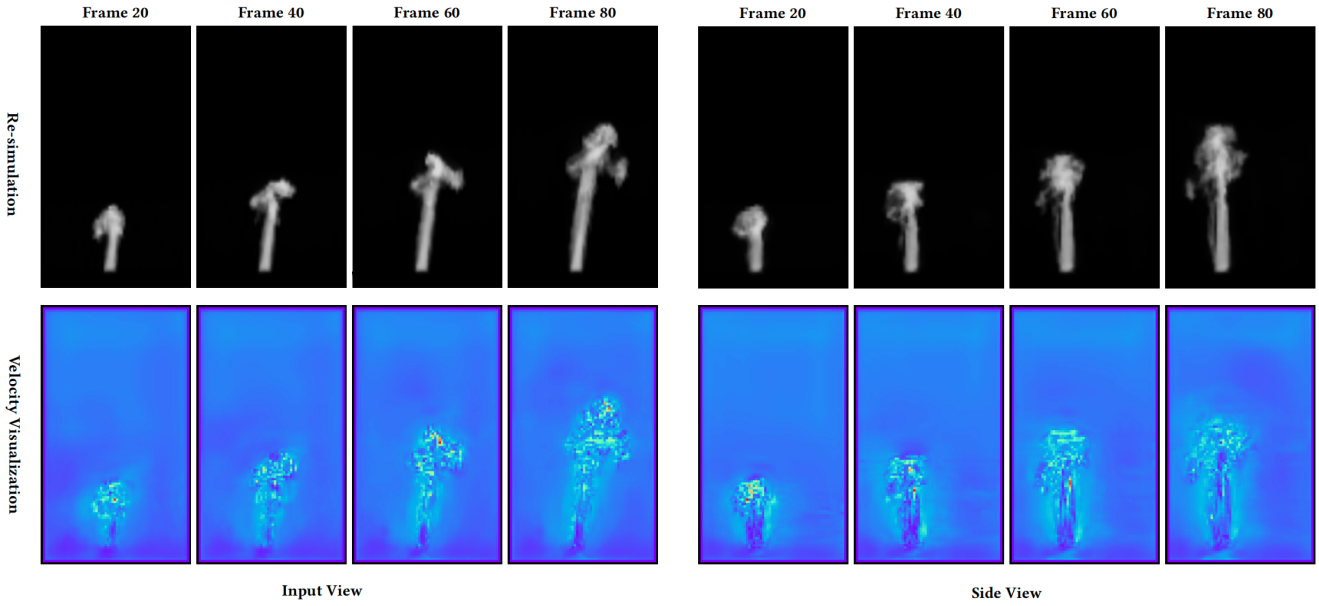

Figure 8. The rendered re-simulation results and velocity estimation visualization at the input view and the side view.

Ablation on Side-view Synthesizer. We also visualized the maximum values and gradient of reconstructed velocity fields in Figs. 14 and 15.

Ablation on Key Components. Figs. 16 and 17 show NGT combined with our refinement and reconstruction. Our approach is compatible with NGT and further enhances its results, achieving high-quality reconstruction.

G. Limitation and Discussion

While our proposed framework demonstrates strong performance in reconstructing dynamic smoke from single-view input, several limitations remain. First, the current method assumes a relatively clean background and consistent lighting conditions; in real-world scenarios with complex backgrounds or varying illumination, the quality of side-view synthesis and subsequent reconstruction may degrade. Second, although our progressive refinement strategy improves multi-view consistency, the approach still relies on the accuracy of the initial side-view synthesis, significant errors in early stages can propagate and affect the final results. Third, our model is primarily evaluated on synthetic and controlled real-world datasets; its generalization to highly diverse or outdoor smoke phenomena remains to be further validated. Additionally, the computational cost, while lower than optimization-based methods, can still be significant when scaling to higher resolutions or longer sequences. Finally, our framework currently focuses on grayscale smoke and does not explicitly handle colored smoke, solid obstacles, or interactions with complex environments. Future work could address these limitations by incorporating more

