Supplementary materials of VipDiff: Towards Coherent and Diverse Video Inpainting via Training-free Denoising Diffusion Models

This is the supplementary material of our VipDiff, the following contents are included,

- Comparisons on video results with SOTA methods.
- More ablation studies on evaluating our **VipDiff**.
- Video results on demonstrating the generalization capability of our **VipDiff**.

1. Video Comparison Results

We present video comparisons results with three SOTA video inpainting methods, namely E2FGVI [4], FGT [6], ECFVI [3] and ProPainter [7]. The video results are shown in the .mp4 file named 'Comparisons_with_SOTA.mp4'. We have listed the video inpainting results for 4 different videos, with 3 videos on object-shaped masks and 1 video on stationery masks. The total video lasts about 29 seconds. Figure. 3 showed a screen shot of our video results, for each video, we first present a static zoom in views (Fig.3) for each methods with 3 seconds, and then followed by the videos. In the video results, we also present 2 different video inpainting samples generated by our **VipDiff**. Readers can stop at one specific frame for more detailed comparisons in need.

As we can see from the video results (in file 'Comparisons_with_SOTA.mp4'), our **VipDiff** is capable of generating stable and temporal-coherent video inpainting results, and much better visual quality over the existing state-of-the-art video inpainting methods. In addition, we also present two different video inpainting samples for each video, both of the samples are temporal coherent, which shows the diverse generalization ability of our **VipDiff**.

2. More Ablation Studies

We present more ablation studies in the .mp4 file named 'More_ablation_study.mp4'. We answer two questions here: 1) why do we need the reverse noise optimization process, can we simply trust the result generated by LDM $\hat{y}^k = \epsilon_\theta^d(z, t, \widetilde{x}_0^k)$ with pixel propagated prior input $\widetilde{x}_0^k$ as condition? One reason is that directly propagating pixels would cause color discrepancy issues along the boundary areas (check the beginning frames in the video file 'More_ablation_study.mp4'), even we adopt error compensation model provide by ECFVI [3], see Fig.1 top right case as an example (Pixel propagation conditioned LDM + without noise optimization), even LDM with pixel propagated prior $\widetilde{x}_0^k$ as condition, it has no strength for reducing the brightness issues caused through pixel propagation steps. Although these issue may not affect the whole video, it still reduces the overall video quality. While our **VipDiff** takes the original masked image x_0^k as input, using $\widetilde{x}_0^k$ as an guidance to optimize the random sampled Gaussian noise z, it trust the original masked image, so the input would not contain the incorrect color pixels, which hepls reduce the color discrepancy issues.And the other reason is that our **VipDiff** can be applied to other unconditional image-level diffusion models for optimizing the noise to generate coherent-video inpainting results (Check next section).

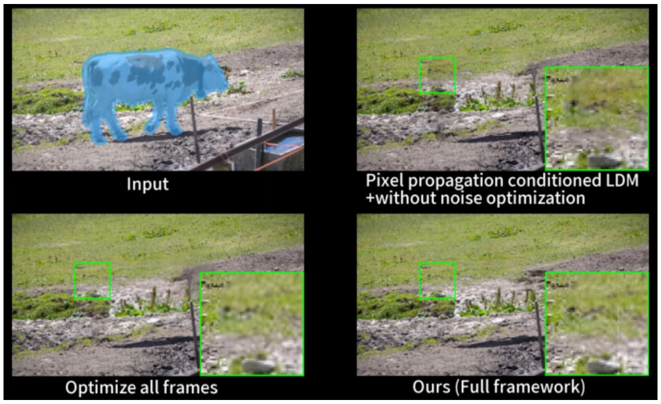
Input
Pixel propagation conditioned LDM +without noise optimization
Optimize all frames
Ours (Full framework)

Figure 1. Screenshot of the ablation study.

2) Why do not optimize every frame? We observe that even with iterative pixel propagation processes, optimize every frame would result in slightly frame blinking issues, we believe generating results frame by frame without mixing or blending pixel information from other reference frames would inevitably face this issues, that's why we propagate the generated results to to other frames and iteratively doing the pixel propagation and reverse noise optimization. The other reason is that optimizing every frame costs much more time than our current framework, since most of the generated results can be transferred to neighbouring frames by optical flows, to reduce the overall computational burdens.

