Keyframe-Guided Creative Video Inpainting

Supplementary Material

A. Implementation Details

A.1. Image-to-Video Backbone

We use Stable Video Diffusion (SVD) [4] as the image-to-video backbone model considering its state-of-the-art visual quality and its open-source support. SVD is initialized with strong text-to-image generation prior, *i.e.*, Stable Diffusion 2.1 [10], with temporal aware model inflation and video finetuning. To create a 3D UNet that can generate a video sequence, temporal convolution and attention layers are inserted into the 2D spatial layers interleave. Temporal embedding is added to the attention layers to embed the frame order information. After that, the model is trained on carefully curated image and video datasets and learns the motion priors, resulting in a model capable of generating high-quality video at a resolution of 576×1024.

We adopt the image-to-video version of SVD, where the model is further finetuned with a still input image as conditioning. The image is firstly encoded with the VAE encoder then temporally repeated and concatenated to the UNet input. There are two variants of the model: the `svd-base` version predicts 14 frames, and the `svd-xt` predicts 25 frames. We use `svd-xt` for main results and `svd-base` for ablative study to save the computational budgets.

A.2. Trainable Parameters

As discussed in the main paper, we opt to maintain the pre-trained image-to-video priors for the inpainting task via efficient model repurposing with minimum trainable parameters scale. We only optimize the following parameters:

Input Convolution. We modified the pixel condition from repeated first frame latent (for image-to-video) to our symmetrically encoded masked pixels (for inpainting). The symmetric condition takes up twice the latent channels compared to only the first frame conditioning in the I2V model. To handle the additional channels, we zero-initialized a new input convolution layer. Since our new condition is symmetric, we treat the new and original convolution layers equally and optimize them together.

Low-Rank Adapter. The low-rank adapter for model repurposing is implemented with the LoRA [5] technique. This strategy is applied to the spatial and temporal self-/cross-attention layers since they better capture long-range correspondence. The decomposition matrices are injected into the linear layers for Q, K, V, and the output projection. We set the LoRA rank to 32 in our implementation, resulting in about 29 M additional parameters.

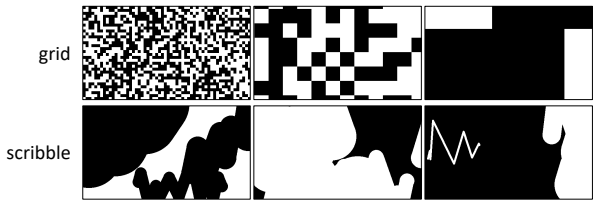

Figure 1. Random mask visualization. We visualize several random mask samples generated on the fly during training.

A.3. Mask Synthesis

To leverage in-the-wild videos for training as well as alleviate the data preprocessing cost, we adopt several strategies to synthesize random masks on the fly. The key idea is to make the mask highly randomized and avoid the model fitting to specific mask patterns. For spatial mask patterns, we leverage full mask, grid mask, scribble mask [11], and square mask [7], as visualized in Fig. 1. By adjusting the controlling hyperparameters, we can synthesize masks with fine-grained details, which encourages the model to improve robustness for arbitrary mask shape input. We additionally adopt mask inversion (by inverse the inpaint and uninpaint regions) and spatiotemporal augmentations to deal with mask changes in location and shape. All these practices result in our model's robustness, as well as efficiency in training on in-the-wild video data.

A.4. Training Configs

Since we freeze most of the pre-trained parameters, there's less risk of model collapse or overfitting. Therefore, we adopt a strategy by first pre-training the model on low resolution (320×576) and then finetuning with high resolution (576×1024), to improve the efficiency. We use AdamW optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$, at a learning rate of $5e-5$. For both training stages, the auto mix-precision is enabled to reduce memory overhead. For high-resolution fine-tuning, we additionally enable gradient checkpointing for each UNet block. We use a self-collected video dataset for training, which contains roughly 300K watermark-free videos. Compared to video foundation model pre-training, our method is lightweight in terms of data scale and is economical to reproduce. The training is done with 16 NVIDIA A100s at a batch size of 16 (1 for each device). The total optimization iteration is 50K (45K pre-training and 5K fine-tuning), taking less than 2 days to complete.

A.5. Image Inpainting Models

Our method leverages a two-stage solution for video inpainting by first allowing users to inpaint an arbitrary keyframe with established image inpainting models and then propagating the changes to other video frames. Such a solution can leverage powerful image models that exhibit better visual quality and text-following ability compared to their video counterparts. In practice, we use the following two image inpainting solution:

FLUX-inpainting. FLUX [3] is a powerful text-to-image model based on a latent rectified flow transformer architecture containing over 12 B parameters. It demonstrates state-of-the-art visual quality and prompt-following ability in image generation. We use an inpainting ControlNet developed by the community [1], where an axillary encoder is trained to take in the pixel condition. Since the model produces the best quality with its training resolution (768×768), we crop a square region containing the region of interest from the keyframe and then paste it back after inpainting.

Generative Fill. We use the generative fill feature in Adobe Photoshop [6] to deal with the background changing usage. The background mask is obtained with Segment Anything [9]. To avoid artifacts, we slightly erode the mask to omit foreground pixels close to the boundary.

B. Additional Results

DAVIS dataset. We demonstrate our method inpainting results on some challenging videos in DAVIS benchmark [8] in Fig. 2. We do not provide keyframe reference and the model operates in the unconditioned object removal use case. As shown in the figure, though the presented videos feature complex motion dynamics and mask shapes, our method produces coherent inpainting results. This robustness can be primarily attributed to our strong mask augmentation strategies.