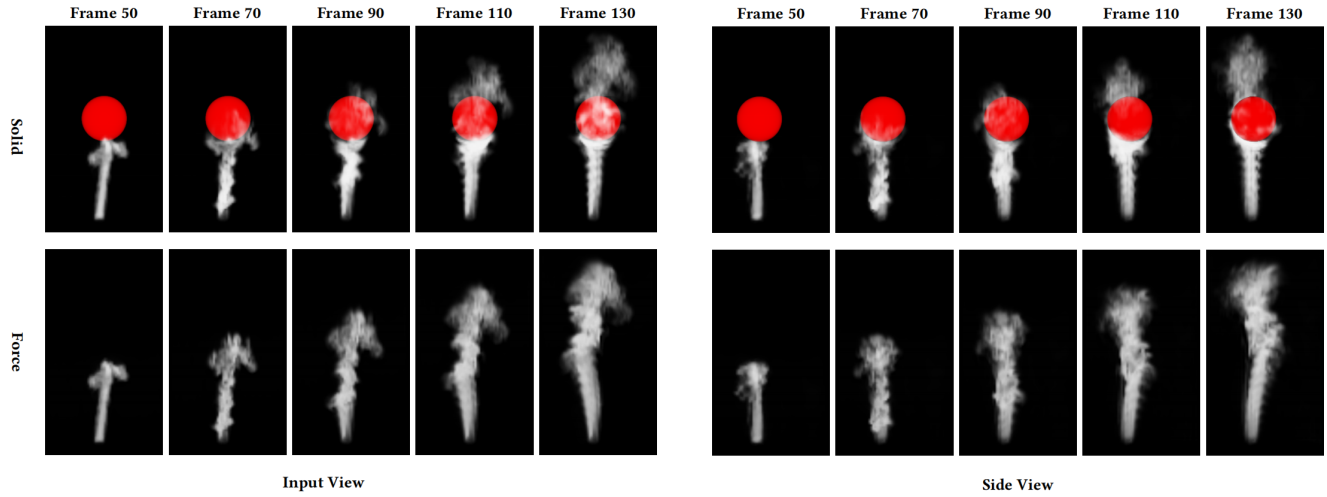

Figure 9. The re-simulation result with added fluid-solid coupling (top row), where we place a sphere obstacle (the red circle) at the 50th time step, and external force field (bottom row).

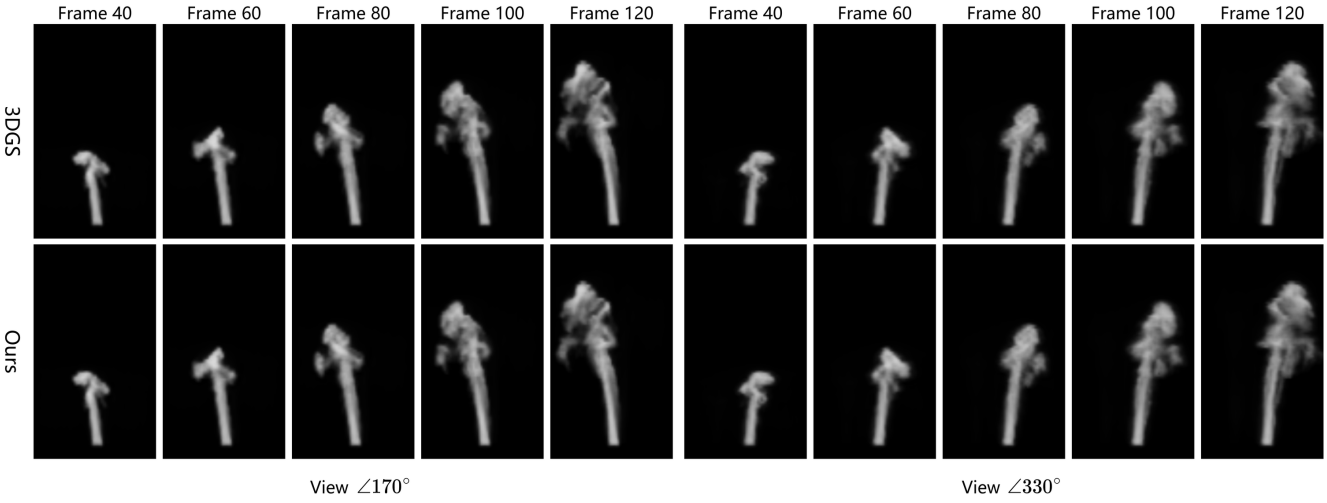

Figure 10. 3DGS results (top) based on our synthesized novel views (bottom).

robust background modeling, exploring domain adaptation techniques, extending the framework to handle color and multi-phase flows, and integrating more advanced physical constraints to further enhance realism and generalization.