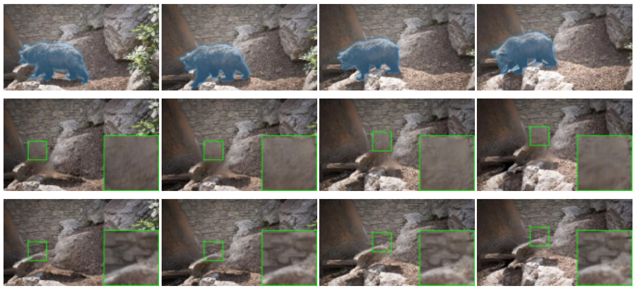

Figure 2. Visual results of adopting FGT' modules. From top row to the bottom: Input, results of FGT, Ours+FGT.

Adopt FGT's flow propagation modules. We conducted experiments by using the flow completion and pixel propagation modules from FGT [6] (shown in Fig 2, Ours+FGT). Our method successfully generates sharp and temporal-consistent results when combined with FGT's flow modules, reaffirming its generalization capabilities. We observed minor issues such as light jittering and slight blurriness after applying them, potentially stemming from their utilization of Poisson blending during pixel propagation. Moreover, FGT exhibits speed improvements, our average processing time can be reduced to 2.74s per image. This outcome highlights the feasibility of our method when employed with different flow modules.

3. Generalization Capability

Our **VipDiff** is not restricted to specific diffusion models, it can be embedded to other unconditional image generation diffusion models for generating temporal-coherent video inpainting results. We choose two variants of diffusion model, DDIM [5] and DDNM [1] with image-level diffusion model pretrained by [2]. The video results are shown in .mp4 file named 'Generalization_Capability.mp4'. For DDIM and DDNM, we present two video inpainting of each methods with different sampled noises, Fig 4 shows as screenshot of the video results. One can see from the video, that for those image-level diffusion models, our **VipDiff** is able to generate temporal-coherent and high-fidelity video inpainting results for all of them, which proves the generalization capability of our training-free framework.

References

[1] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 2021. 2

[2] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 2

[3] Jaeyeon Kang, Seoung Wug Oh, and Seon Joo Kim. Error compensation framework for flow-guided video inpainting. In *European Conference on Computer Vision*, 2022. 1

[4] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *IEEE/CVF conference on computer vision and pattern recognition*, 2022. 1

[5] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[6] Kaidong Zhang, Jingjing Fu, and Dong Liu. Flow-guided transformer for video inpainting. In *European Conference on Computer Vision*, 2022. 1, 2

[7] Shangchen Zhou, Chongyi Li, Kelvin C.K Chan, and Chen Change Loy. ProPainter: Improving propagation and transformer for video inpainting. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1

