NS-Diff: Fluid Navier–Stokes Guided Video Diffusion via Reinforcement Learning

1. Detailed Hyperparameter Sensitivity Analysis

The hyperparameters of the NS-Diff framework are grounded in clear physical interpretations and are not the result of arbitrary tuning. As detailed in the main manuscript, K represents the standard temporal smoothing window, β governs the rigidity probability threshold, λ_1 balances the rigid prior, and α scales the Reinforcement Learning (RL) activation based on the diffusion noise schedule. To demonstrate the robustness of our method, we conducted extensive sensitivity experiments. The results, summarized in Table 1, show that the model performance remains stable across a wide range of values. This stability arises because our physical rewards act as directional guides rather than rigid constraints, allowing the diffusion model's strong generative priors to self-correct within plausible physical scales. Crucially, a single set of hyperparameters was utilized across three different datasets, which further underscores the generalization capability of our approach across diverse physical scenarios.

Table 1. Hyperparameter sensitivity analysis. The colors indicate different variations from the base configuration.

Hyperparameter	Var. 1 / Base / Var. 2	FVD ↓	ΔJ ↓	$\mathcal{L}_{div}$ (10^{-3}) ↓
Window Size K	1 / 2 / 3	179 / 183 / 187	0.39 / 0.33 / 0.34	3.2 / 2.9 / 2.9
RL Rigid Weight λ_1	0.5 / 1.0 / 2.0	184 / 183 / 190	0.34 / 0.33 / 0.30	2.7 / 2.9 / 3.3
Rigidity Threshold β	0.4 / 0.5 / 0.6	183 / 183 / 185	0.36 / 0.33 / 0.37	2.9 / 2.9 / 2.8

2. Qualitative Ablation and Semantic Stability

To further illustrate the necessity of each component in NS-Diff, we provide an extended qualitative ablation study. Without the RL framework, physical interactions often appear inert or unrealistic; for instance, a rigid object like a mango might land flatly on a granular surface without displacing it. The removal of condition injection or physics guidance terms frequently leads to semantic collapse, where the model hallucinates non-physical events, such as a "slicing" effect, to resolve the interaction. Specific prior terms also play a critical role: omitting the fluid prior L_{fluid} causes granular materials to behave like static solids, while the absence of the rigid prior L_{rigid} results in unnatural deformations of solid objects upon impact. These results confirm that our physics-guided RL framework is essential for maintaining both physical plausibility and semantic integrity.

3. Flow Estimator Benchmarking on Decoded Proxies

A key design choice in our framework is the use of the ARFlow estimator for motion analysis. While state-of-the-art (SOTA) models like RAFT and GMFlow are highly optimized for high-resolution, clean video data, they exhibit reduced robustness when applied to the noisy, low-resolution "decoded proxies" generated during the diffusion process. Our evaluation, presented in Table 2, demonstrates that ARFlow achieves a superior balance of accuracy and efficiency for this specific task. Compared to RAFT and GMFlow, ARFlow yields lower FVD and more accurate physical metrics (ΔJ, $\mathcal{L}_{div}$) with a significantly lower computational overhead of only 8%. This justifies our choice of a well-adapted lighter model over more complex alternatives for this noisy proxy domain.