References

[1] 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 Transactions on Graphics (ToG)*, 41(4):1–14, 2022. 4

[2] 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 Transactions on Graphics (TOG)*, 38(6):1–16, 2019. 4

[3] 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. 4

[4] Erik Franz, Barbara Solenthaler, and Nils Thuerey. Learning to estimate single-view volumetric flow motions without 3d supervision. *arXiv preprint arXiv:2302.14470*, 2023. 4

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[6] Theodore Kim, Nils Thürey, Doug James, and Markus

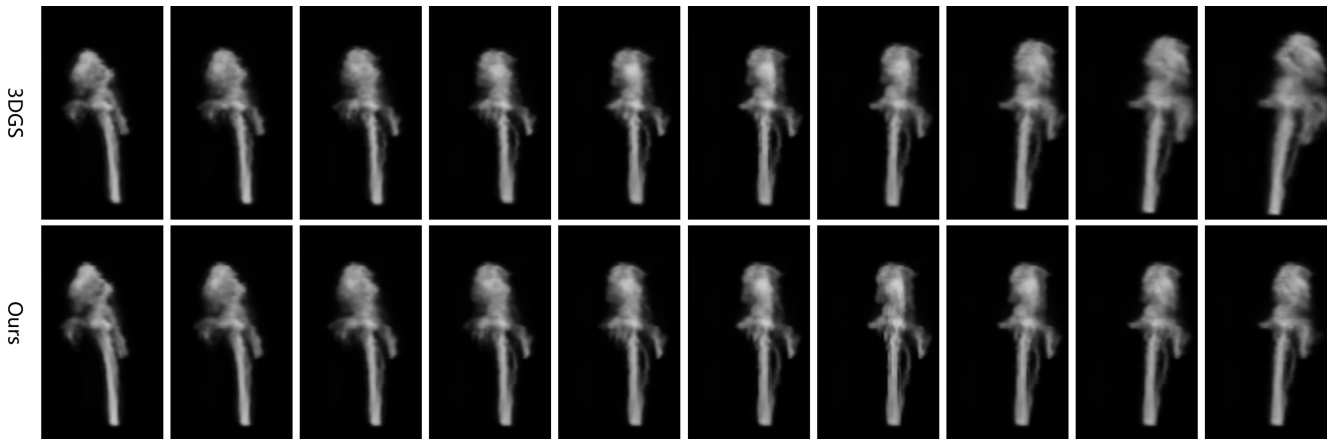

Figure 11. 3DGS results (top) and our reconstruction result (bottom) under rotating views from $\angle 210^\circ$ to $\angle 300^\circ$.

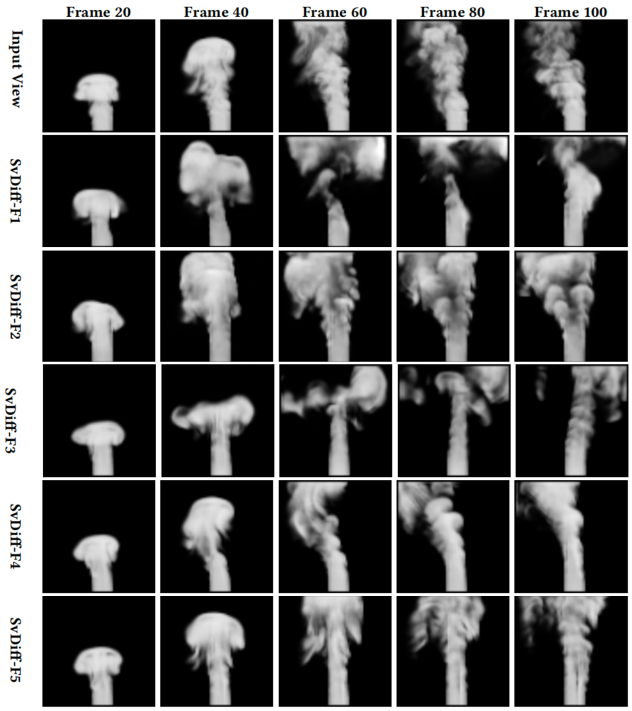

Figure 12. Qualitative comparison of side view synthesis with different frame numbers on the synthetic dataset.

Gross. Wavelet turbulence for fluid simulation. *ACM Transactions on Graphics (TOG)*, 27(3):1–6, 2008. 4

[7] Sheng Qiu, Chen Li, Changbo Wang, and Hong Qin. A rapid, end-to-end, generative model for gaseous phenomena from limited views. *Computer Graphics Forum*, 40(6):242–257, 2021. 4

[8] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4

[9] 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, New York, NY, USA, 2024. Association for Computing Machinery. 4

[10] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[11] Richard Zhang, Phillip Isola, Alexei A Efros, and et al. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 5