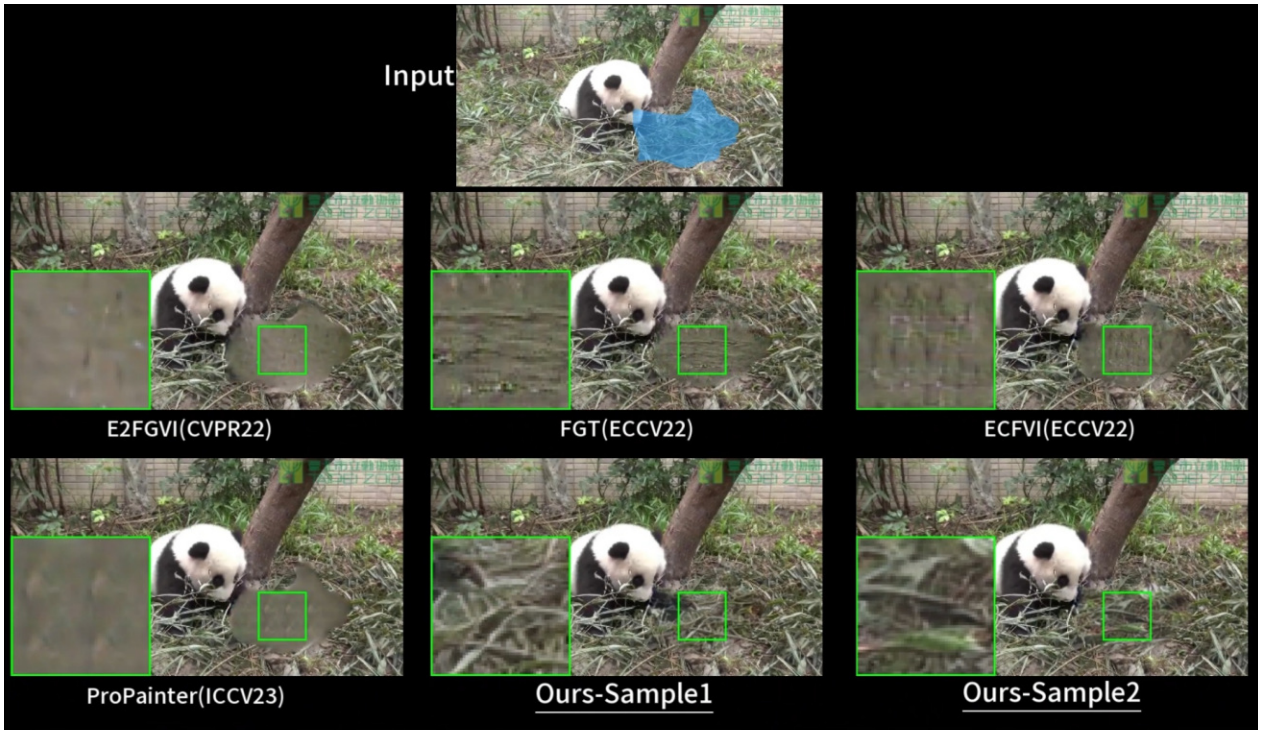

Figure 3. Screenshot of the video results with other competing methods. In 'Comparisons_with_SOTA.mp4', we also present two different video inpainting results by our **VipDiff**, demonstrating its diverse generalization ability.

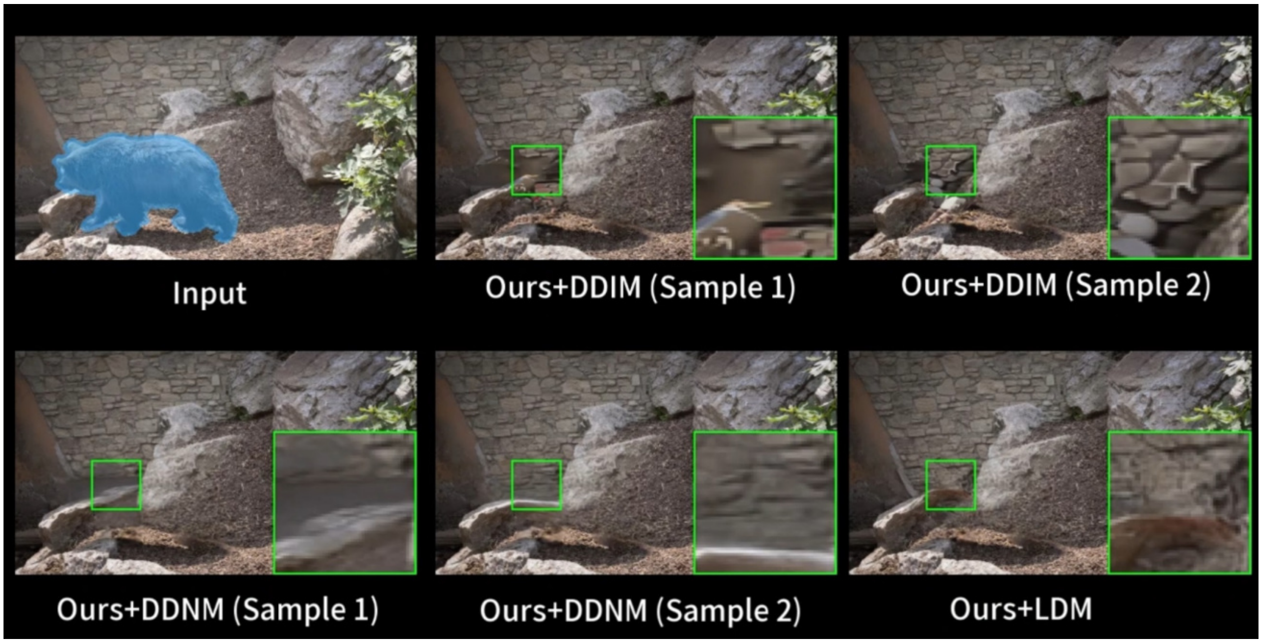

Figure 4. Screenshot of the video generation results of our **VipDiff** applied with other diffusion models.