Symmetric mask condition. We present additional ablative results examining our symmetric mask condition design in Fig. 3. Our analysis reveals that utilizing directly downsampled masks as conditions introduces boundary ambiguity, leading to back pixel leakage from the masked region to the final video output. This degradation is particularly pronounced in areas exhibiting complex mask shape and motion dynamics. Our proposed approach effectively addresses these limitations.

Longer sequence extension. In Fig. 4, we present a comparative analysis between the conventional temporal MultiDiffusion [2] and our proposed coarse-to-fine sampling strategy for extended sequence processing. While T-MultiDiff exhibits limitations in maintaining content fidelity, our method preserves content coherence through anchor frame placement across the entire sequence duration.

References

[1] alimama creative. Flux-controlnet-inpainting, 2024. 2

[2] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023. 2

[3] black-forest labs. flux, 2024. 2

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1

[5] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 1

[6] Adobe Inc. Adobe photoshop, 2023. 2

[7] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5792–5801, 2019. 1

[8] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 2

[9] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 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. 1

[11] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4471–4480, 2019. 1

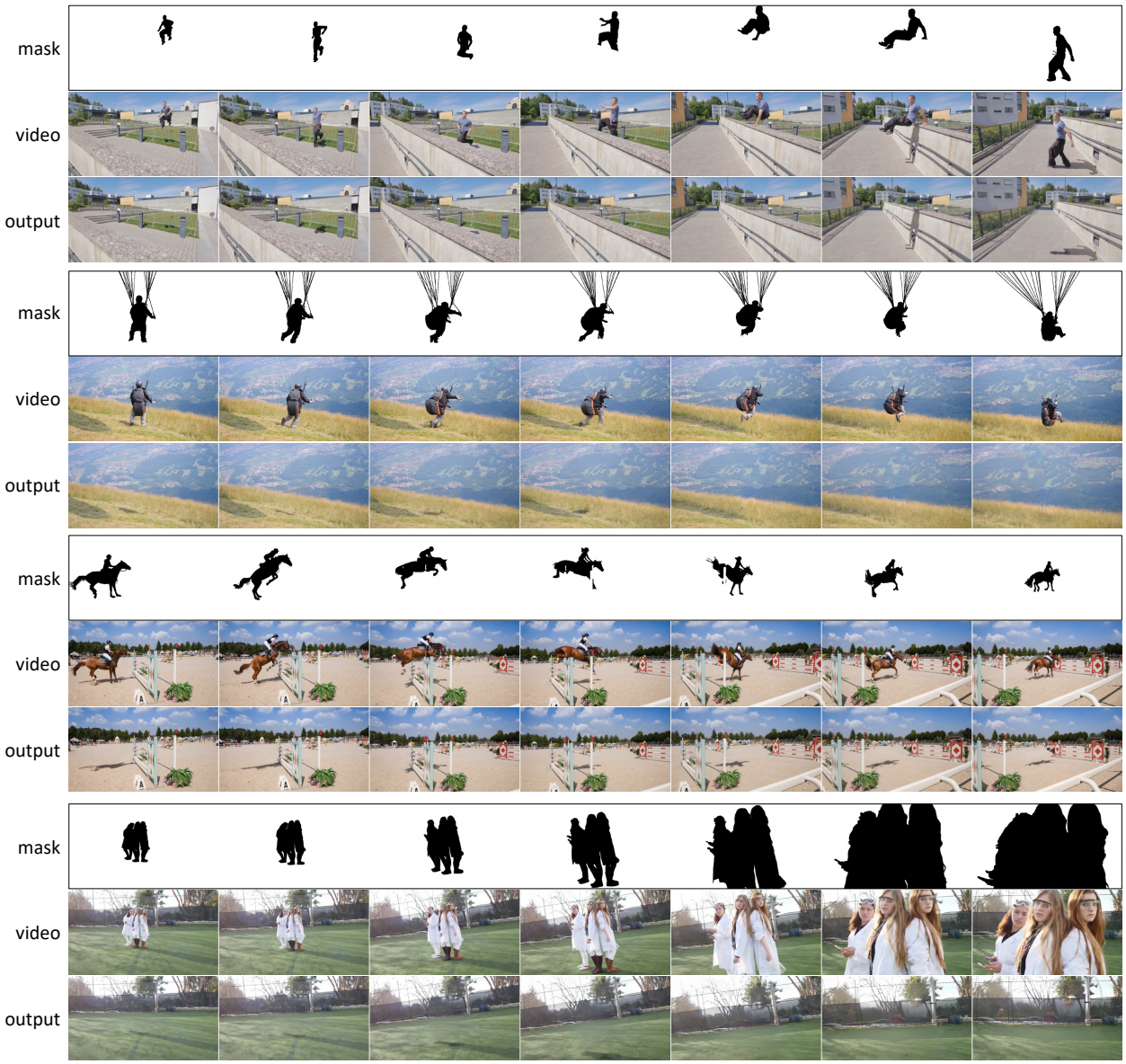

Figure 2. Video inpainting results on DAVIS dataset.

Figure 3. Ablative comparisons for symmetric mask condition. The red bounding boxes highlight the artifact region, where the black pixels from the masked area bleed into the final inpainting results.

Figure 4. Ablative comparisons for longer sequence extension. The red bounding boxes highlight the vanishing details.