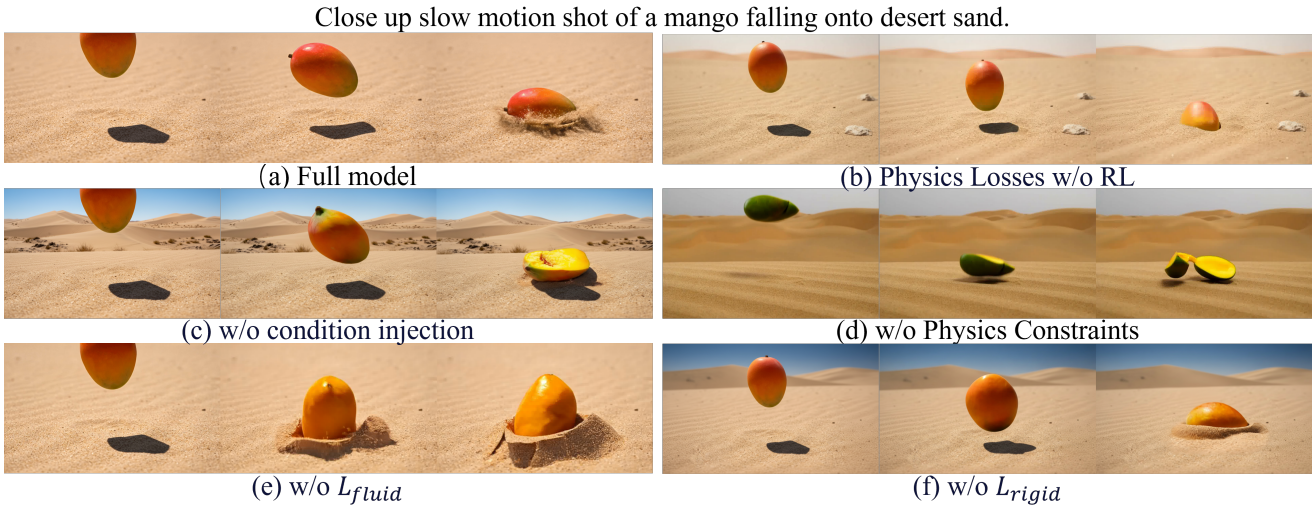

Figure 1. Qualitative ablation results.

Table 2. Comparison of motion estimators on decoded proxies.

Flow Estimator	**FVD** ↓	ΔJ ↓	$(L_{div} \times 10^{-3})$ ↓	**Overhead**
RAFT	191.0	0.38	3.20	+15%
GMFlow	188.4	0.36	3.09	+25%
ARFlow	**183.3**	**0.33**	**2.90**	**+8%**

4. PPO Training Stability and Seed Variance

To ensure the reliability of the Reinforcement Learning process, we conducted a stability analysis using five random seeds. The results indicate low variance across all key metrics, including FVD (183.3 ± 2.5), ΔJ (0.33 ± 0.02), and $\mathcal{L}_{div}$ (2.9 ± 0.1). No mode collapse or reward hacking was observed during training. This stability is largely due to the retention of the standard diffusion loss, which anchors the model to the learned data manifold while the physical rewards provide directional optimization. The lightweight 3-layer MLP used for the value function also prevents overfitting to the reward signal, ensuring efficient and robust convergence.

5. Generalization to Higher Fidelity and Long Horizons

NS-Diff maintains its performance advantage when scaled to higher resolutions and longer temporal horizons. While primary comparisons were conducted at a standard 256×256 resolution, we evaluated our framework on PhysVideoBench at a 32-frame, 512×512 resolution. As shown in Table 3, NS-Diff significantly outperforms the OpenSora2 (11B) baseline across both physical and perceptual metrics. This demonstrates that our physics-guided RL approach generalizes effectively to state-of-the-art large-scale models and remains robust during extended video synthesis.

Table 3. High-resolution (512px) and long-horizon (32 frames) evaluation on PhysVideoBench.

Method	ΔJ ↓	L_{div} $(\times 10^{-3})$ ↓	**Appear.** ↑	**Motion** ↑
OpenSora2 11B	0.81	4.8	69.5	88.1
Ours 11B	**0.28**	**2.6**	**73.3**	**92.5**

6. Failure Case Analysis and Physical Approximations

We emphasize that NS-Diff utilizes efficient, differentiable proxies to improve physical plausibility rather than attempting to guarantee strict correctness via costly simulators. Our Navier-Stokes-inspired terms effectively penalize temporal discontinuities and enforce incompressibility, while the Minimum Jerk prior reduces jittery motion. However, certain limitations

remain. In scenarios involving fast-moving, turbulent fluids or semi-transparent rigid bodies—such as cola being poured over ice—2D optical flow may prove insufficient for modeling 3D rotations or splashing effects. These upstream estimation errors can lead to misapplied rigid constraints and minor object deformations. Despite these challenges, our RL framework maintains motion continuity and prevents severe artifacts, demonstrating significant robustness even when guidance is imperfect.

Figure 2. Failure Case "A close-up shot of refreshing cola being poured into a clear glass filled with square ice cubes, carbonated bubbles fizzing and splashing, studio lighting, high resolution."