FlowVid: Taming Imperfect Optical Flows for Consistent Video-to-Video Synthesis

Supplementary Material

1. Webpage Demo

We highly recommend looking at our demo web by opening the https://jeff-liangf.github.io/projects/flowvid/ to check the video results.

2. Quantitative comparisons

2.1. CLIP scores

Inspired by previous research, we utilize CLIP [5] to evaluate the generated videos' quality. Specifically, we measure 1) Temporal Consistency (abbreviated as Tem-Con), which is the mean cosine similarity across all sequential frame pairs, and 2) Prompt Alignment (abbreviated as Pro-Ali), which calculates the mean cosine similarity between a given text prompt and all frames in a video. Our evaluation, detailed in Table 1, includes an analysis of 116 video-prompt pairs from the DAVIS dataset. Notably, CoDeF [4] and Rerender [6] exhibit lower scores in both temporal consistency and prompt alignment, aligning with the findings from our user study. Interestingly, TokenFlow shows superior performance in maintaining temporal consistency. However, it is important to note that TokenFlow occasionally underperforms in modifying the video, leading to outputs that closely resemble the original input. Our approach not only secures the highest ranking in prompt alignment but also performs commendably in temporal consistency, achieving second place.

Table 1. **CLIP score comparisons**. 'Tem-Con' stands for temporal consistency, and 'Pro-Ali' stands for prompt alignment.

Method	Tem-Con ↑	Pro-Ali ↑
CoDeF [4]	96.98	30.83
Rerender [6]	96.88	31.84
TokenFlow [1]	**97.30**	33.11
Ours	97.08	**33.20**

2.2. Runtime breakdown

We benchmark the runtime with a 512 × 512 resolution video containing 120 frames (4 seconds video with FPS of 30). Our runtime evaluation was conducted on a 512 × 512 resolution video comprising 120 frames, equating to a 4-second clip at 30 frames per second (FPS). Both our methods, FlowVid, and Rerender [6], initially create key frames followed by the interpolation of non-key frames. For these techniques, we opted for a keyframe interval of 4. FlowVid demonstrates a marked efficiency in keyframe generation, completing 31 keyframes in just 1.1 minutes, compared to Rerender's 5.2 minutes. This significant time disparity is attributed to our batch processing approach in FlowVid, which handles 16 images simultaneously, unlike Rerender's sequential, single-image processing method. In the aspect of frame interpolation, Rerender employs a reference-based EbSynth method, which relies on input video's non-key frames for interpolation guidance. This process is notably time-consuming, requiring 5.6 minutes to interpolate 90 non-key frames. In contrast, our method utilizes a non-reference-based approach, RIFE [3], which significantly accelerates the process. Two other methods are CoDeF [4] and TokenFlow [1], both of which necessitate per-video preparation. Specifically, CoDeF involves training a model for reconstructing the canonical image, while TokenFlow requires a 500-step DDIM inversion process to acquire the latent representation. CoDeF and TokenFlow require approximately 3.5 minutes and 10.4 minutes, respectively, for this initial preparation phase.

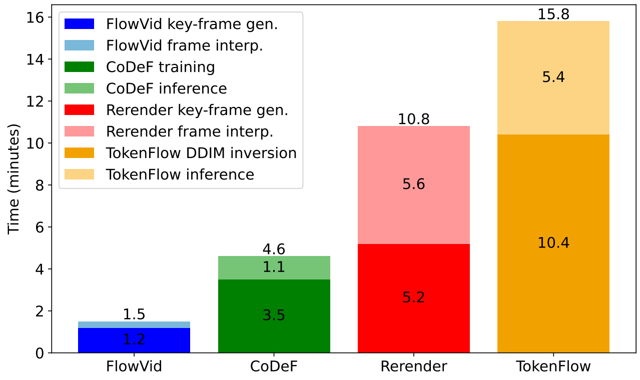

Figure 1. **Runtime breakdown** of generating a 4-second 512 × 512 resolution video with 30 FPS. Time is measured on one A100-80GB GPU.

3. Additional ablation study

v-prediction and ϵ-prediction While ϵ-prediction is commonly used for parameterization in diffusion models, we found it may suffer from unnatural global color shifts across frames, as shown in Figure 2. Even though all these two methods use the same flow warped video, the ϵ-prediction introduces an unnatural grayer color. This phenomenon is also found in Imagen-Video [2].

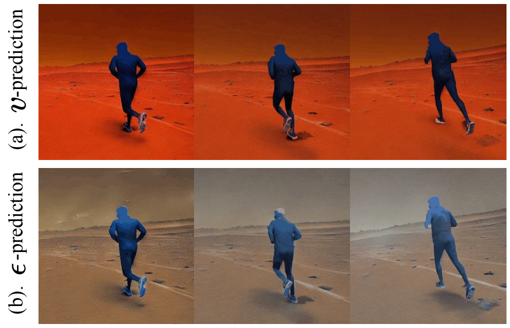
(a). $\boldsymbol{v}$-prediction
(b). $\boldsymbol{\epsilon}$-prediction

Figure 2. **Ablation study of different parameterizations.** ϵ-prediction often predicts unnatural global color while v-prediction doesn't. Prompt: `'a man is running on Mars'`.

References

[1] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 1

[2] 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. 1

[3] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1

[4] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. Codef: Content deformation fields for temporally consistent video processing. *arXiv preprint arXiv:2308.07926*, 2023. 1

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[6] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. *arXiv preprint arXiv:2306.07954*, 2023. 1