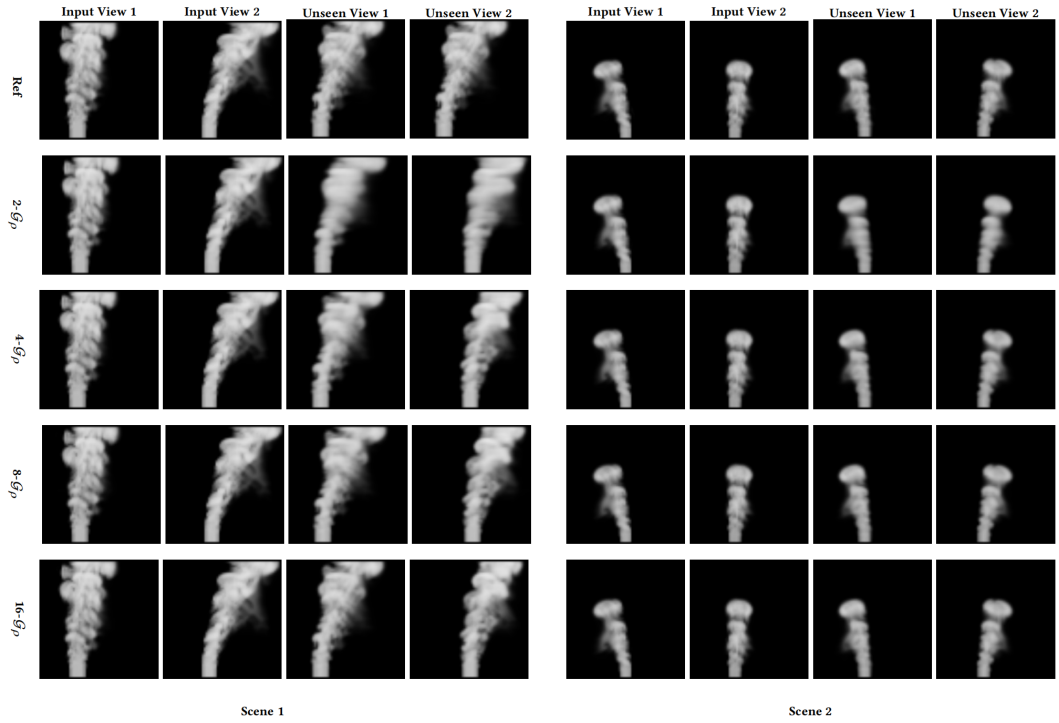

Figure 13. Qualitative comparison of density generators with different numbers of views on the synthetic dataset.

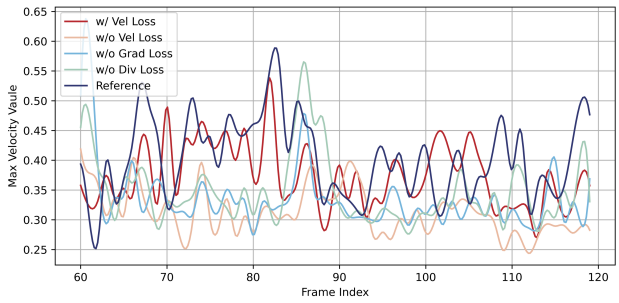

Figure 14. Comparison of the maximum values of reconstructed velocity fields by SvDiff with different loss functions at various time steps.

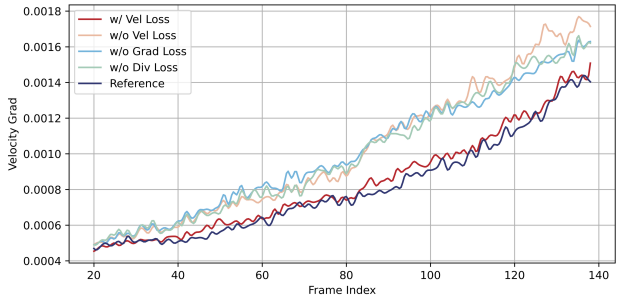

Figure 15. Comparison of the gradient of reconstructed velocity fields by SvDiff with different loss functions at various time steps.

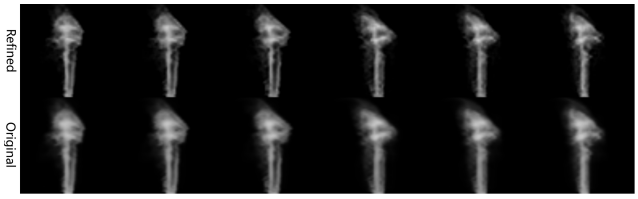

Figure 16. NGT combined with our refinement model.

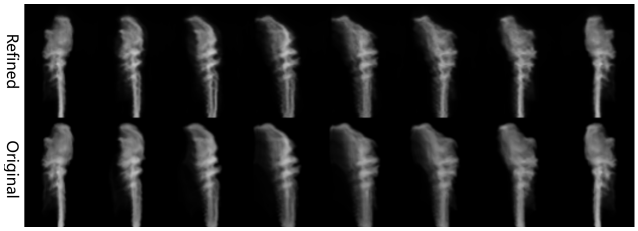

Figure 17. NGT combined with our reconstruction model.